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

July 22, 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 1's comments. We also include our comment-by-comment responses. In Section 3 we list Referee 2's comments. We address this referee's comments in a point-by-point fashion as well. 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 referees for their time, thoughtfulness, and feedback. Their 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, restructured the section to make it more clear for the reader. 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 typically 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

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

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 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 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.

   R: 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 there are more restrictions on the possible values the evaporation rates can 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 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."

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?

   R: We add the following sentence in line 236: "Note that the estimated lower limits of monomer evaporations from all the clusters except for the most stable ones are far above the 10^{-10} s^{-1} 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.

   R: The authors decided to keep the figures separately to make the visual inspection of each of the individual histograms corresponding to different estimated parameters more convenient.

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

R: We add the corresponding labels for the subplots.

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

R: To clarify the statement, we replace the sentence in line 284 with the two following sentences: " Notice that the evaporation rates of monomers for the cluster $\left(H_2SO_4\right)_2NH_3$ display strong inverse linear relationship, which is indicated by the pairwise marginal posterior distribution of the coefficients $\left(H_2SO_4\right)_2NH_3 \rightarrow \left(H_2SO_4\right)_2 + NH_3$ and $\left(H_2SO_4\right)_2NH_3 \rightarrow H_2SO_4NH_3 + H_2SO_4$, (see Figure **??**). Also, the estimated rate coefficients $\left(H_2SO_4\right)_2 \rightarrow H_2SO_4 + H_2SO_4$ and $H_2SO_4NH_3 \rightarrow H_2SO_4 + NH_3$ exhibit linear correlation."

(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 cal $K^{-1}mol^{-1}$ 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).

R: In line 65 we change the sentence to "In this study, we test which combinations of experimental data and fitted parameters lead to the best identification of the evaporation rates."

(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!

3 Referee 2 comments and our responses

Referee 2's summary: The author proposes to use the Markov chain Monte Carlo (MCMC) algorithm to solve the problem of cluster evaporation rate based on cluster distribution, and this is a novel idea for us to evaluate the thermal stability of clusters. But I have a question about the cluster distribution. The author uses ACDC to simulate the cluster distribution (from 1SA.1NH3 to 5SA.5NH3 box) instead of experimental data. Is this simulation result good enough to replace the experimental data? Simulation results are affected by accurate structure, calculation method and basis set. So I suggest that first the author expand the SA.NH3 system to a larger size (1.7 nm). Before using MCMC, simulate the SA.NH3 formation rate and compare it with the experiment data (Nature 502, 359-363, 2013) to illustrate the reliability of the simulation cluster distribution.

1. **R: The answer to reviewer's summary:** The objective of the present study is to investigate if we can extract evaporation rates from the type of data generated by experiments. Here we search to identify the combination of estimated parameters and experimental data which enables to obtain the estimates for evaporation rates with fair accuracy (i.e., the estimates with the variances comprising less then one order of magnitude).

   In Besel et all, 2020 (J. Phys. Chem. A.) is was shown that the 5x5 simulation box (which is used for generation of the synthetic data is the present study) produces results in a good agreement with the measurements obtained from the CLOUD chamber experiment. Howevere, the quality of data is not a major issue for our parameter estimation procedure, since the main point is not here to reproduce CLOUD data with the quantum chemical calculations, but to find the settings which

will give fair estimates of the evaporation rates in case if the data are available.

The MCMC results are not specific for the simulation box considered in the present study, but rather general. This is supported by the fact that although the size of the system (the number of clusters included into simulations) has impact on the particle formation rates at high temperatures (> 278 K), the particle formation rates and cluster concentrations produced using different simulation boxes are qualitatively similar. Thus the changes of the ACDC outputs due to the difference in the simulation box does not change for MCMC parameter estimation results.

The experimental data can differ from the synthetic data in the sense that they contain noise which originate from measurement instruments and uncertainies associated with experimental conditions (e.g., in CLOUD chamber experiments). Treating the noise inherent for experimental data will be the topic of our future studies.

2. **C:** "time-independent steady-state" in abstract could be revised to be "steady-state"

   R: We have made this change of wording.

3. **C:** The motivation and test results about the case of single temperature steady-state cluster distributions should be mentioned in the abstract;

   R: At the end of line 12, we have added:

   "We also estimated the evaporation rates using synthetic steady-state cluster concentration data at one temperature (which has appeared in previous literature) and compared our two study cases to this setting. Both the transient and two-temperature steady-state concentration data estimated the evaporation rates with less variance than the steady-state one temperature case. "

4. **C:** The best result in this study is the case for steady-state concentration with two temperatures. Is this conclusion general or very specific? How sensitive towards the number of ammonia concentrations and the box size (referring to the cluster types here) is this conclusion?

   R: The MCMC results are not specific for the simulation box considered in the present study, but rather general. This is supported by the fact that although the size of the system (the number of clusters included into simulations) has impact on the particle formation rates at high

temperatures (> 278 K), the particle formation rates and cluster concentrations produced using different simulation boxes are qualitatively similar. Thus the changes of the ACDC outputs due to the difference in the simulation box does not change for MCMC parameter estimation results. In Besel et all, 2020 (J. Phys. Chem. A.) is was shown that the 5x5 simulation box (which is used for generation of the synthetic data) produces reasonable results with a good agreement with the measurements obtained from the CLOUD chamber experiment. Additionally, the boundary conditions for the outgrowing clusters (the choice of the clusters that are considered as formed particles) has only minor influence on the simulation results, given that the simulated system of clusters is defined in a reasonable way (see Besel at al., 2020, J. Phys. Chem. A).

In general, the accuracy of the MCMC results increases when we include additional data. In particular, including more concentration data measured at different ammonia concentrations will yield better estimates for the evaporation rates. The sensitivity of the estimates to the number of ammonia concentrations will be considered in the future work. In the present study we rather focus on the question which combination of estimated parameters and concentration data will produce an accurate estimates for the evaporation rate.

The data of steady-state concentration with two temperatures allowed us to apply two general principles of inverse problems/Bayesian estimation to the problem of estimating evaporation rates. First, the two temperature data set enabled us to reformulate the problem in a numerically effective way (in terms of enthalpy and entropy) that reduced the number of unknown parameters we sought to estimate. Second, the reformulated differential equation describing the time evolution of the concentrations was more numerically stable than the original expression (the stiffness of the equation was reduced in the reformulated form). This made our estimates for the rates less sensitive to small perturbations/errors.

In addition, the fact that the entropies and enthalpies were strongly correlated made them an effective parametrization. The strong inverse correlations have a physical explanation. Firstly, both enthalpy and entropy follow from the partition function of the molecular complex, and their functional forms are partly similar. Practically, if a cluster has really strong bonds between the molecules, then that means the formation enthalpy is very negative, and also the intermolecular vibrational frequencies corresponding in a broad sense to vibrations involving those

bonds (note that these frequencies dominate the "variable part" of the formation entropy, as the entropy effect from the loss of translational and rotational degrees of freedom is almost a constant factor) are fairly high, meaning that the entropy loss in forming the cluster is large. So if the formation enthalpy is very negative so is also the formation entropy. Conversely, if the cluster is only quite weakly bound, the formation enthalpy is only slightly negative, and the intermolecular frequencies can be very low, leading to a less negative (though still negative of course) formation entropy.

In line 343 we add the Section 3.5."Discussion and future work", where we place the above-written answer to the reviewer's question.

At the end of the "Discussion and future work" section we add the paragraph:

"Note that experimental data can differ from the synthetic data in the sense that they contain noise which originate from measurement instruments and uncertainties associated with experimental conditions (e.g., in CLOUD chamber experiments). Treating the noise inherent for experimental data will be the topic of our future studies."

5. **C:** VODE mentioned in L107 may be different from the solver used in McGrath et al. (2012) (ode15s). If so, "A detailed description of this program was published in McGrath et al. (2012)." should be deleted and a simple benchmark should be made to compare different solvers.

   R: We compared the ode15s with those for the vode when creating synthetic data, and they were producing practically identical results.

6. **C:** For table 3, why the minimal values of H and S are set to be -400?

   R:

   (a) A narrower range could have been used for the formation enthalpies, since the upper limit correspond to evaporation which in practice almost always happens before growth. The lower limit formally corresponds to zero evaporation. Physically, an upper limit of 0 can be justified by the fact that > 0 formation enthalpies would mean no attractive interactions at all, which is obviously physically wrong for polar, H-bonding molecules such as H_2SO_4 and NH_3. For the lower limit (-400) we mean that on average each H_2SO_4 cluster is bound more strongly than in the (extremely strongly bound) $HSO_4^- * H_2SO_4$ cluster, for which the best available computational studies indicate a binding enthalpy

roughly around -40 kcal/mol. So it seems unlikely that the average binding per H2SO4 could be tens of kcal/mol stronger than that in the larger clusters where the effect of charge should be much smaller. In any case, a formation enthalpy below -400 kcal/mol means practically zero evaporation so it makes no difference if this is set to a lower value. On the other hand, the largest cluster included into the system has 5 H_2SO_4 and 5 NH_3, so 10 molecules, and -400 kcal/mol would mean -40 kcal/mol per molecule, which 1) corresponds to the strongest known cluster in the system and 2) means evaporation of practically zero.

(b) For the formation entropies, the 0 cal/Kmol upper limit can be justified as follows: clustering has to have a negative ΔH, as we are reducing the number of gas molecules (and converting translational and rotational degrees of freedom into much more constrained vibrational degrees of freedom). Probably a much lower upper limit could have been used, but certainly the ΔS values can never be > 0. For the lower limit, we state that the typical per-molecule ΔS for clustering is around -30 cal/Kmol, with a typical variation of up to +-10 cal/mol K, see Kürten, 2019. So for the largest clusters the upper limit corresponds to a per-molecule ΔS of -40 cal/Kmol. In this case, all the new vibrational degrees of freedom formed in the product clusters are quite rigid, i.e. have very low entropy.

(c) After the line 153 we edit an explanation on the sampling limits selected for the thermodynamic parameters: "Next we justify the limits selected for data setting 2, where we sample thermodynamic parameters. For the formation enthalpies an upper limit of 0 kcal/mol is chosen by the fact that a positive ΔH would mean an absence of attractive interactions in the molecular cluster, which is physically incorrect for polar, H-bonding molecules such as H_2SO_4 and NH_3. For the lower limit (-400 kcal/mol) we mean that on average each H_2SO_4 is bound substantially stronger than in the $HSO_4^- * H_2SO_4$ cluster, for which the most recent computational studies indicate a binding enthalpy roughly around -40 kcal/mol. Another motivation for the prior distribution selected for the cluster formation enthalpies comes from the fact that the largest cluster included into the system has 5 H_2SO_4 and 5 NH_3, so 10 molecules, and -400 kcal/mol would give an enthalpy of -40 kcal/mol per molecule, which 1) corresponds to the strongest known cluster in the system and 2) which implies that the evap-

oration rate is zero for all purposes of measurement.

Next, we set the upper limit for the formation entropies to 0 cal/K/mol, since molecule clustering must have a negative ΔH,as the number of gas molecules is reduced (and translational and rotational degrees of freedom are converted into much more constrained vibrational degrees of freedom). For the lower limit of -400 cal/K/mol, we state that the typical per-molecule ΔS for clustering is around -30 cal/K/mol, with a typical variation of up to +-10 cal/mol K, see Kürten, 2019. So for the largest clusters the upper limit corresponds to a per-molecule ΔS of -40 cal/Kmol. In this situation,all the new vibrational degrees of freedom formed in the product clusters are quite rigid, i.e. have very low entropy."

7. **C:** L156, "ACDC plus VODE" should be revised to be "ACDC based on VODE"

   R: We have rewritten this paragraph for clarity, and this emphasis for ACDC has been redirected to Section 2.1. The new paragraph which includes the old line 156 is as follows:

   " We make our initial guess $\boldsymbol{\theta} = \boldsymbol{\theta}_{old}$, where $\boldsymbol{\theta}_{old}$ is the flat distribution which obeys the estimates in Tabs. 3-4. The limits are explained in Section 2.2.3. 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 likelihood probability distribution for $\boldsymbol{\theta}_{old}$, a new candidate for the unknown parameter values, $\boldsymbol{\theta}_{new}$, is sampled using the proposed Gaussian likelihood 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 ... "

8. **C:** L233, "upper limit" needs to be explained further.

   R: We have edited the sentence to read "... all the parameter chains for the evaporation rates have values bounded above by an upper limit which differs for different evaporation rates.'

9. **C:** L244, "well-defined" need to be defined.

   R: We have rewritten the sentence to state:

   "All the evaporation rates larger than $10^{-3}s^{-1}$ are well-identified (see subfigures labelled 1, 2, 4, 5, 7, 10, 12, 16, 18, 22, 27, 31 and 35

in Figures 3- 4), in the sense that their estimated variances are well within our accepted error range of less then one order of magnitude."

Identification of molecular cluster evaporation rates, cluster formation enthalpies and entropies by Monte Carlo method

Anna Shcherbacheva[1], Tracey Balehowsky[2], Jakub Kubečka[1], Tinja Olenius[3], Tapio Helin[4], Heikki Haario[4,5], Marko Laine[5], Theo Kurtén[6,1], and Hanna Vehkamäki[1]

[1]Institute for Atmospheric and Earth System Research, P.O. Box 64 00014 University of Helsinki, Finland
[2]Department of Mathematics and Statistics Subunit, P.O. Box 64 00014 University of Helsinki, Finland
[3]Department of Environmental Science and Analytical Chemistry & Bolin Centre for Climate Research, Stockholm University, Svante Arrhenius väg 8, SE-11418 Stockholm, Sweden
[4]LUT School of Engineering Science, Lappeenranta-Lahti University of Technology, P.O.Box 20 FI-53851 Lappeenranta, Finland
[5]Finnish Meteorological Institute, P.O. Box 503, FI-00101 Helsinki, Finland Finland
[6]Department of Chemistry, P.O. Box 55 FI-00014 University of Helsinki, Finland

Correspondence: Anna Shcherbacheva (anna.shcherbacheva@helsinki.fi)

Abstract. We address the problem of identifying the evaporation rates for neutral molecular clusters from synthetic (computer-simulated) cluster concentrations. We applied Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) algorithm to determine cluster evaporation/fragmentation rates from known cluster distributions, assuming that the cluster collision rates are known. We used the Atmospheric Cluster Dynamic Code (ACDC) with evaporation rates based on quantum chemical calculations to generate cluster distributions for a set of electrically neutral sulphuric acid and ammonia clusters. We then treated these concentrations as synthetic experimental data, and tested two approaches for estimating the evaporation rates. First we have studied a scenario where at one single temperature time-dependent cluster distributions are measured before the system reaches a ~~time-independent~~ steady-state. In the second scenario only steady-state cluster distributions are measured, but at several temperatures. This allowed us to use multiple sets of concentrations at different temperatures. Additionally, in the latter case the evaporation rates were represented in terms of cluster formation enthalpies and entropies which were considered to be free parameters. This reparametrization reduced the number of unknown parameters, since several evaporation rates depend on the same cluster formation enthalpy and entropy values. We also estimated the evaporation rates using synthetic steady-state cluster concentration data at one temperature (which has appeared in previous literature) and compared our two study cases to this setting. Both the transient and two-temperature steady-state concentration data estimated the evaporation rates with less variance than the steady-state one temperature case.

We show that in the second setting, even if only two temperatures were used, the temperature-dependent steady-state data outperforms the first setting for parameter ~~identification~~estimation. We can thus conclude that for experimentally determining evaporation rates, cluster distribution measurements at several temperatures are recommended over time-dependent measurements at one temperature.

1 Introduction

The formation of molecular clusters, and their subsequent growth to aerosol particles, is an important yet poorly understood process in our atmosphere. Clusters and aerosols affect both climate, air chemistry (?)(?), evapotranspiration in forest environments (?)(?), and many other atmospheric processes (?)(?).

Recent developments in mass spectrometers have enabled the detection, quantification, and chemical characterization of ionic clusters containing between one and some tens of molecules at atmospherically relevant mixing ratios [1] (?; ?; ?; ?; ?; ?)(??????). Molecular clusters in atmospheric conditions are predominantly electrically neutral, and must thus be charged prior to mass spectrometric detection. This may affect the measurement results, as only part of the sample molecules or clusters may be charged (?)(?), and the charging may also alter cluster compositions. For example, for sulfuric acid base clusters, negative charging tends to lead to loss of base molecules, and positive charging to loss of acid molecules (?)(?). Modelling is thus needed to connect measured ion cluster distributions to the original neutral population.

Even when the atmospheric cluster distribution can be accurately deduced from experimental data, this does not quantify the individual kinetic parameters, such as the cluster collision and evaporation rates (?)(?). Collision rates may be computed from kinetic gas theory or classical trajectory simulations with reasonable accuracy (?)(?), although recent research has shown that long-range attractive interactions may enhance collision rates (?)(?), for example by around a factor of 2-3 for H2SO4-H2SO4 collisions (?)$H_2SO4 - H_2SO_4$ collisions (?). These relatively minor uncertainties in the collision rates are dwarfed by the error margins of cluster evaporation rates. In computational applications, evaporation rates are usually computed using the detailed balance assumption together with the free energies of cluster formation, which can in turn be computed using quantum chemical (QC) methods(?; ?; ?; ?; ?), (?????). Unfortunately, the evaporation rates depend exponentially on the free energies variations of several kcal/mol between different QC methods thus translate into orders of magnitude differences in evaporation rates , (?, ?)(??).

Despite uncertainties involved in computational estimates of collision and evaporation rates, cluster population dynamic models based on Becker-Döring equations have been able to predict the sulphuric acid concentration dependence of cluster concentrations (?)(?), and even absolute particle formation rates (?) (?) in sulphuric acid-ammonia and sulphuric acid-DMA systems, without empirical model calibration or parameter tuning. The Becker-Döring equations are a system of Ordinary Differential Equations (ODE), which account for cluster birth and death processes (which depend on the collision and evaporation rates), as well as external cluster sinks and sources. In both studies (**?** and **?**), these equations were implemented through the Atmospheric Cluster Dynamic Code (ACDC) (?)(?), using kinetic gas theory collision rates, and standard quantum chemistry techniques for computing cluster formation free energies (and thus evaporation rates).

In mathematical terms, the prediction of cluster concentrations using known collision and evaporation rates is called the forward problem. The associated inverse problem is to use known cluster concentrations to deduce the collision and evaporation rates. The inverse problem can be addressed with Bayesian approaches such as Markov chain Monte Carlo (MCMC) methods. In a recent paper (?)by ?, Differential Evolution (DE) MCMC (see ?) (?) was applied to determine evaporation rates for nega-

[1] around or below one part per trillion (ppt)

tively charged sulphuric acid and ammonia clusters (containing up to five of each type of molecules, with the HSO_4^- ion here defined as an "acid"). This study used steady-state cluster concentrations measured in the CLOUD [2] chamber experiment at constant temperature, with varying sulphuric acid and ammonia concentrations (we refer to **?** for details relevant to the experimental data). Collision rates were taken from kinetic gas theory. **?** concluded that these data were insufficient for ~~identification~~ estimation of all the evaporation rate coefficients. Another recent paper ~~(?)~~ (**?**) reported thermodynamic data (cluster formation enthalpies and entropies) for 11 neutral sulphuric acid and ammonia clusters. In the CLOUD experiment, these were deduced from new particle formation (NPF) rates measured at 5 different temperatures, over a wide range of sulphuric acid and ammonia concentrations. Most of the thermodynamic parameters could not be narrowly constrained, as the ranges of cluster formation enthalpies and entropies that reproduced the measured NPF rates were quite wide. However, for each cluster only one monomer evaporation rate was taken into account (either acid or base). Furthermore, the NPF rates obtained using the fitted parameters were systematically lower than the measured ones for warmer temperatures ($\geq$ 248 K).

In this study, we test which combinations of experimental data and fitted parameters ~~leads~~ lead to the best identification of ~~cluster~~ the evaporation rates. As experiments are expensive and time-consuming to perform, we use synthetic cluster concentration data created from ACDC simulations to test if the use of time-dependent cluster distribution data would significantly improve the accuracy of the evaporation rates. Use of synthetic data also allows us to know for sure if our inverse modelling actually produces the correct kinetic parameters or not, which would not be possible with experimental concentration data. As in the **?** study, we compute collision rates from kinetic gas theory, while the evaporation rates used to generate our synthetic data are calculated from Gibbs free energies published by **?**. Note that the conclusions of this study are not sensitive to the accuracy of the quantum chemical data, as our focus is on the inverse problem of how to determine evaporation rates from known concentrations rather than the forward problem.

For simplicity, we consider the case of neutral sulphuric acid-ammonia clusters containing up to five of each type of molecules. Studying neutral clusters has the advantage that we can restrict ourselves to a smaller set of kinetic parameters, and ignore uncertainties related to charging and neutralization processes. In situations where a large fraction of the clusters are charged, accurate modelling would require at least three times as many parameters, as both the negative, positive and neutral cluster populations interact with each other. The downside of this simplification is that we lose the direct connection to potential real-life experiments, as neutral atmospheric clusters cannot currently be measured without first charging them.

We investigate two different scenarios for estimating evaporation rates. First, we test the use of time-dependent cluster concentrations measured before the system has attained a steady state. This is motivated by the fact that this transient data should provide additional information about the speed of the processes, which is missing from the steady-state data. Second, we apply the approach of **?**, and express the evaporation rates as parameterized functions of the temperature, with the cluster formation enthalpies and entropies (assumed here to be temperature-independent) as the unknown parameters. This reparametrization is useful for two reasons. First, since the formation enthalpies and entropies of the monomers can be set to zero, and since several evaporation rates depend on the same enthalpy and entropy values, the dimension of the unknown parameter space for our problem is actually reduced, despite the apparent doubling of the number of parameters. Second, utilizing the temperature

[2] Cosmics Leaving OUtdoor Droplets

dependence allows us to produce and use arbitrarily many synthetic data sets at various temperatures, which mathematically has a regularizing effect on the problem. Note that unlike in **?**, all possible evaporation processes, including cluster fissions into two daughter clusters, are taken into consideration.

2 SIMULATION METHODS

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.

2.1 Generation of synthetic data

The 16 cluster types included in our study are summarized in Table **??**. To save computational time, we have excluded clusters where the number of acid and base molecules differs significantly from each other. Irrespective of the level of theory, quantum chemical data predict that these clusters will have very high evaporation rates, leading to negligibly small concentrations. This is also supported by mass spectrometric measurements showing that the clusters with highest concentrations have roughly the same number of acid and base molecules ~~(see ?, ?, ?, ?)~~(????). The ammonia monomer mixing ratio is assumed to remain constant in each individual simulation, and varied between 5 and 200 ppt. (These correspond to concentrations of 1.3×10^8 and 5.0×10^9 molecules per cm^3 for the temperature ranges studied here, respectively). The sulfuric acid monomer source rate is kept constant at $\mathrm{Q} = 6.3 \times 10^4\ \mathrm{cm}^{-3}\mathrm{s}^{-1}$ in all simulations~~(see Table ??)~~. See Table **??** for the summary of ammonia mixing ratio and the source of sulphuric acid monomer used for the ACDC simulations.

Synthetic concentration data for such neutral clusters were generated by the following method.

~~The evaporation rate coefficients computed in ?, the associated collision rates as determined by~~ First, we computed the collision rates using the Eq. **??** ~~-?? ,~~ from kinetic gas theory. Then, we used these values for the collision rates along with Eq. **??** and the Gibbs free energies computed from Eq. **??** to obtain the evaporation rates. Next, to compute the Gibbs free energies, we substituted the values for cluster formation enthalpies and entropies given by **?** into Eq. A5. Additionally, we consider the losses on the CLOUD chamber walls which depend on the cluster size computed with Eq. **??** (**?**) and a dilution loss of $S = 9.6 \times 10^{-5}\ \mathrm{s}^{-1}$. These values for the rates and losses were substituted into the ~~wall losses calculated by Eq. ??, and dilution losses of ($S_i = 9.6 \times 10^{-5}\mathrm{s}^{-1}$), are substituted in to the~~ ACDC algorithm ~~?, which~~ (**?**), 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. **??**. ~~Similarly to the earlier paper ?, we~~ We then integrate the system produced by ACDC using the Fortran ordinary differential equation solver VODE ~~(?)~~(**?**). A detailed description of this ~~program~~ strategy for solving the forward-problem of finding the cluster concentration rates from Eq. A1 was published in **?**. 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 as it was previously mentioned.

The above method we used for producing synthetic concentration rates is similar to the one described in **?**. We note that unlike ~~in **?**, the **?**~~, in this paper, our particle system is considered at various temperatures~~in this paper~~.

~~Two data setswere generated~~Using the above algorithm, model configuration and parameters, we generated two data sets. First, time evolution of the concentrations $\mathrm{Y_i(t)}$ is computed for time values less than the time at which the system has attained the steady state. The maximum time we run is 60 minutes from beginning of the simulation, in the above model configurations. 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 ~~transient concentration measurements~~ time-dependent concentration data for each of the cluster types i measured from beginning to the end of each simulation, before the system has attained a steady state.

Secondly, we solve for time-independent steady-state concentrations for all the cluster types for two temperatures comprising 278 K and 292 K. In both data configurations, the steady-state cluster concentrations are calculated as the average of the concentrations determined for time instances $t_1 := 50$ min and $t_2 := 60$ min. The measure of how close the system has reached to the steady state is monitored by a convergence parameter, which is the ratio of the concentrations at times t_2 and t_1, taken in each case for the cluster for which this ratio deviated most from unity, ~~**?**~~(**?**).

In both data settings, the simulation outputs are amended with the measurement errors sampled from a multivariate, non-correlated, Gaussian distribution, where the variance of the distribution depends on cluster type i, temperature T and time instance t. While a simplification of noise characteristics of the real data obtained from a mass spectrometer, we impose that the standard deviation of the noise comprises 0.001% of the original concentration.

Note that apart from generation of synthetic data, we apply the ACDC as a ~~kinetics~~ kinetic model of cluster population in the MCMC simulations. The ACDC outputs are compared to the synthetic measurements and explained in Section 2.2.

Table 1. Neutral molecular clusters included into model system. The first column indicates the number of sulphuric acid molecules, the second column stands for the number of ammonia in the cluster.

Number of H_2SO_4 molecules	Number of NH_3 molecules	Number of clusters
0	1	1
1	0-1	2
2	0-2	3
3	1-3	3
4	2-5	4
5	3-5	3

2.2 Markov chain Monte-Carlo simulations

The evaporation rate coefficients $\gamma_{i+j\rightarrow i,j}$ appearing in the ACDC simulation of **??** are treated as unknown parameters. Now we describe how we estimate the evaporation rates from the noisy synthetic data sets obtained by the method described

Table 2. Monomer concentrations used in simulations

$[H_2SO_4]$ monomer source	$[NH_3]$ concentration
6.3×10^4 cm^{-3}s^{-1}	5 ppt
6.3×10^4 cm^{-3}s^{-1}	35 ppt
6.3×10^4 cm^{-3}s^{-1}	100 ppt
6.3×10^4 cm^{-3}s^{-1}	200 ppt

in Section 2.1. We first give a general overview of the basic Metropolis algorithm (**?**), 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. Our purpose is to determine all the parameter sets that reproduce the synthetic data within their noise level (which is known). We do this using Markov Chain Monte Carlo (MCMC) sampling.

The ~~MCMC approach computes a~~ objective of MCMC in parameter estimation is to identify possible parameter values which yield the best fit with the experimental data. Unlike optimization algorithms that produce one best combination of parameter values, 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 converge 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 ~~of the parameters as point-wise likelihood approximations across the~~ which gives the likelihood of observing each of the parameters given the concentration data.

2.2.1 The Metropolis algorithm

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 typically selected based on the previous knowledge of 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 ~~space. The algorithm samples the candidate parameter points from a predefined proposal distribution~~ 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, ~~and then either accept or reject it,~~ according to ~~how closely the output model fits the data. The fundamental technique is the Metropolis algorithm (**?**). The sets of parameters which produce cluster concentrationswithin the allotted noise level of~~ the ~~data are kept in the sampled distribution. Finally,~~ least-squares fit of

the output to the data, which measures the difference between the modelled $\mathbf{Y}_{\mathrm{mod}}$ and measured $\mathbf{Y}_{\mathrm{exp}}$ cluster concentrations:

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

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 $F(\boldsymbol{\theta}_{\mathrm{new}})$ is compared to the least-square sum from the previous step $F(\boldsymbol{\theta}_{\mathrm{old}})$. 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 $F(\boldsymbol{\theta}_{\mathrm{new}}) > F(\boldsymbol{\theta}_{\mathrm{old}})$, the point will be accepted with the probability

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

If the candidate point is accepted, the parameter combination $\boldsymbol{\theta}_{\mathrm{new}}$ is added to the chain, in the ~~approximation of the posterior distribution is constructed from the retained parameter sets. We remark that to create a reliable sample from the underlying parameter distribution, many different parameter combinations must be tested; that is , the length of the MCMC chain must be large enough (?, ?). In both our studies, the MCMC chain length typically comprised 3 million samples. The MCMC acceptance probabilities (defined below) in each of the cases were about 88.0%, which is a typical level of acceptance since the "forward" ACDC model (in which the rate coefficients are known) is deterministic.~~ opposite case the old value is replicated in the chain. Finally, the value $F(\boldsymbol{\theta}_{\mathrm{old}})$ is replaced with $F(\boldsymbol{\theta}_{\mathrm{new}})$ 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 ~~(?)~~(?). This variant is called the Delayed Rejection Adaptive Metropolis (DRAM), introduced in **?**. We briefly explain our approach below.

~~Parameter identification is conducted using the **'mcmcstat'** toolbox implemented for FORTRAN (see **?**, **?**). See the description and the examples of usage on the web page~~

2.2.2 The DRAM algorithm

Similar to the basic Metroplois algorithm, the DRAM is initialized with a chosen prior distribution and initial guess for parameter values.

~~First, an initial 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. In our case, we chose the flat prior, but impose some domain~~

restrictions for sampling from this prior to exclude unphysical parameters (see Tables ??-??). We make our initial guess $\boldsymbol{\theta} = \boldsymbol{\theta}_{old}$, where $\boldsymbol{\theta}_{old}$ is the flat distribution which obeys the estimates in Tabs. 3-4. The limits are explained in Section 2.2.3. We also assume that the conditional probability distributions for the parameters given the concentration data are of Gaussian type.

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.* However, we assume that the evaporation rates with orders of magnitude less than $10^{-10}\mathrm{s}^{-1}$ are irrelevant in practise, since such an evaporation event is highly improbable, and it is very likely that instead the cluster will grow further by collisions. Similarly, when the evaporation rate is of the order of magnitude more than $10^{+10}\mathrm{s}^{-1}$, it is reasonable to expect that the cluster will most certainly evaporate before it has a chance to grow further. With these assumptions, the prior distribution of the evaporation rates spans over several orders of magnitude, and the base 10 logarithm of evaporation rates was sampled from the range of -12 to 12.

Domain limitations for two data settings under consideration imposed to exclude non-physical parameters in parameter identification procedure. Data settings Estimated parameters Minimal value Maximal value Data setting 1 Base 10 logarithms of -12 12 evaporation rates (in s^{-1})Data setting 2 Cluster formation enthalpies (kcal mol^{-1}) and -400 0entropies (cal K^{-1} mol^{-1}) -400 0

Additional domain limitations for the data setting 2 from Table ?? (identification of thermodynamic data), where the cluster formation enthalpy of the i-th cluster is denoted by $\Delta \mathrm{H}_i$ and the symbols A and N stand for ammonia and sulphuric acid, respectively. $\Delta \mathrm{H}_{2\mathrm{A}} > \Delta \mathrm{H}_{2\mathrm{A1N}}$ $\Delta \mathrm{H}_{3\mathrm{A2N}} > \Delta \mathrm{H}_{4\mathrm{A2N}}$ $\Delta \mathrm{H}_{1\mathrm{A1N}} > \Delta \mathrm{H}_{2\mathrm{A1N}}$ $\Delta \mathrm{H}_{4\mathrm{A2N}} > \Delta \mathrm{H}_{4\mathrm{A3N}}$ $\Delta \mathrm{H}_{2\mathrm{A1N}} > \Delta \mathrm{H}_{3\mathrm{A1N}}$ $\Delta \mathrm{H}_{4\mathrm{A3N}} > \Delta \mathrm{H}_{4\mathrm{A4N}}$ $\Delta \mathrm{H}_{4\mathrm{A4N}} > \Delta \mathrm{H}_{5\mathrm{A5N}}$ $\Delta \mathrm{H}_{3\mathrm{A1N}} > \Delta \mathrm{H}_{3\mathrm{A2N}}$ $\Delta \mathrm{H}_{4\mathrm{A4N}} > \Delta \mathrm{H}_{4\mathrm{A5N}}$

Once initialized, the following iterative steps take place. From the proposed distribution , a guess for the parameter values sampled, denoted likelihood probability distribution for $\boldsymbol{\theta}_{old}$. Then, a new candidate for the unknown parameter values, $\boldsymbol{\theta}_{new}$, is sampled from the old point using Gaussian proposal using the proposed Gaussian likelihood distribution. We use ACDC plus VODE to simulate concentration outputs with parameter 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

$$\mathrm{p}_{\mathrm{acc}}(\boldsymbol{\theta}_{\mathrm{old}}, \boldsymbol{\theta}_{\mathrm{new}}) = \min\left\{1, \frac{\mathrm{p}(\mathbf{Y}_{\mathrm{exp}}|\boldsymbol{\theta}_{\mathrm{new}}))}{\mathrm{p}(\mathbf{Y}_{\mathrm{exp}}|\boldsymbol{\theta}_{\mathrm{old}}))}\right\}, \tag{3}$$

where $\mathbf{Y}_{\mathrm{exp}}$ is the array of synthetic cluster concentration data, and $\mathrm{p}(\mathbf{Y}_{\mathrm{exp}}|\boldsymbol{\theta}_{\mathrm{old}})$, $p(\mathbf{Y}_{exp}|\boldsymbol{\theta}_{new})$ denote the likelihood (conditional) probabilities for the old and new parameter values, respectively. These likelihood probabilities quantify how closely the kinetic model with parameters $\boldsymbol{\theta}$ reproduce the data, as they depend on the sum of squared residuals (see Eqs. ?? and ??) between the given data and the concentrations obtained from the ACDC and VODE simulations with parameters $\boldsymbol{\theta}\boldsymbol{\theta}_{old}$ and $\boldsymbol{\theta}_{new}$, respectively. This relationship is explained further in Appendix A1.

In DRAM we allow for partial modification of the proposed parameters (the "delayed rejection" "delayed rejection" component of DRAM). This second stage of sampling improves the computational time needed to obtain an estimate for $\boldsymbol{\theta}$; it is

performed as follows. If the proposed $\boldsymbol{\theta}_{\rm new}$ is rejected, a nearby proposal is created, $\boldsymbol{\theta}_{\rm new2}$. We accept this second proposal keeping in mind the rejection probability of the first, according to

$$\mathrm{p}_{\rm acc2} = \min\left\{1, \frac{\mathrm{p}(\mathbf{Y}_{\rm exp}|\boldsymbol{\theta}_{\rm new})\mathrm{p}(\mathbf{Y}_{\rm exp}|\boldsymbol{\theta}_{\rm new},\boldsymbol{\theta}_{\rm new2})[1-\mathrm{p}_{\rm acc}(\boldsymbol{\theta}_{\rm new},\boldsymbol{\theta}_{\rm new2})]}{\mathrm{p}(\mathbf{Y}_{\rm exp}|\boldsymbol{\theta}_{\rm old})\mathrm{p}(\mathbf{Y}_{\rm exp}|\boldsymbol{\theta}_{\rm old},\boldsymbol{\theta}_{\rm new})[1-\mathrm{p}_{\rm acc}(\boldsymbol{\theta}_{\rm old},\boldsymbol{\theta}_{\rm new})]}\right\}. \tag{4}$$

At the start of the MCMC simulations, the proposal covariances for both stages are initialized using arbitrary diagonal matrices with equal variances. It is assumed that the proposals of the form $\mathrm{p}(\mathbf{Y}_{\rm exp}|\cdot)$ and $\mathrm{p}(\mathbf{Y}_{\rm exp}|\cdot,\cdot)$ are Gaussian. They are updated at each successive iteration of the MCMC algorithm to improve the mixing of the chains.

The first-stage proposal covariance is recomputed via the Adaptive Metropolis (AM) procedure ~~(see ?)~~(?). Let d be the dimension of the parameter space, and $\{\mathbf{X_0},\ldots,\mathbf{X_n}\} \subset \mathbb{R}^d$ be a set of d-dimensional vectors containing the sampled values of free parameters. Then the first-stage proposal is centred at the current position of the Markov chain $\mathbf{X_n}$, whereas the corresponding proposal covariance $\mathbf{C}^1_{\rm n}$ is updated using the path of the previously sampled MCMC chain:

$$\mathbf{C}^1_{\rm n} = \begin{cases} \mathbf{C}_0, & n \le n_0 \\ \mathrm{s_d Cov}(\mathbf{X_0},\ldots,\mathbf{X_{n-1}}), & \mathrm{n} > \mathrm{n}_0, \end{cases} \tag{5}$$

where C_0 is the initial covariance assigned at the beginning of the MCMC runs, n_0 stands for the length of the initial non-adaptation period, $s_d = 2.4/d$ is the scaling parameter, and $\mathrm{Cov}(\mathbf{X_0},\ldots,\mathbf{X_{n-1}})$ is the empirical covariance matrix for the vectors $\mathbf{X_0},\ldots,\mathbf{X_{n-1}}$:

$$\mathrm{Cov}(\mathbf{X_0},\ldots,\mathbf{X_{n-1}}) = \frac{1}{\mathrm{n}-1}\Big(\sum_{\mathrm{i}=0}^{\mathrm{n}-1}\mathbf{X}_{\mathrm{i}}\mathbf{X}_{\mathrm{i}}^{\mathrm{T}} - \mathrm{n}\overline{\mathbf{X}}_{\mathrm{n}-1}\overline{\mathbf{X}}_{\mathrm{n}-1}^{\mathrm{T}},\Big), \tag{6}$$

where ~~$\overline{\mathbf{X}}^T_{n-1} = \frac{1}{n}\sum_{i=0}^{n-1} X_i$~~ $\overline{\mathbf{X}}^T_{n-1} = \frac{1}{n}\sum_{i=0}^{n-1}\mathbf{X}_i$ and $\mathbf{X}_i \in \mathbb{R}^d$ are column vectors. In our study and all runs therein, we set n_0 to be 100 iterations.

Simultaneously, the second-stage proposal covariance is computed as a scaled version of the first-stage proposal covariance:

$$\mathbf{C}^2_{\rm n} = \gamma\mathbf{C}^1_{\rm n}, \tag{7}$$

with the scaling factor $\gamma = 5$ borrowed from **?**. This value was chosen to increase the acceptance at the second stage.

Then, if both $\boldsymbol{\theta}_{old}$ and $\boldsymbol{\theta}_{new}$ are rejected at this stage, a new parameter candidate is sampled and the process is repeated. If the parameter candidate is accepted, the Markov chain is advanced one step and sampling as above is repeated. The process stops once the chain length is exhausted.

~~Further, observe that the sampled parametersof 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~~Parameter estimation is conducted using the **'mcmcstat'** toolbox implemented for FORTRAN (**??**). See the description and the examples of usage on the web page helios.fmi.fi/~lainema/.

2.2.3 Overview of the MCMC runs

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 Tabs. **??**-**??**).

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.* However, we assume that the evaporation rates with orders of magnitude less than $10^{-10}\mathrm{s}^{-1}$ are irrelevant in practise, since such an evaporation event is highly improbable, and it is very likely that instead the cluster will grow further by collisions. Similarly, when the evaporation rate is of the order of magnitude more than $10^{+10}\mathrm{s}^{-1}$, it is reasonable to expect that the cluster will most certainly evaporate before it has a chance to grow further. With these assumptions, the prior distribution of the evaporation rates spans over several orders of magnitude, and the base 10 logarithm of evaporation rates was sampled from the range of -12 to 12.

Next we justify the limits selected for data setting 2, where we sample thermodynamic parameters. For the formation enthalpies an upper limit of 0 kcal/mol is chosen by the fact that a positive ΔH would mean an absence of attractive interactions in the molecular cluster, which is physically incorrect for polar, H-bonding molecules such as H_2SO_4 and NH_3. For the lower limit (-400 kcal/mol) we mean that on average each H_2SO_4 is bound substantially stronger than in the $HSO_4^- * H_2SO_4$ cluster, for which the most recent computational studies indicate a binding enthalpy roughly around -40 kcal/mol, (**??**). Another motivation for the prior distribution selected for the cluster formation enthalpies comes from the fact that the largest cluster included into the system has 5 H_2SO_4 and 5 NH_3, so 10 molecules, and -400 kcal/mol would give an enthalpy of -40 kcal/mol per molecule, which 1) corresponds to the strongest known cluster in the system and 2) which implies that the evaporation rate is zero for all purposes of measurement (**?**).

Next, we set the upper limit for the formation entropies to 0 cal/K/mol, since molecule clustering must have a negative ΔH,as the number of gas molecules is reduced (and translational and rotational degrees of freedom are converted into much more constrained vibrational degrees of freedom). For the lower limit of -400 cal/K/mol, we state that the typical per-molecule ΔS for clustering is around -30 cal/K/mol, with a typical variation of up to +-10 cal/K/mol (**?**). So for the largest clusters the upper limit corresponds to a per-molecule ΔS of -40 cal/Kmol. In this situation, all the new vibrational degrees of freedom formed in the product clusters are quite rigid, i.e. have very low entropy (**?**).

An outline of the ~~above~~ sampling procedure is illustrated in Figure **??** below.

We next explicitly describe what synthetic data ($\mathbf{Y}_{exp}$~~and~~) and parameters ($\boldsymbol{\theta}$) which give the acceptance probability in **??** represent in the two study cases.

In the first study, the free parameters $\boldsymbol{\theta}$ represent the evaporation rates. The data $\mathbf{Y}_{exp}$ is either the time-independent steady-state or transient cluster concentrations measured at temperature 278 K.

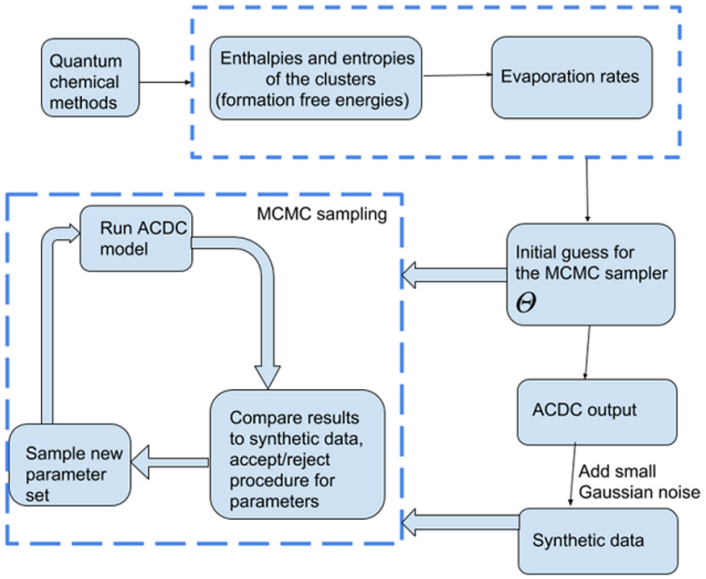


Figure 1. Schematic representation of the study methods.

In the second study, we use Eq. **??** and **??** to express the evaporation rates as functions of thermodynamic data, parametrized by temperature:

$$\gamma_{\mathrm{i+j\to i,j}} = \mathrm{f}(\mathrm{T}, \{\Delta \mathrm{H_k}, \Delta \mathrm{S_k}\}_{\mathrm{k}\in\{\mathrm{i+j,i,j}\}}). \tag{8}$$

In Eq. **??**, we set $T = 278$ K or $T = 292$ K. We emphasize that the rates $\gamma_{i+j\to i,j}$ now depend on temperature and six parameters: the cluster formation enthalpy $\Delta \mathrm{H_{i+j}}$ and entropy $\Delta \mathrm{S_{i+j}}$ of the evaporating cluster $i+j$, and the formation enthalpies $\Delta \mathrm{H_i}, \Delta \mathrm{H_j}$ and entropies $\Delta \mathrm{S_i}, \Delta \mathrm{S_j}$ of the clusters i and j respectively. In this setting $\boldsymbol{\theta}$ represents the array of quantities $\Delta \mathrm{H_{i+j}}$, $\Delta \mathrm{S_{i+j}}$, $\Delta \mathrm{H_i}$, $\Delta \mathrm{H_j}$, $\Delta \mathrm{S_i}$, $\Delta \mathrm{S_j}$ with $i+j \in \{1, 2, \ldots, 16\}$.

At either temperature $\mathrm{T} = 278$ K or $\mathrm{T} = 292$ K, the smaller clusters for certain combinations of ammonia and sulphuric acid may arise from the evaporation of several larger clusters. This implies that several of the pairs $\Delta \mathrm{H_i}, \Delta \mathrm{S_i}$ appear in expression **??** for the evaporation rates of different cluster types. Additionally, the Gibbs formation free energies of monomers are fixed to be zero, and their associated enthalpies and entropies do not vary in our simulations. This imposes additional constraints on possible parameter values. One can calculate that of the 39 evaporations that are involved in the dynamics of the neutral cluster system under consideration, only 28 distinct entropy and enthalpy values appear. Consequently, in this case the number of free parameters has been reduced from 39 to 28. This information is summarized in Table **??**. Moreover, from this table one can see that the entropy and enthalpy values lie within two orders of magnitude. This feature of the cluster formation entropies and enthalpies has the effect of reducing the *stiffness* of the differential system in **??** (computed via ACDC) which allows for easier integration via VODE.

For the setting above, the data $\mathbf{Y}_{exp}$ are the time-independent steady-state cluster concentrations measured at temperature 278 K or 292 K. We note that several experiments conducted at different temperatures are needed to obtain state information concerning the specific evaporation rate associated with each temperature level ~~(?)~~(?). In this work we consider two temperatures, which is one such minimal configuration that contains information sufficient for determination of thermodynamic data. Similar approaches were applied for the inverse problem of chemical kinetics modelled by the Arrhenius equation, where chemical reaction rates are temperature dependent ~~(?)~~ (?).

Note that to create a reliable sample from the underlying parameter distribution, the length of the MCMC chain must be "large enough" in an appropriate sense (??), that is, many different parameter combinations must be tested. We remark here that in both our studies, the MCMC chain length typically comprised of 3 million samples. The MCMC acceptance probabilities (defined below) in each of the cases were about 88.0%, which is a typical level of acceptance since the "forward" ACDC model (in which the evaporation and collision rates are known) is deterministic.

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. Specifically, the sampled parameters 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.

3 RESULTS AND DISCUSSION

3.1 Identification of the evaporation rate coefficients from steady-state data

First, we generate synthetic steady-state data by the method in Section **??**, for varying initial ammonia monomer concentrations, previously summarized in Table **??**; the sulphuric acid monomer is supplied to the system at a constant rate comprising 6.3×10^4 s^{-1} at the temperature $\mathrm{T} = 278$ K. As an output, we obtain the concentrations for all cluster types considered (listed earlier in Table **??**), measured when the system has attained the steady-state. A graphical representation of the data set is given above in Figure **??**~~-??~~.

~~Steady-state cluster concentrations for the clusters containing sulphuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for $[\mathrm{NH_3}]$ concentrations comprising 200 ppt at temperature T=278 K. The concentrations have been amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. The source of sulphuric acid monomers is $[\mathrm{H_2SO_4}] = 6.3 \times 10^4\ \mathrm{s}^{-1}$.~~

Next, from the steady-state data we determine the base 10 logarithms of the evaporation rate coefficients. Since the noise added to cluster concentrations results in a random bias towards an increase (or decrease) from the original values produced from the ACDC, the estimates of parameters derived from synthetic data are likely to be biased. In order to average the effects attributed to the random bias, we generated 3 sets of synthetic data by adding random increments to original concentration measurements. Utilizing these data sets, three independent MCMC runs were conducted, each run containing 3 million parameter samples. An example of one of the sampled chains is depicted in ~~Figures~~ Figs. **??**-**??**. We omit the initial one million samples

~~Steady-state cluster concentrations for the clusters containing sulphuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for $[NH_3]$ concentrations comprising 5 ppt at temperature T=278 K. The concentrations have been amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. The source of sulphuric acid monomers is $[H_2SO_4] = 6.3 \times 10^4$ s^{-1}.~~

~~Steady-state cluster concentrations for the clusters containing sulphuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for $[NH_3]$ concentrations comprising 35 ppt at temperature T=278 K. The concentrations have been amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. The source of sulphuric acid monomers is $[H_2SO_4] = 6.3 \times 10^4$ s^{-1}.~~

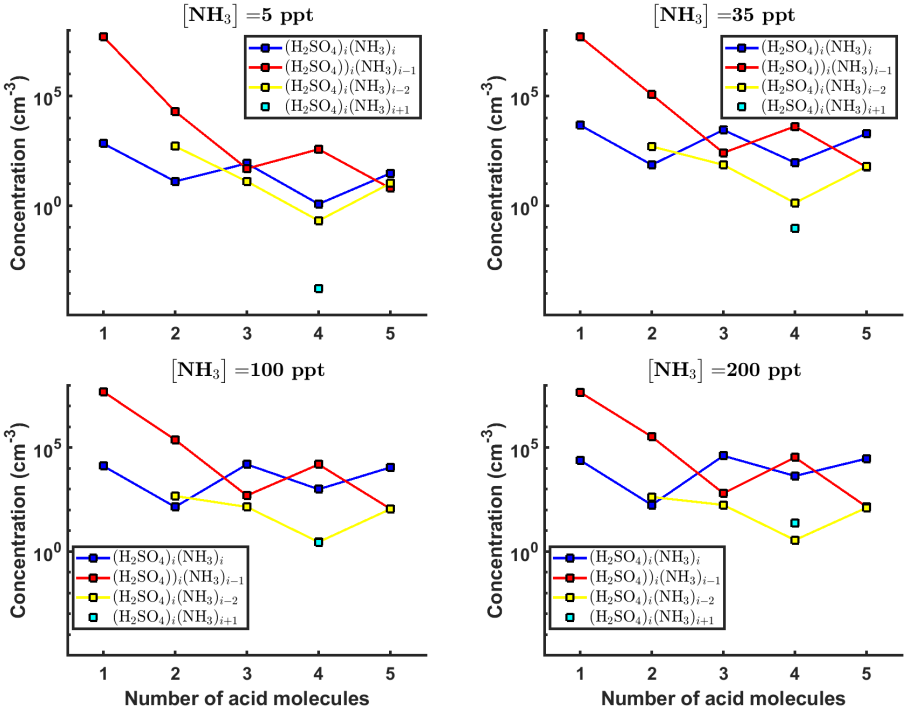


Figure 2. Steady-state cluster concentrations for the clusters containing sulphuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for $[NH_3]$ concentrations comprising (a) 5 ppt, (b) 35 ppt, (c) 100 ppt and (d) 200 ppt at temperature T=278 K. The concentrations have been amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. The source of sulphuric acid monomers is $[H_2SO_4] = 6.3 \times 10^4$ s^{-1} in each of the simulations.

and plot the stationary[3] parts of the chains. As we observe from the plots in ~~Figures~~ Figs. **??**-**??**, all the parameter chains ~~feature~~ for the evaporation rates have values bounded above by an upper limit which differs for different evaporation rates. However, only 15 out of 39 evaporation rates are limited from below (see subfigures labelled 1-5, 7, 10, 12, 16, 18, 22, 27, 31, 33 and 35 in ~~Figures~~ Figs. **??**-**??**). This subset of evaporation parameters is comprised of the evaporation rates of monomers, with the

[3]Here stationary means that the probability of transitioning from the current state at position j to the new state at position $j+1$ is independent of j.

exception of monomer evaporation rates for: H_2SO_4 from $(H_2SO_4)_5(NH_3)_4$ and $(H_2SO_4)_5(NH_3)_5$, and the evaporation rate of NH_3 from ~~$(H_2SO_4))_5(NH_3)_5$~~$(H_2SO_4)_5(NH_3)_5$. These excluded parameters correspond to the evaporations of monomers from the largest and most stable clusters. Note that the estimated lower limits of monomer evaporations from all the clusters except for the most stable ones are far above the 10^{-10} s^{-1} as defined for complete growth.

For each evaporation parameter, we calculate the one dimensional (that is, depending only on the evaporation rate) marginal posterior distribution as the position-wise average of the stationary parts of the three sampled chains. This procedure is needed to average the bias originating from random noise. The resulting distributions are given in ~~Figures~~ Figs. **??**-**??**. We use the maximum (also called the mode in the statistics literature) of the posterior marginal distribution function as our parameter estimate in the case when the marginal posterior distributions have precisely one maximum value. In the cases where we have multiple estimators, we provide a range for the evaporation rate values.

All the evaporation rates larger than 10^{-3} s^{-1} are well-identified (see subfigures labelled 1, 2, 4, 5, 7, 10, 12, 16, 18, 22, 27, 31 and 35 in ~~Figures~~ Figs. **??**- **??**), in the sense that their estimated variances are well within our accepted error range of less then one order of magnitude. The estimates for the remaining evaporation rates can take values within ranges spanning several orders of magnitude and are thus uncertain. Also, notice that most of the marginal posterior distributions are non-uniform, except for the evaporation rate of $(H_2SO_4)_2(NH_3)_2$ from $(H_2SO_4)_5(NH_3)_5$. In five cases (refer to subfigures labelled 6, 21, 28, 32 and 36 in ~~Figures~~ Figs. **??**- **??**), the estimated parameter values are not unique; that is the marginal posterior distributions feature multiple modes. The results of our parameter ~~identification~~ estimation are summarized in ~~Tables~~ Tabs. **??**- **??** and in subfigures labelled (a) and (b) in Figure **??**.

The pairwise marginal posterior distributions for the estimated evaporation rates are illustrated in ~~Figures~~ Figs. **??**-**??**. From these plots one can see that the majority of parameters are not correlated. However, the evaporation of monomers from ~~$(H_2SO_2)_5NH_3$~~$(H_2SO_4)_5NH_3$, $(H_2SO_4)_3(NH_3)_2$ and $(H_2SO_4)_5(NH_3)_4$ display non-linear inverse correlations. This implies that either H_2SO_4 rarely evaporates (at the rate less then 10^{-4} s^{-1}) and that NH_3 evaporates often, or the evaporation rates of H_2SO_4 and NH_3 are of comparable magnitude in these cases. Additionally, it can be seen from the pairwise posteriors that most of the estimated parameters are highly uncertain. Therefore, we conclude that in the situation where we determine parameters from the synthetic steady-state data, parameter ~~identification~~ estimation is not unique.

From a mathematical perspective, the existence of multiple distinct parameter estimates indicates that the problem of recovering evaporation rates from the synthetic steady-state concentration data is ill-posed. In these situations, one seeks to regularize the problem; that is, add more data or information to the model to reduce the number of possible estimates.

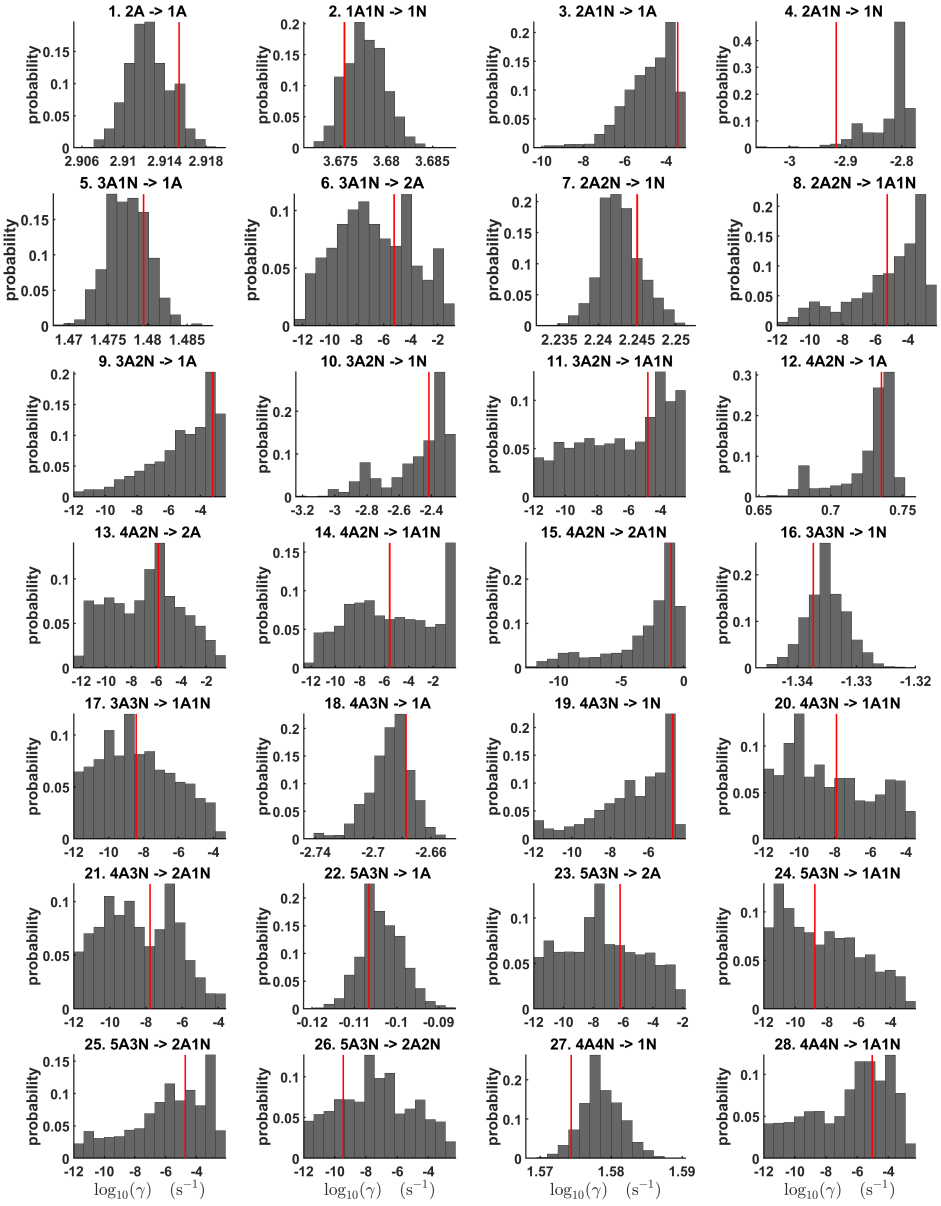


Figure 3. One-dimensional marginal posterior distributions (for parameter indexes ranging from 1 to 28) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Table 3. Domain limitations for two data settings under consideration imposed to exclude non-physical parameters in parameter estimation procedure.

Data settings	Estimated parameters	Minimal value	Maximal value
Data setting 1	Base 10 logarithms of evaporation rates (in s^{-1})	-12	12
Data setting 2	Cluster formation enthalpies (kcal mol^{-1}) and	-400	0
	entropies (cal K^{-1} mol^{-1})	-400	0

Table 4. Additional domain limitations for the data setting 2 from Table **??** (estimation of thermodynamic data), where the cluster formation enthalpy of the i-th cluster is denoted by ΔH_i and the symbols A and N stand for ammonia and sulphuric acid, respectively.

$\Delta H_{2A} > \Delta H_{2A1N}$	$\Delta H_{3A2N} > \Delta H_{4A2N}$
$\Delta H_{1A1N} > \Delta H_{2A1N}$	$\Delta H_{4A2N} > \Delta H_{4A3N}$
$\Delta H_{2A1N} > \Delta H_{3A1N}$	$\Delta H_{4A3N} > \Delta H_{4A4N}$
$\Delta H_{2A2N} > \Delta H_{3A2N}$	$\Delta H_{4A4N} > \Delta H_{5A5N}$
$\Delta H_{3A1N} > \Delta H_{3A2N}$	$\Delta H_{4A4N} > \Delta H_{4A5N}$

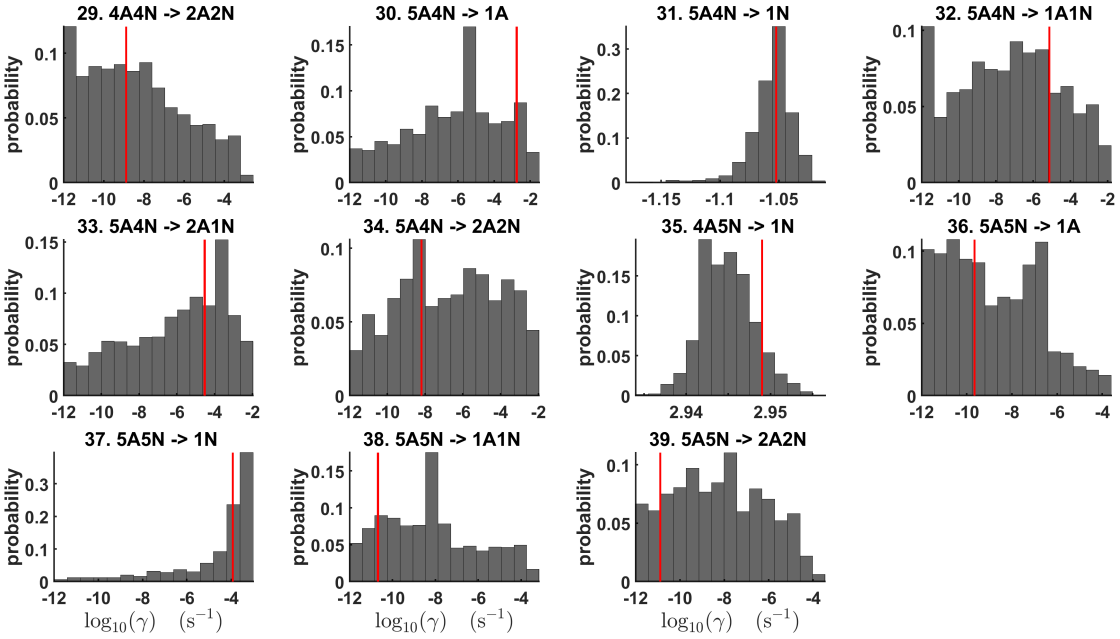


Figure 4. One-dimensional marginal posterior distributions (for parameter indexes ranging from 29 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

3.2 Identification of the evaporation rate coefficients from transient data

~~In this section and next, we consider two methods of regularizing our problem.~~ First, we ~~change~~ extend the synthetic measurement data from steady state concentrations to transient concentrations. ~~We then conduct analogous MCMC runs (as described in Section 2 using this extended data set~~ 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 **??**.), 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 ~~Figures ??-~~**??**.

From this transient cluster concentration data set, we then conduct analogous MCMC runs (as described in Section 2.2). As in the steady-state setting, we conduct three independent MCMC runs to determine the base 10 logarithms of the evaporation rates. One of these runs is presented in ~~Figures~~ Figs. **??**-**??**. Again, we omit the first one million samples, which are the samples before the chains have obtained their stationary distributions.

It is shown in ~~Figures~~ Figs. **??-??**, that all the chains have the upper limits. Most of the chains are bounded from below, with five exceptions. Specifically, the evaporation rates of ~~$(H_2SO_2)_2(NH_3)_2$ from $(H_2SO_2)_4(NH_3)_4$ and $(H_2SO_2)_5(NH_3)_3$~~ $(H_2SO_4)_2(NH_3)$ from $(H_2SO_4)_4(NH_3)_4$ and $(H_2SO_4)_5(NH_3)_3$, the evaporation rates of H_2SO_4, ~~$H_2SO_2NH_3$ and $(H_2SO_2)_2(NH_3)_2$ from $(H_2SO_2)_5(NH_3)_5$~~ $H_2SO_4NH_3$ and $(H_2SO_4)_2(NH_3)_2$ from $(H_2SO_4)_5(NH_3)_5$ have arbitrarily large magnitude.

We examine the one-dimensional marginal posterior distributions for the estimated parameters in ~~Figures~~ Figs. **??-??**. From these plots, one sees that most of the estimates are close to the baseline values used for generation of the synthetic data. However, the estimated evaporation parameters still feature substantial uncertainties, as their marginal posterior distributions span several orders of magnitude (see subfigures 6, 8, 9, 11, 13, 14, 17, 21, 23-26, 30, 32-34, 37-39 in ~~Figures~~ Figs. **??-??**). Three parameters (subfigures 20, 29 and 36 in ~~Figures~~ Figs. **??-??**) have multimodal marginal posterior distributions. We also note that the evaporation rate of $(H_2SO_4)_2(NH_3)_2$ from $(H_2SO_4)_5(NH_3)_3$ (which corresponds to subfigure 26) has a uniform posterior distribution. Further, we can only specify that the upper limits for the evaporation rates depicted in subfigures 20 and 36 are less than 1.96×10^{-5} s^{-1}. However, given the reliable upper estimates, the evaporation processes $(H_2SO_4)_4(NH_3)_3 \rightarrow (H_2SO_4)_4(NH_3)_2 + NH_3$ and $(H_2SO_4)_5(NH_3)_5 \rightarrow (H_2SO_4)_4(NH_3)_5 + H_2SO_4$ can be neglected, as they are relatively slow when compared with the other competing processes.

Pairwise marginal posterior distributions for the evaporation rates are plotted in ~~Figures~~ Figs. **??-??**. Notice that the evaporation rates of monomers for ~~clusters $(H_2SO_2)_2$~~ the cluster $(H_2SO_4)_2NH_3$ display strong inverse linear relationship, which is indicated by the pairwise marginal posterior distribution of the coefficients $(H_2SO_4)_2NH_3 \rightarrow (H_2SO_4)_2 + NH_3$ and ~~$(H_2SO_2)_2NH_3$ display inverse linear correlations.~~ $(H_2SO_4)_2NH_3 \rightarrow H_2SO_4NH_3 + H_2SO_4$, (see Figure **??**). Also, the estimated rate coefficients $(H_2SO_4)_2 \rightarrow H_2SO_4 + H_2SO_4$ and $H_2SO_4NH_3 \rightarrow H_2SO_4 + NH_3$ exhibit linear correlation. Additionally, the uncertainties in all the correlated parameters are relatively small (less then an order of magnitude). We also remark that from these plots one can see that most of the evaporation rates do not display any substantial correlations.

In ~~Tables~~ Tabs. **??-??** we summarize the results of parameter ~~identification~~ estimation for the above-discussed two data settings. Note that the estimated upper limits for some of the small evaporation rates (less than 10^{-5} s^{-1}) determined from the steady-state data can be as large as 1.55×10^{-2} s^{-1}. This is a poor estimate, since the uncertainties in the synthetic data are small. For example, see the results for parameters shown in subfigures 32 and 34 of Figure **??**. In these cases the identification has improved when we extended the data set with time-dependent measurements. Overall one observes that the transient data enabled us to determine the lower bounds for most of the parameters, with the exception of those parameters shown in subfigures numbered 26 and 29. Moreover, the additional time dependent data enabled us to reduce the uncertainties in the estimates of parameters in subfigures 15, 19 and 37. As a result, with the aid of time-dependent data we have improved the estimates of minimal and maximal values for the evaporation rate parameters (see comparison of the 95 % confidence intervals plotted in Figure **??**).

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

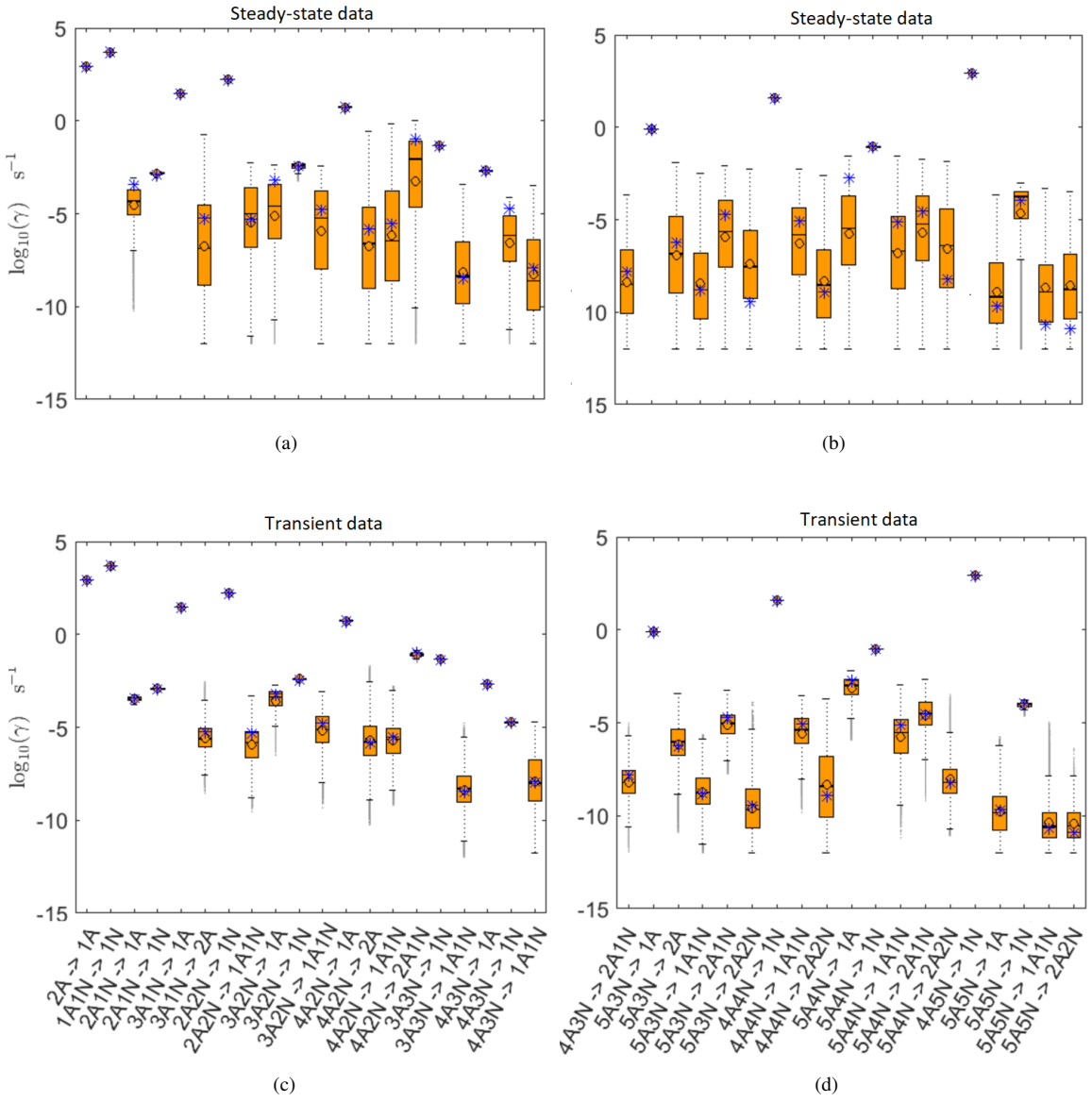


Figure 5. Comparison of 95 % confidence intervals (orange box plots) of base 10 logarithms of the evaporation rates determined from (a)-(b) steady-state and (c)-(d) time-dependent synthetic data measured at temperature 278 K. In reactions "A" stands for H_2SO_4 and "N" for NH_3. Here blue asterisks denote the baseline values used for creating the synthetic data (borrowed from **?**). Black circle and horizontal line markers indicate the mode and the mean value of the distribution, respectively.

has attained the steady state together with the time-step concentration data measured from the starting point to the end of the ACDC simulation.

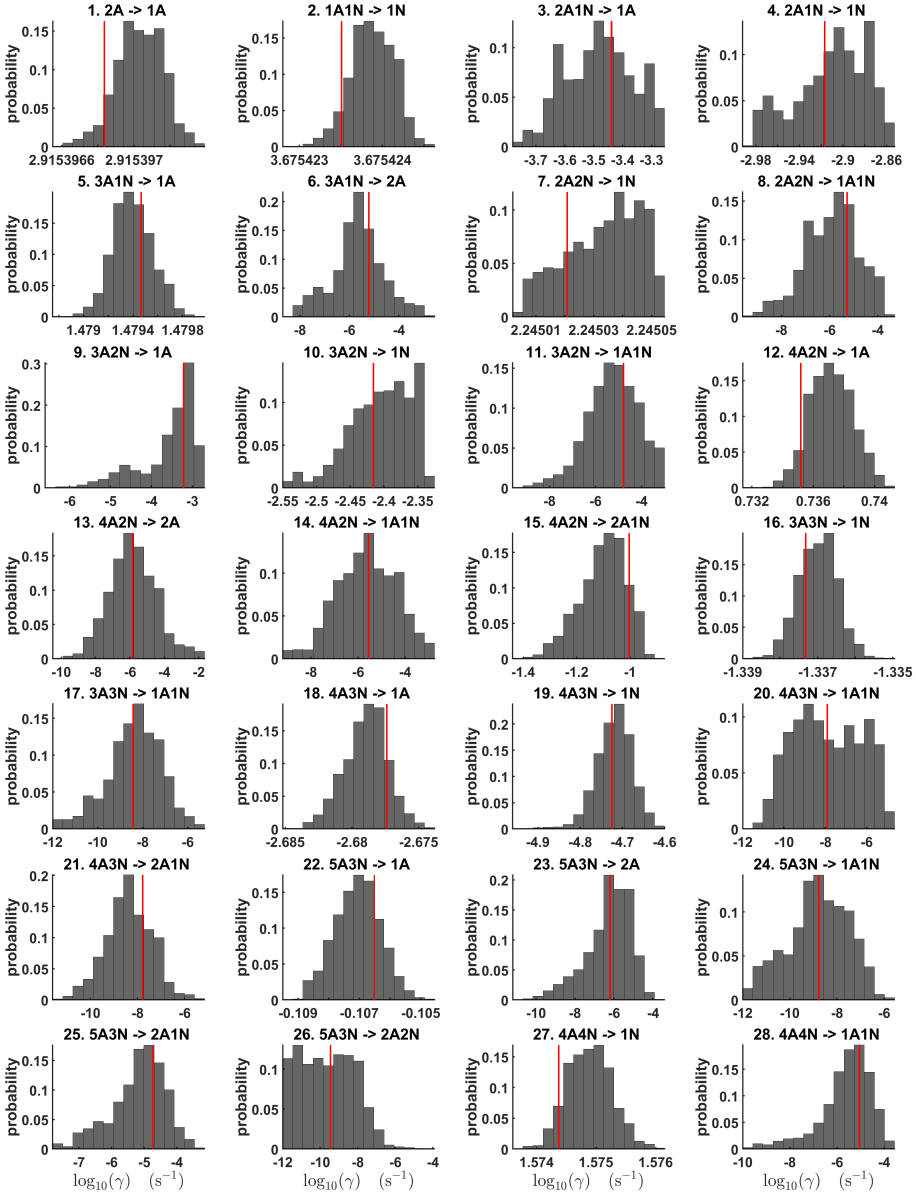


Figure 6. One-dimensional marginal posterior distributions (for parameter indexes ranging from 1 to 28) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

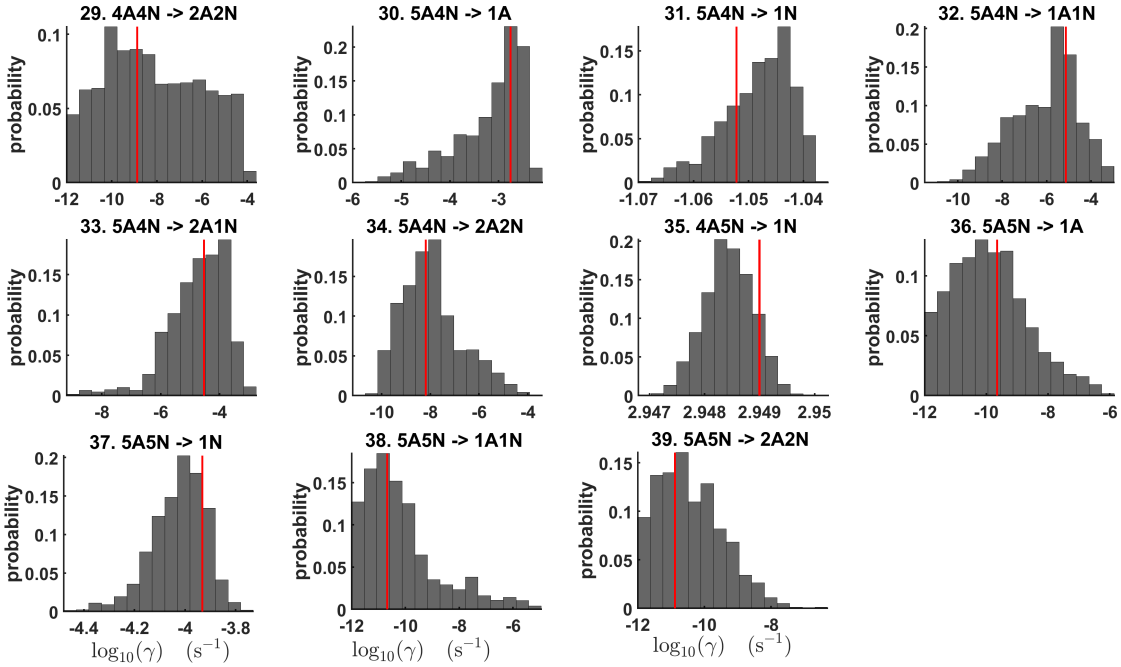


Figure 7. One-dimensional marginal posterior distributions (for parameter indexes ranging from 29 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

3.3 Estimating thermodynamic data from steady-state concentration measurements

In this section we describe another method for regularizing our problem of estimating evaporation rates from steady-state concentration data. We will determine the cluster formation enthalpies and entropies from two sets of synthetic, steady-state cluster concentrations, now measured at two temperatures: 278 and 292 K. This data set is plotted in ~~Figures~~ Figs. **??** and **??** for 278 K and 292 K, respectively.

We will demonstrate that reparameterization (in terms of thermodynamic data) plus the extended data set transforms our parameter ~~identification~~ estimation problem from an ill-posed problem to a well-posed one. We use synthetic steady-state cluster concentrations generated for two temperatures to recover the thermodynamic parameters. This is done to improve the identification by using the temperature dependence of the Gibbs free energies (and the evaporation rates).

For each temperature choice, we use the methods described in Section 2 to obtain synthetic steady-state cluster concentration data. We summarize this data in Table **??**; the data sets are plotted in Figure **??** for 278 K and **??** for 292 K. Three MCMC runs were conducted to average the bias attributed to random noise added to the data, as discussed in the previous section. An example of one of the sampled chains is illustrated in Figure **??**. It can be seen that all the chains are bounded, with the

exception of the formation enthalpy and entropy of the biggest cluster (~~$(H_2SO_2)_5(NH_3)_5$~~$(H_2SO_4)_5(NH_3)_5$).

Next we consider the one-dimensional (depending on the particular cluster formation entropy or enthalpy parameters) marginal posterior distributions of free parameters built from the stationary parts of the three sampled chains merged together, see Figure **??**. It can be seen that for all the clusters except ~~$(H_2SO_2)_5(NH_3)_5$ the~~ $(H_2SO_4)_5(NH_3)_5$ the variance for the estimated formation enthalpies ~~vary at most by 1 kcal mol^{-1}~~are less than 0.46 kcal mol^{-1}, while the ~~variance for the formation entropies is less than 1~~ estimated formation entropies vary at most by 5.4 cal ~~K^{-1} mol^{-1}.~~ K^{-1}mol^{-1}. The estimated free parameters together with the baseline quantum chemistry-based values from **?** used for generation of the synthetic data are summarized in Table **??**.

Although the posterior distributions of sampled thermodynamic parameters for ~~$(H_2SO_2)_5(NH_3)_5$~~ $(H_2SO_4)_5(NH_3)_5$ feature higher uncertainties in comparison to the corresponding posterior distributions identified for the smaller clusters, the ~~Gibbs free energy of cluster formation for $(H_2SO_2)_5(NH_3)_5$~~evaporation rates for evaporations from $(H_2SO_4)_5(NH_3)_5$, as calculated from the aforementioned posterior distributions, ~~has low variance. This is due to the fact that formation enthalpies and entropies of the molecular clusters exhibit strong linear correlations, as we see from our MCMC simulations in Figure ?? and Figures ??-??. As a result, the evaporation rates of $(H_2SO_2)_5(NH_3)_5$ calculated from a posterior distribution of sampled thermodynamic parameters have low uncertainties, i.e., they vary within one order of magnitude, see Figure~~ have low variances, see Table **??**.

Notice that the evaporation rates for all the molecular clusters calculated from a posterior distribution of sampled thermodynamic parameters for the temperature 278 K are close to the baseline values from **?** used for generation of the synthetic data and their variances are less than one order of magnitude, see ~~Figures~~ Figs. **??**-**??**.

Additionally, strong correlations are observed between formation enthalpies (entropies) of the clusters containing same number of ammonia molecules larger then 2, except the case of ~~$(H_2SO_2)_5(NH_3)_5$~~$(H_2SO_4)_5(NH_3)_5$. Since our parameters are strongly correlated, we may alternatively consider just cluster formation enthalpies or the ratios of cluster formation entropies and enthalpies as our free parameters.

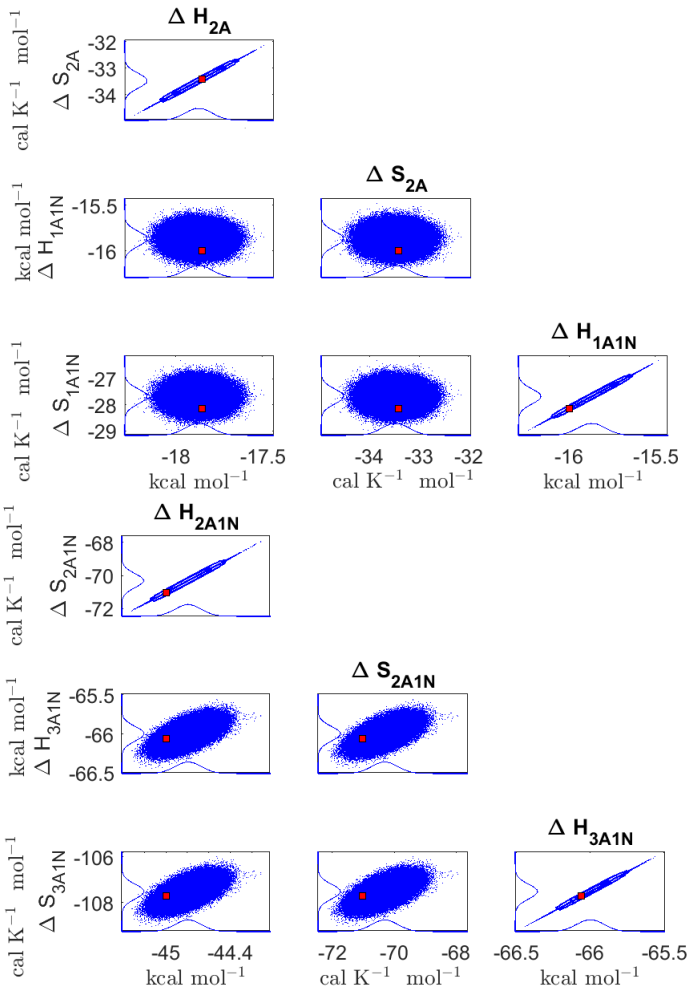


Figure 8. Pairwise marginal posterior distributions (for parameter indexes ranging from 1 to 8) of the cluster formation enthalpies and entropies determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

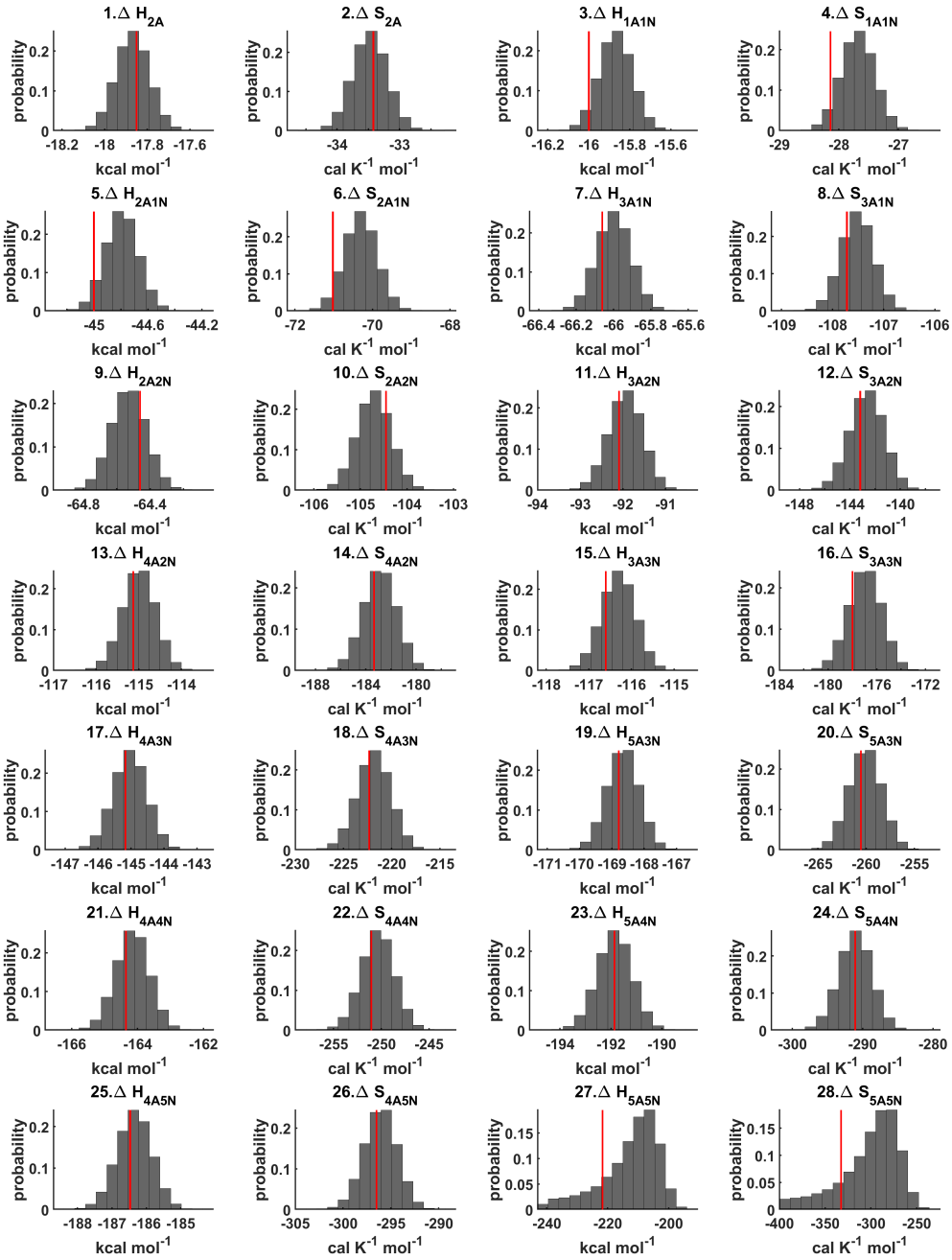


Figure 9. One-dimensional marginal posterior distributions of the cluster formation enthalpies (units given in kcal/mol) and entropies (units given in $\mathrm{cal\ K^{-1}\ mol^{-1}}$)) determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red lines denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote $\mathrm{H_2SO_4}$ and "$\mathrm{NH_3}$", correspondingly.

3.4 Comparison to previous evaporation rate determinations

The evaporation rates can be obtained either experimentally or computationally, when applying the Quantum Chemical (QC) methods, (**?**). Experimental detection was conducted from the measurements in a flow tube (**???**) and in the CLOUD chamber (**?????**). The summary of thermodynamic parameters obtained from different methods has previously been published in **?**. These parameters can be employed to calculated the evaporation rates at different temperatures.

In this study we determine the evaporation rates and thermodynamic data from measurements of cluster concentrations. Supplementary to the methodology presented in **?**, our first method enables to determine parameters from the time-dependent cluster concentrations measured before the system has attained the steady state. The transient data improved the estimates for all the evaporation rates.

In the second method we identify thermodynamic parameters from the steady-state cluster concentrations measured at two different temperatures. This approach is similar to **?**, but our model takes into account all the possible evaporation processes. In **?** the thermodynamic parameters had been determined from the New Particle Formation Rates (NPFs) measured at different temperatures. Instead of the NPFs, we employ the measurements of cluster concentrations. By so doing, we find the combination of data and fitted parameters which enables to determine the evaporation rates with the variances comprising less that one order of magnitude.

Although the transient data have improved the estimates, the temperature-dependent data have been demonstrated to yield the most accurate estimates of the evaporation rates, when we treat cluster formation enthalpies and entropies as free parameters.

3.5 Discussion and future work

The MCMC results are not specific for the simulation box considered in the present study, but rather general. This is supported by the fact that although the size of the system (the number of clusters included into simulations) has impact on the particle formation rates at high temperatures (> 278 K), the particle formation rates and cluster concentrations produced using different simulation boxes are qualitatively similar. Thus the changes of the ACDC outputs due to the difference in the simulation box does not change for MCMC parameter estimation results. In **?** is was shown that the 5x5 simulation box (which is used for generation of the synthetic data) produces reasonable results with a good agreement with the measurements obtained from the CLOUD chamber experiment. Additionally, the boundary conditions for the outgrowing clusters (the choice of the clusters that are considered as formed particles) has only minor influence on the simulation results, given that the simulated system of clusters is defined in a reasonable way (**?**).

In general, the accuracy of the MCMC results increases when we include additional data. In particular, including more concentration data measured at different ammonia concentrations will yield better estimates for the evaporation rates. The sensitivity of the estimates to the number of ammonia concentrations will be considered in the future work. In the present study we rather focus on the question which combination of estimated parameters and concentration data will produce an accurate estimates for the evaporation rate.

The data of steady-state concentration with two temperatures allowed us to apply two general principles of inverse problems/Bayesian estimation to the problem of estimating evaporation rates. First, the two temperature data set enabled us to reformulate the problem in a numerically effective way (in terms of enthalpy and entropy) that reduced the number of unknown parameters we sought to estimate. Second, the reformulated differential equation describing the time evolution of the concentrations was more numerically stable than the original expression (the stiffness of the equation was reduced in the reformulated form). This made our estimates for the rates less sensitive to small perturbations/errors.

In addition, the fact that the formation entropies and enthalpies were strongly correlated made them an effective parametrization. The strong inverse correlations have a physical explanation. Firstly, both formation enthalpy and entropy follow from the partition function of the molecular complex, and their functional forms are partly similar (**?**). Practically, if a cluster has really strong bonds between the molecules, then that means the formation enthalpy is very negative, and also the intermolecular vibrational frequencies corresponding in a broad sense to vibrations involving those bonds (note that these frequencies dominate the "variable part" of the formation entropy, as the entropy effect from the loss of translational and rotational degrees of freedom is almost a constant factor) are fairly high, meaning that the entropy loss in forming the cluster is large. So if the formation enthalpy is very negative so is also the formation entropy. Conversely, if the cluster is only quite weakly bound, the formation enthalpy is only slightly negative, and the intermolecular frequencies can be very low, leading to a less negative (though still negative of course) formation entropy (**?**).

Note that experimental data can differ from the synthetic data in the sense that they contain noise which originate from measurement instruments and uncertainties associated with experimental conditions (e.g., in CLOUD chamber experiments). Treating the noise inherent for experimental data will be the topic of our future studies.

4 Conclusions

We applied a Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) algorithm to identify cluster evaporation/fragmentation rates from known cluster distribution data and known cluster collision rates. We used Atmospheric Cluster Dynamic Code (ACDC) with quantum chemistry based evaporation rates to generate synthetic data for the purpose of validating the parameter ~~identification~~estimation.

First, we sought to determine the cluster evaporation rates from both steady-state and time-dependent cluster concentration data at one temperature. In this first scenario, we sought to determine the cluster evaporation rates from both steady-state and time-dependent cluster concentration data. Due to the mathematical stiffness of the ordinary differential equations describing the time evolution of the cluster concentrations, we were only able to identify a subset of the free parameters (evaporation rates) from the available data. This stiffness originates from the vastly different timescales of some of the key evaporation rates.

In the second scenario, we used only steady-state concentration data but for two different temperatures. We introduced a reparametrization expressing the evaporation rates in terms of cluster formation enthalpies and entropies, and temperature. This reduced the number of parameters we sought to identify. It also lessened the stiffness of the system, as the cluster formation enthalpies and entropies for our system have comparable orders of magnitude. We demonstrated that steady-state concentration data at two different temperatures could be used to determine all the unknown formation enthalpies and entropies, and thus the evaporation rates, to within acceptable accuracy.

The approach presented here can also be applied to infer evaporation rates from mass spectrometric measurements of molecular cluster concentrations. This naturally requires accounting for the process of charging neutral clusters, with its associated uncertainties. A clear conclusion of our proof-of-concept study is that steady-state data at different temperatures is more useful for determining evaporation rates than time-dependent data at a single temperature. Determining very low (below 10^{-5} s^{-1}) evaporation rates may also require additional measurements at low vapor concentrations, which naturally require longer timescales to reach a steady state.

Code availability. The code is available via GitHub repository: http://doi.org/10.5281/zenodo.3766925

Appendix A: Supplementary mathematical material

A1 Cluster kinematics

The kinetics of cluster formation is described by Becker-Döring equations ~~(see ?, ?),~~ (??), which model cluster birth and death which arises from collisions of the smaller clusters into larger ones and evaporations from the bigger clusters into smaller ones. Precisely, labelling the clusters by $i \in \{1, 2, \ldots, N\}$, the time derivative of the ith cluster concentration $\mathrm{Y_i}$ is governed by

$$\frac{\mathrm{dY_i}}{\mathrm{dt}} = \frac{1}{2}\sum_{\mathrm{j<i}} \beta_{\mathrm{i,(i-j)}}\mathrm{Y_i Y_{i-j}} + \sum_{\mathrm{j}} \gamma_{\mathrm{i+j\to i,j}}\mathrm{Y_{i+j}} - \sum_{\mathrm{j}} \beta_{\mathrm{i,j}}\mathrm{Y_i Y_j} - \frac{1}{2}\sum_{\mathrm{j<i}} \gamma_{\mathrm{i\to j,i-j}}\mathrm{Y_i} + \mathrm{Q_i} - \mathrm{S_i}, \tag{A1}$$

where $\beta_{i,j}$ is the collision coefficient of clusters i with j, and $\gamma_{i+j\to i,j}$ is the evaporation coefficient of cluster $i+j$ into clusters i and j, $\mathrm{Q_i}$ is an external source term of i, and $\mathrm{S_i}$ represents the total possible types of losses for the cluster of type i. These last two terms, which stand for external supply and destruction mechanisms, depend on the system under consideration.

We now specify the quantity and type of sinks and sources included in our studies. We assume that the concentration of ammonia monomers is constant, while sulphuric acid monomers are supplied to the system at a constant rate comprising $\mathrm{Q} = 6.3 \times 10^4\ \mathrm{cm^{-3}s^{-1}}$. This settings are selected to imitate the conditions inside of the CLOUD chamber, ~~(see ?, ?)~~(??). Further, we include wall losses arising from clusters sticking on the walls of the experimental chamber~~(see ?)~~, (?). These wall losses are parametrized by the size of the cluster

$$\mathrm{S_{wall,i}} = 10^{-12}/(2\mathrm{r_i} + 0.3 \times 10^{-9}) \quad \mathrm{s^{-1}}, \tag{A2}$$

where $\mathrm{r_i}$ is the mass radius of the cluster (in cm). From Eq. **??**, wall loss rates decrease with cluster size; in practise it also varies with respect to cluster position in the chamber and time. We neglect any uncertainties attributed to the wall losses. However, we do account for dilution losses, with size-independent value comprising $\mathrm{S_{dil,i}} = 9.6 \times 10^{-5}\mathrm{s^{-1}}$, which had previously been determined in the CLOUD chamber, ~~(see ?, ?)~~(??).

Let T denote the temperature of the system of molecular clusters. Using classical kinetic gas theory, the collision rates $\beta_{i,j}$ in Eq. **??** obey

$$\beta_{\mathrm{i,j}} = \sqrt{\mathrm{T}}\left(\frac{3}{4\pi}\right)^{1/6}\left[6\mathrm{k_B}\left(\frac{1}{\mathrm{m_i}} + \frac{1}{\mathrm{m_j}}\right)\right]^{1/2}\left(\mathrm{V_i^{1/3}} + \mathrm{V_j^{1/3}}\right)^2, \tag{A3}$$

where $\mathrm{m_i}$ and $\mathrm{V_i}$ are respectively the mass and volume of cluster i, and $\mathrm{k_B}$ is Boltzmann's constant. In this paper, we assume that the masses and volumes are temperature-independent.

The cluster evaporation rates $\gamma_{i+j\to i,j}$ in Eq. **??** are given by the expression

$$\gamma_{\mathrm{i+j\to i,j}} = \beta_{\mathrm{i,j}}\frac{\mathrm{P_{ref}}}{\mathrm{k_B T}}\exp\left(\frac{\Delta\mathrm{G_{i+j}} - \Delta\mathrm{G_i} - \Delta\mathrm{G_j}}{\mathrm{k_B T}}\right), \tag{A4}$$

where $\mathrm{P_{ref}}$ is the reference pressure and $\Delta\mathrm{G_i}$ is the Gibbs free energy of formation for cluster i. We may further describe the ith Gibbs free energy in terms of the cluster formation enthalpy $\Delta\mathrm{H_i}$ and entropy $\Delta\mathrm{S_i}$:

$$\Delta\mathrm{G_i} = \Delta\mathrm{H_i} - \mathrm{T}\Delta\mathrm{S_i}. \tag{A5}$$

We neglect here the weak temperature dependence of real cluster formation enthalpies and entropies.

A2 Likelihood, data and cost function

The likelihood of observing the data $\mathbf{Y}_{exp}$ given the parameter values $\boldsymbol{\theta}$ is

$$\mathrm{p}(\mathbf{Y}_{\mathrm{exp}}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\mathrm{n_{out}}/2}} \exp(-\frac{1}{2}\mathrm{F}(\boldsymbol{\theta})), \tag{A6}$$

where $\mathrm{n_{out}}$ is the number of measurements and $F(\boldsymbol{\theta})$ is the cost function. We elucidate the cost function below. In our first study in which simulations are conducted with time-dependent data, the number of measurements is $\mathrm{n_{out}} = 4 * (\mathrm{N_c} * \mathrm{N_t} + 1)$, where $N_c = 16$ is the number of cluster types whose concentrations are measured and $\mathrm{N_t} = 41$ is the number of time-step measurements available for each of the cluster types. As explained in Section 2.1, after each VODE integration, a convergence coefficient is computed from the steady-state cluster concentrations to ensure that the system has attained the steady-state.

In our first study, the parameter fit to the data was evaluated by the sum of squared residuals of the model outputs $\mathbf{Y}_{mod}$ and the measurements, $\mathbf{Y}_{exp}$. The *cost function* (sum of squared residuals) measures how far our model outputs are from the "true" experimental outputs. Precisely,

$$\mathrm{F}(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{N_c}} \sum_{\mathrm{j}=1}^{\mathrm{N_t}} \frac{(\mathrm{Y_{exp,i}}(\mathrm{t_j}) - \mathrm{Y_{mod,i}}(\boldsymbol{\theta}, \mathrm{t_j}))^2}{\sigma_{\mathrm{ji}}^2}. \tag{A7}$$

Since concentrations of molecular clusters span a large range (from 10^{-5} to 10^{9} particles per cm^3), we normalize the residuals by the measurement error variance σ_{ji}^2. Normalization in this way avoids overfitting to the larger concentration values. Note also that the error variance σ_{ji}^2 is matched separately for each cluster type and every time instance. We assume that the instrument is capable of detecting all the cluster types represented in the system at arbitrary small levels of concentration. This simplification was considered in order to illustrate the proposed approach.

When parameter estimation is conducted with steady-state cluster concentrations (as is considered in our second study), we use the following cost function:

$$\mathrm{F}(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{N_c}} \sum_{\mathrm{j}=1}^{\mathrm{N_T}} \frac{(\mathrm{Y_{exp,i}}(\mathrm{T_j}) - \mathrm{Y_{mod,i}}(\boldsymbol{\theta}, \mathrm{T_j}))^2}{\sigma_{\mathrm{ji}}^2}. \tag{A8}$$

Now $\mathrm{N_T} = 2$ denotes the number of steady state configurations at different *temperatures* (not times!) and $\mathrm{T_j}$ stands for the measured *temperature*. In this study, the number of measurements for the likelihood given by Eq. ~~??~~. ?? is $n_{\mathrm{out}} = 4 * (N_c * N_T + 1)$ (again $N_c = 16$ cluster types).

Appendix B: ~~Identification~~ Estimation of the evaporation rates from steady-state data

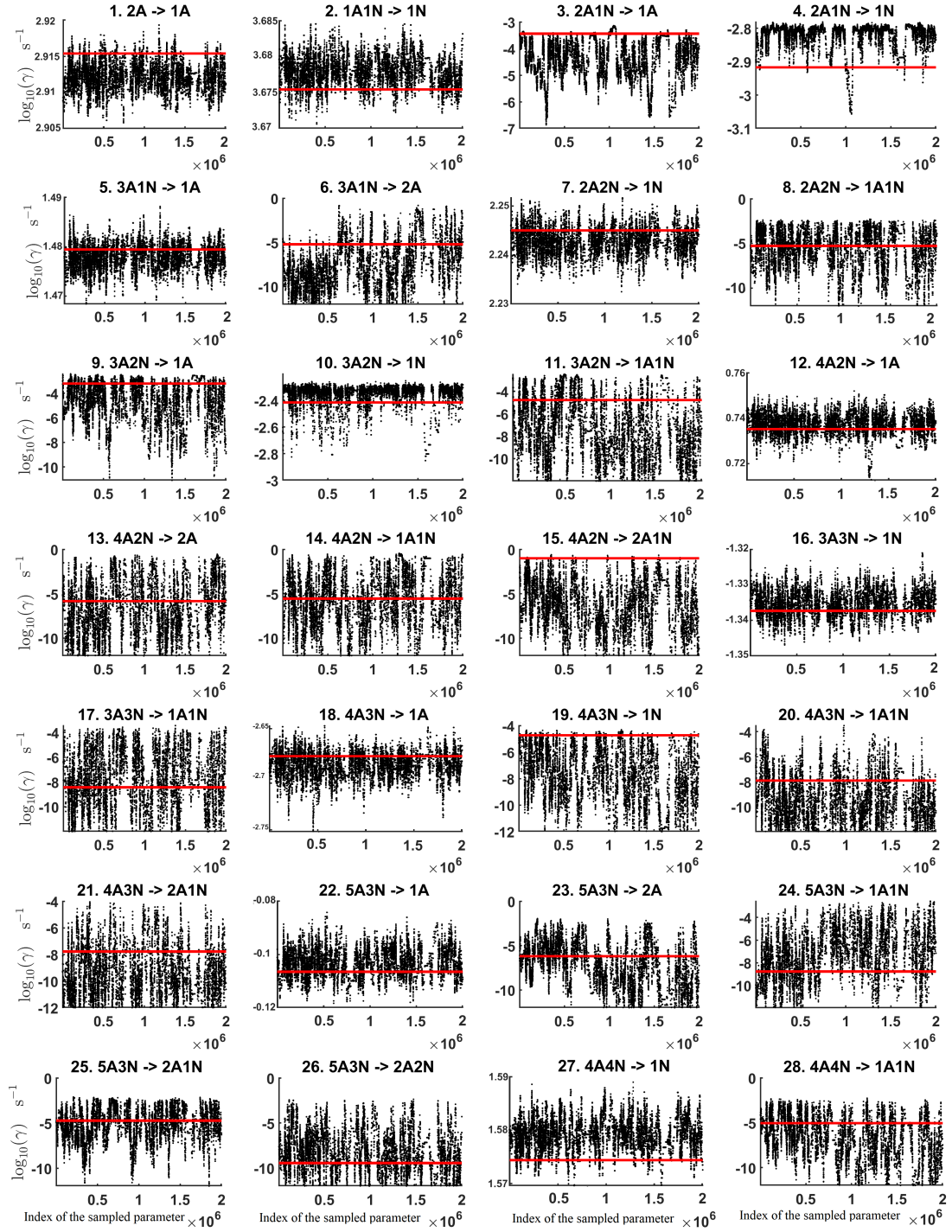


Figure B1. Parameter chains (for parameter indexes ranging from 1 to 28) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

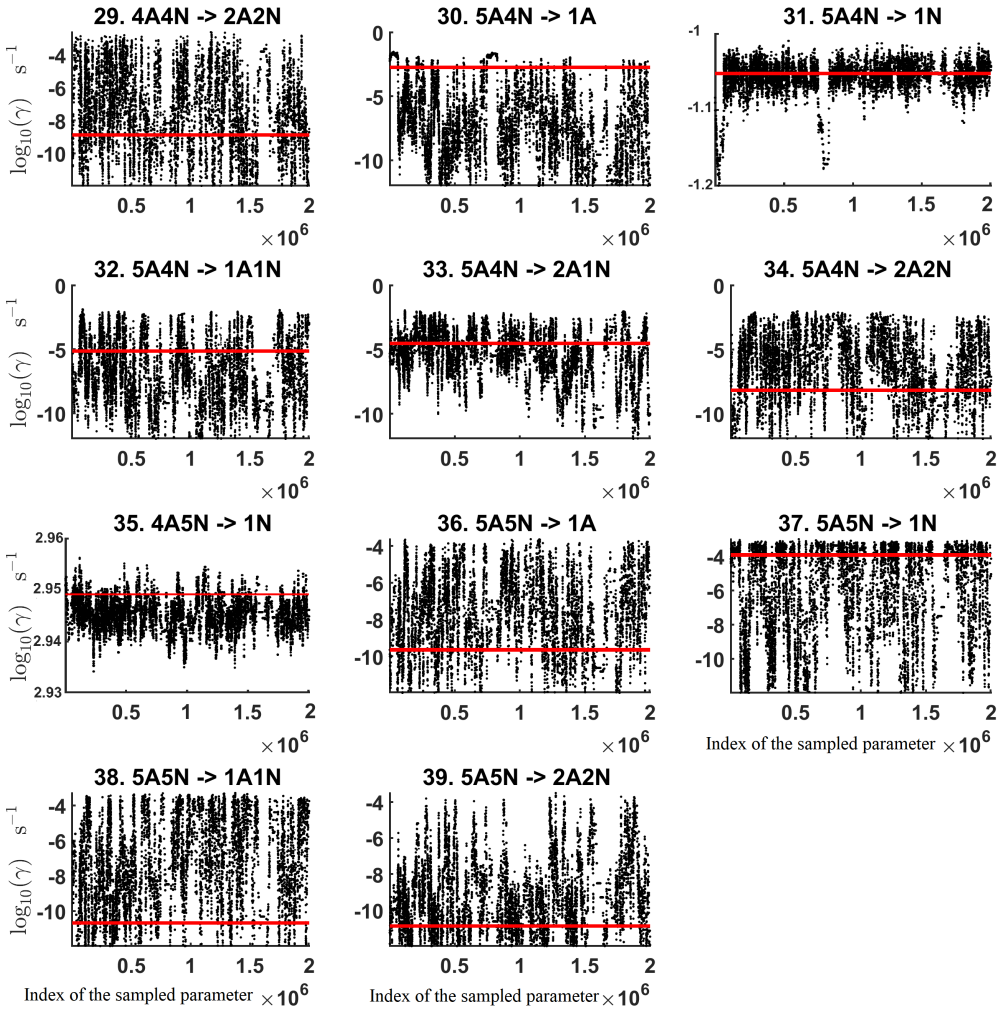


Figure B2. Parameter chains (for parameter indexes ranging from 29 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

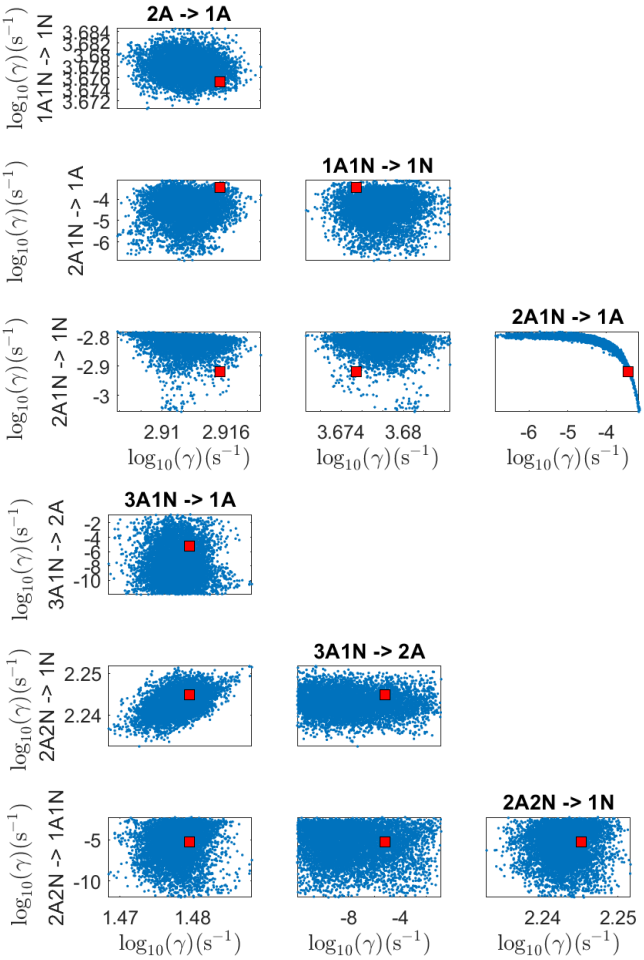


Figure B3. Pairwise marginal posterior distributions (for parameter indexes ranging from 1 to 8) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

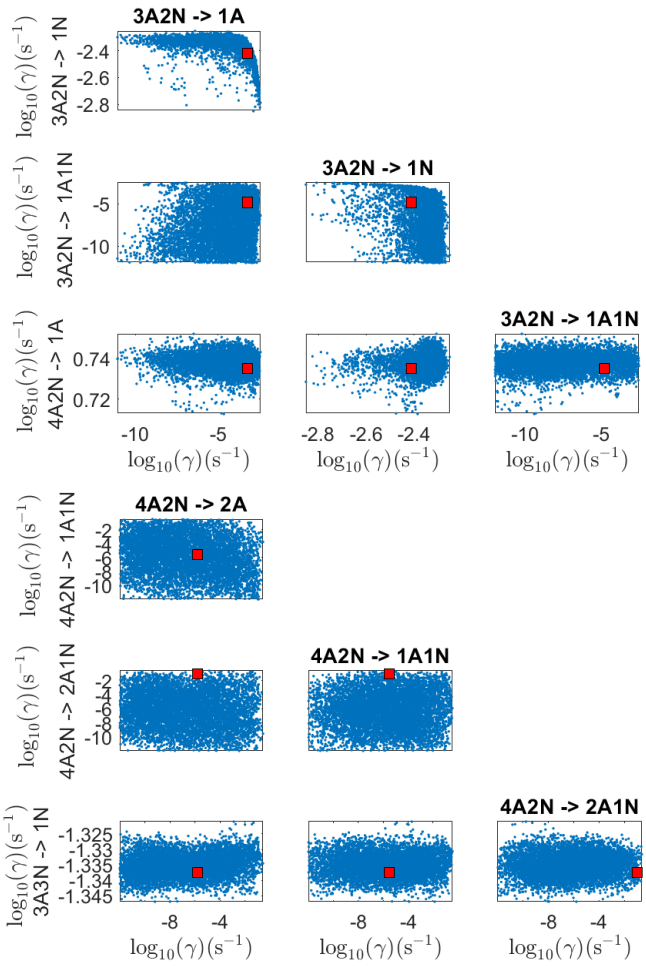


Figure B4. Pairwise marginal posterior distributions (for parameter indexes ranging from 9 to 16) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

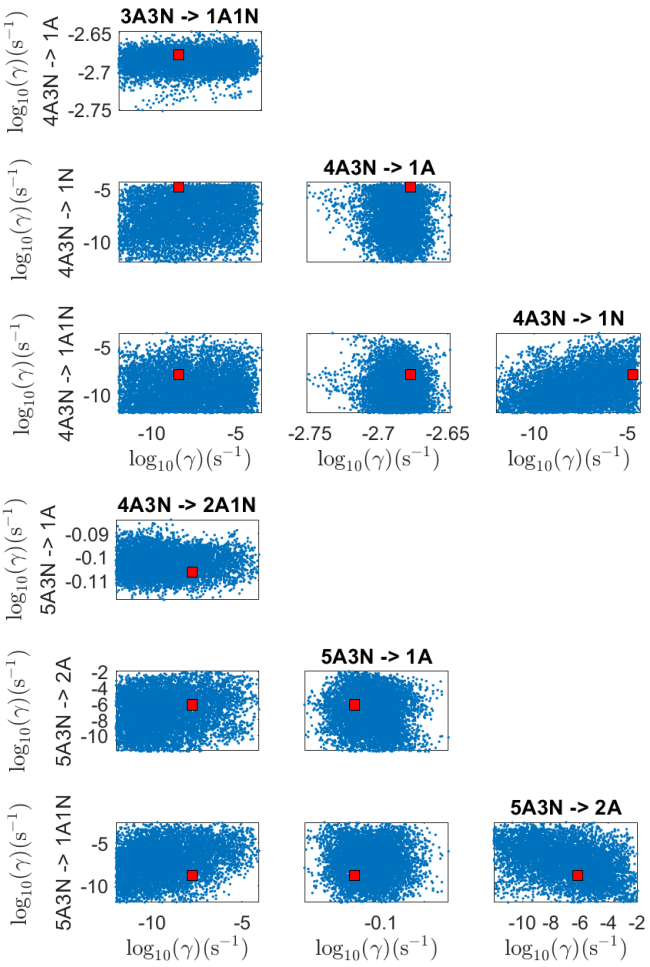


Figure B5. Pairwise marginal posterior distributions (for parameter indexes ranging from 17 to 24) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

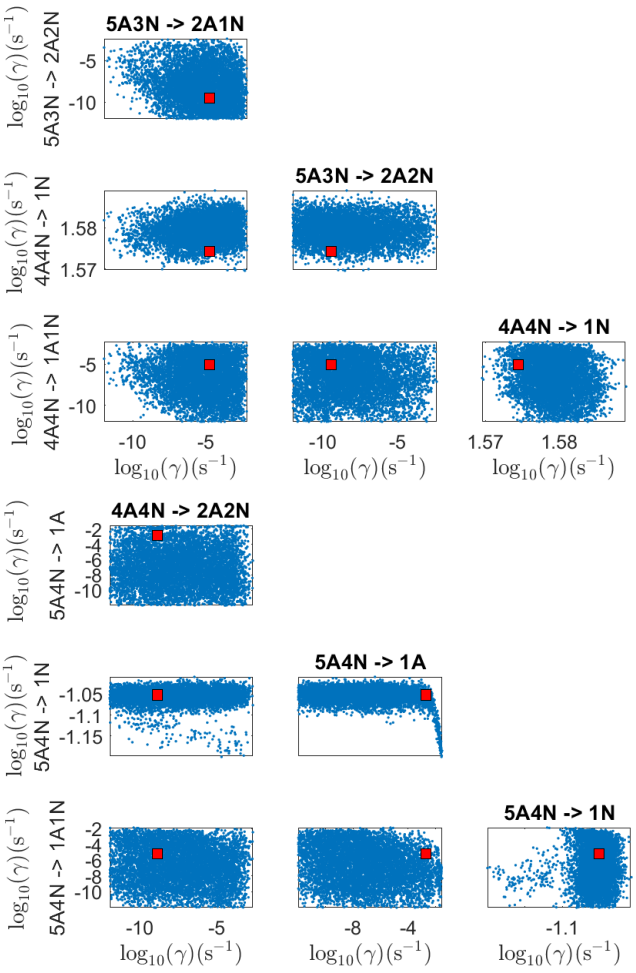


Figure B6. Pairwise marginal posterior distributions (for parameter indexes ranging from 25 to 32) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

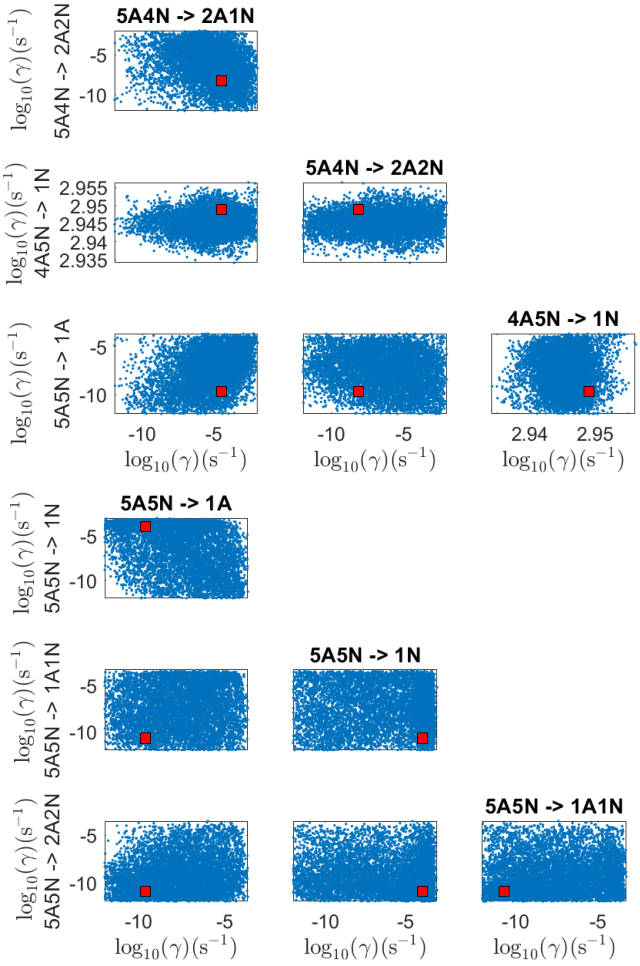
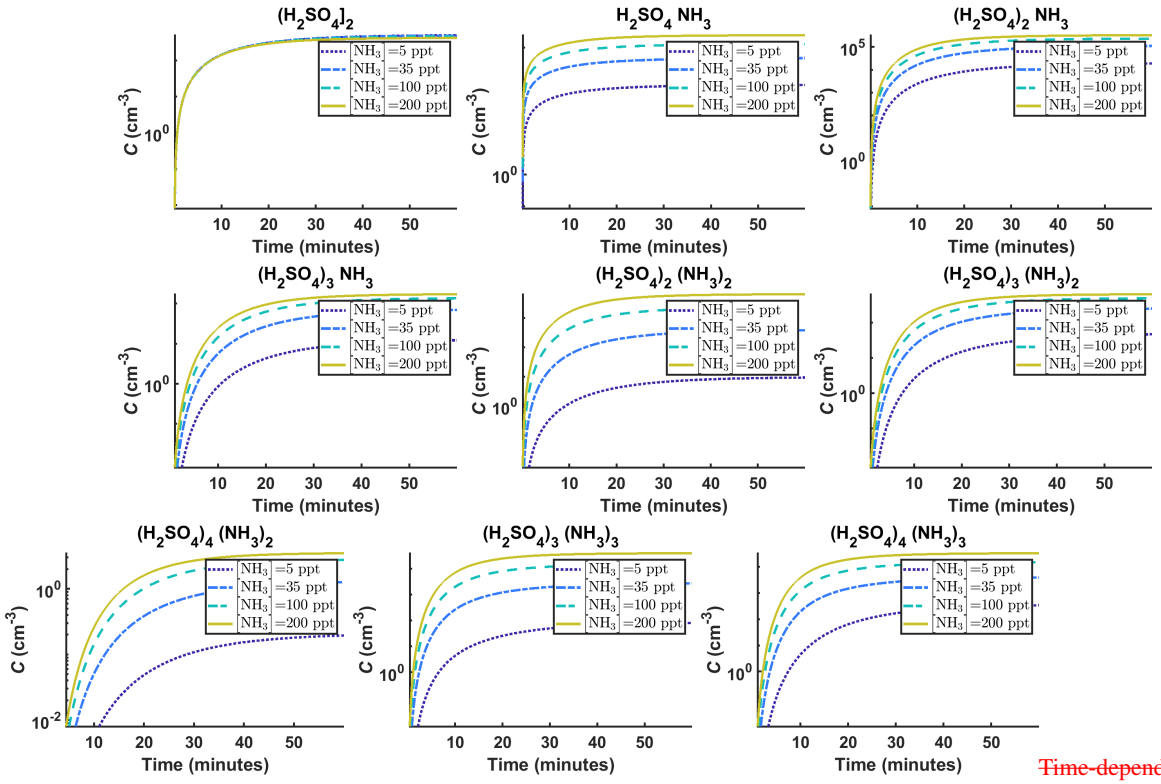


Figure B7. Pairwise marginal posterior distributions (for parameter indexes ranging from 33 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from steady-state cluster concentration measurements at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Appendix C: ~~Identification of the evaporation rates from transient data~~

Appendix C: Estimation of the evaporation rates from transient data



~~Time-dependent cluster concentrations, part 1. Simulated time evolution of concentrations for different cluster types at temperature T=278 K for varying $[NH_3]$ concentration: 5 ppt, 35 ppt, 100 ppt and 200 ppt (see the legend). All the model outputs are amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. Time resolution comprises 1.5 minutes. The source of sulphuric acid monomer is $[H_2SO_4] = 6.3 \times 10^4\ s^{-1}$ in all simulations.~~

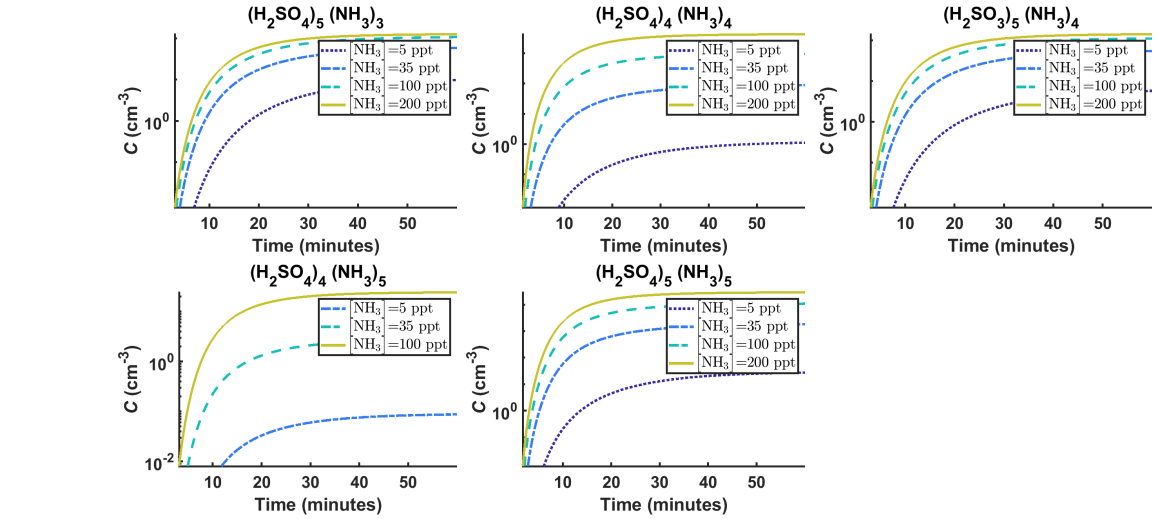


Figure C1. Time-dependent cluster concentrations~~, part 2.~~. Simulated time evolution of concentrations for different cluster types at temperature T=278 K for varying $[NH_3]$ concentration: 5 ppt, 35 ppt, 100 ppt and 200 ppt (see the legend). All the model outputs are amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. Time resolution comprises 1.5 minutes. The source of sulphuric acid monomer is $[H_2SO_4] = 6.3 \times 10^4\ s^{-1}$ in all simulations. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

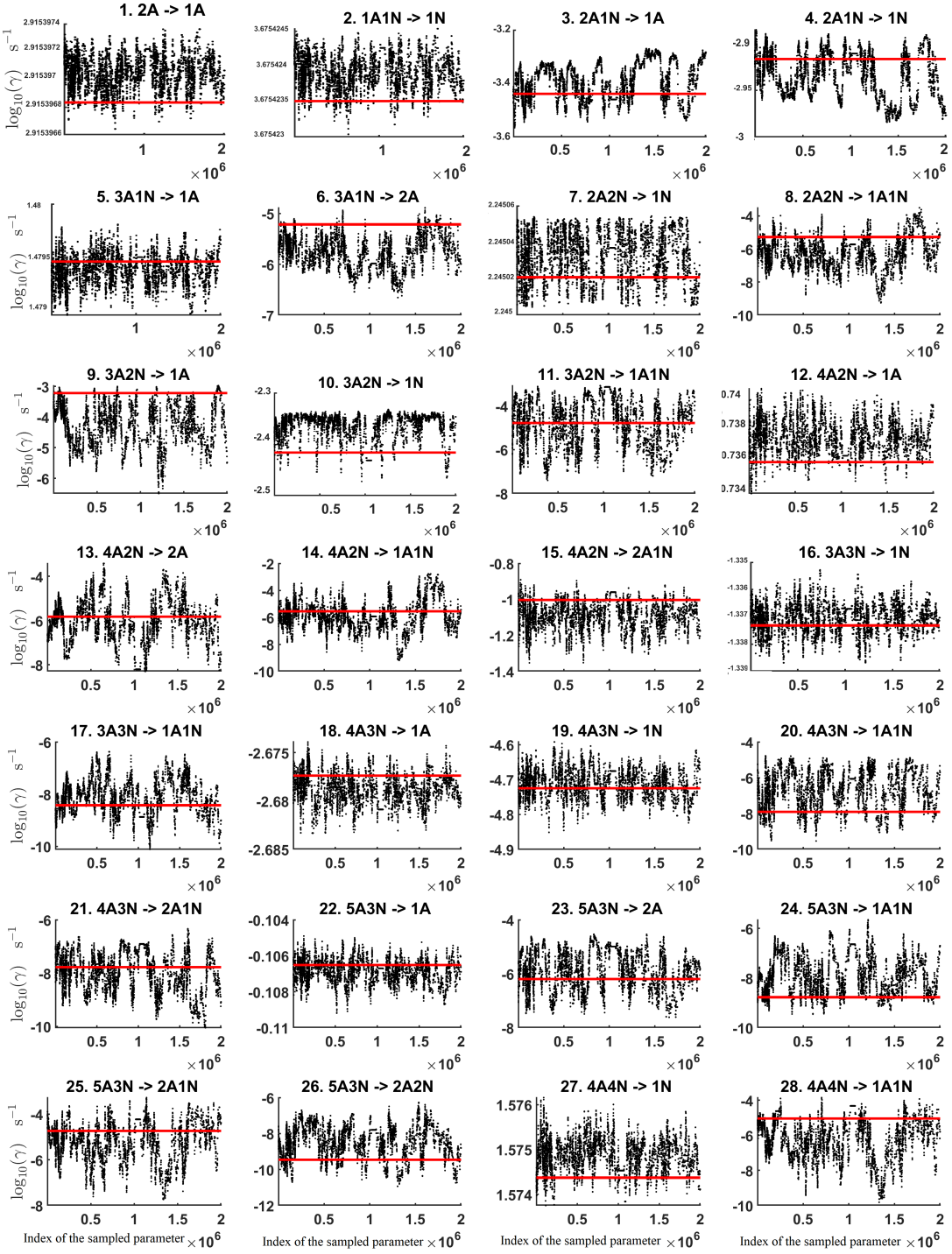


Figure C2. Parameter chains (for parameter indexes ranging from 1 to 28) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data.

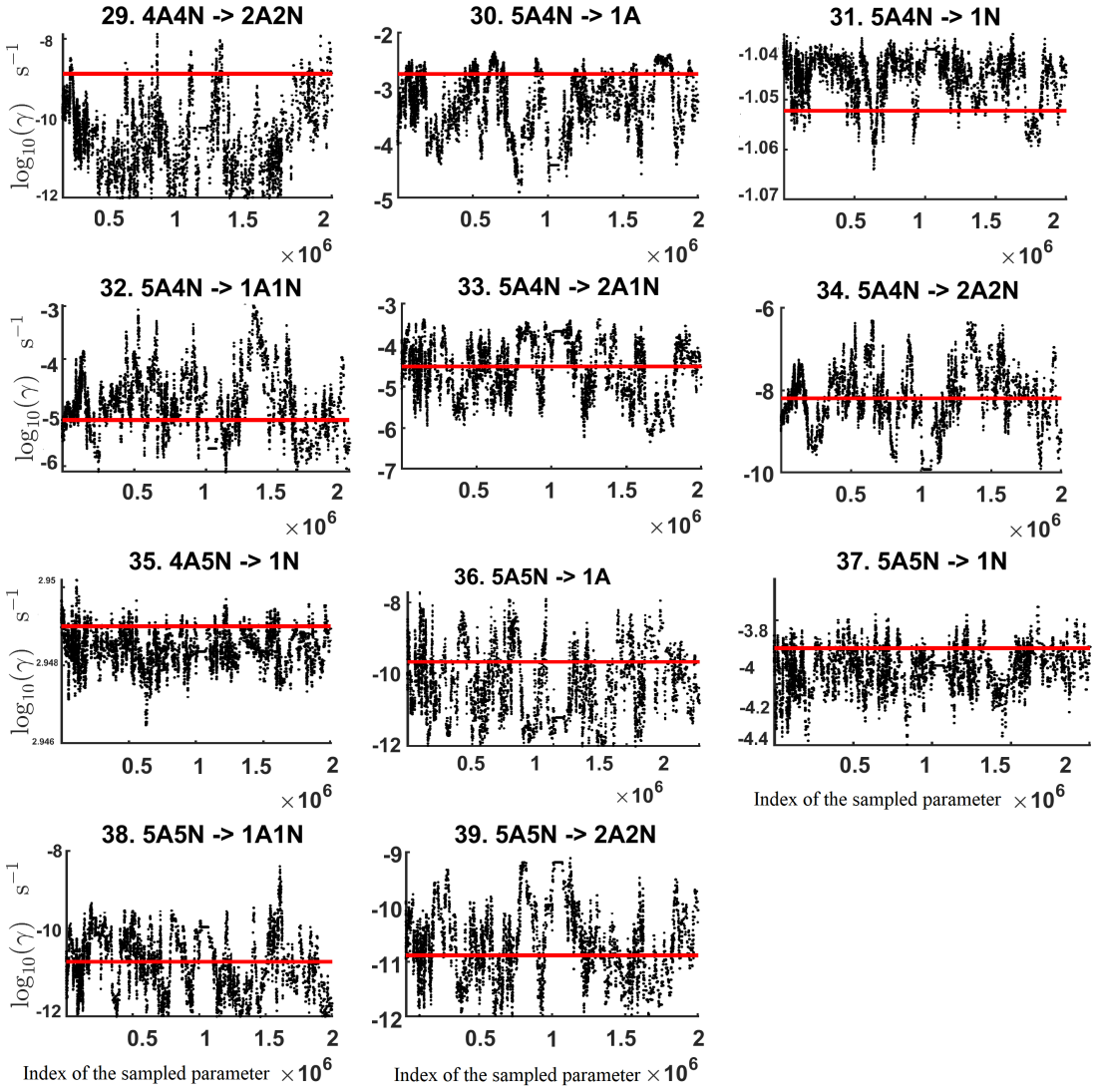


Figure C3. Parameter chains (for parameter indexes ranging from 29 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

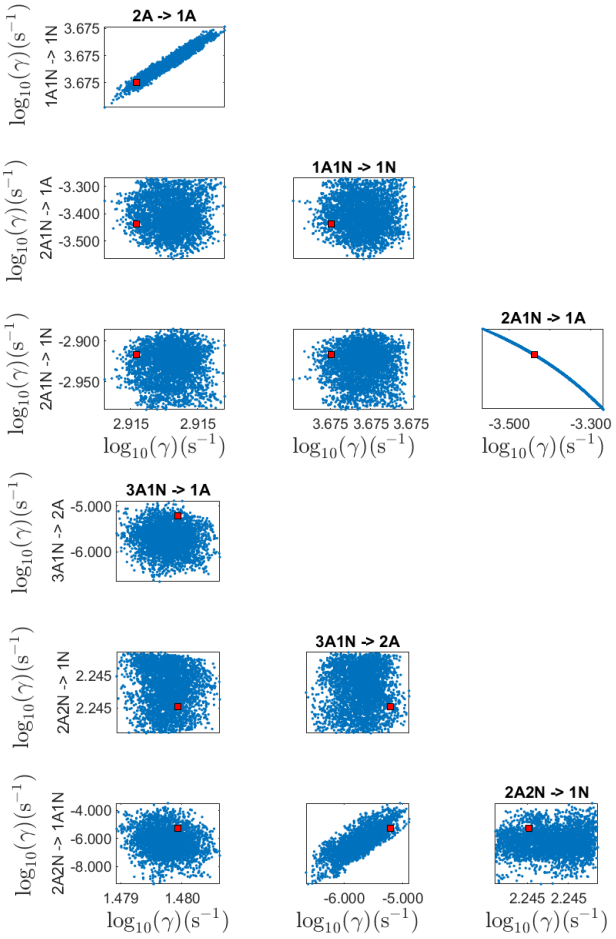


Figure C4. Pairwise marginal posterior distributions (for parameter indexes ranging from 1 to 8) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

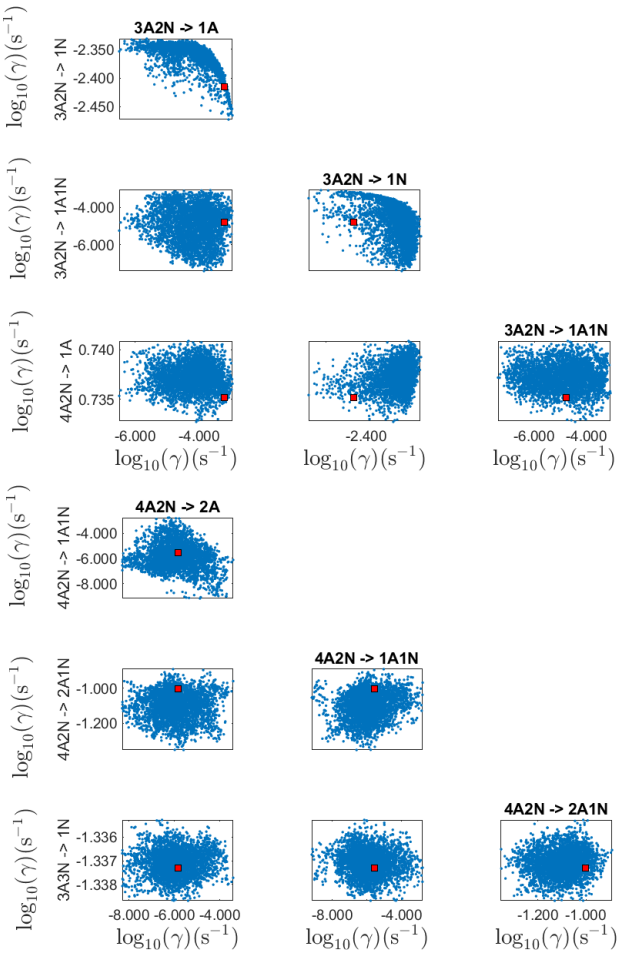


Figure C5. Pairwise marginal posterior distributions (for parameter indexes ranging from 9 to 16) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) determined from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

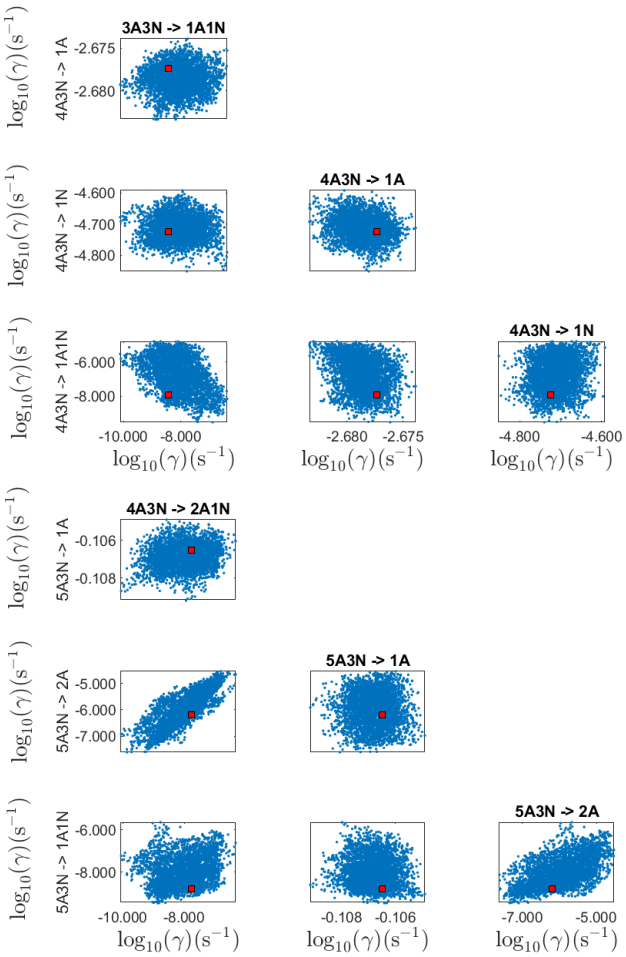


Figure C6. Pairwise marginal posterior distributions (for parameter indexes ranging from 17 to 24) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

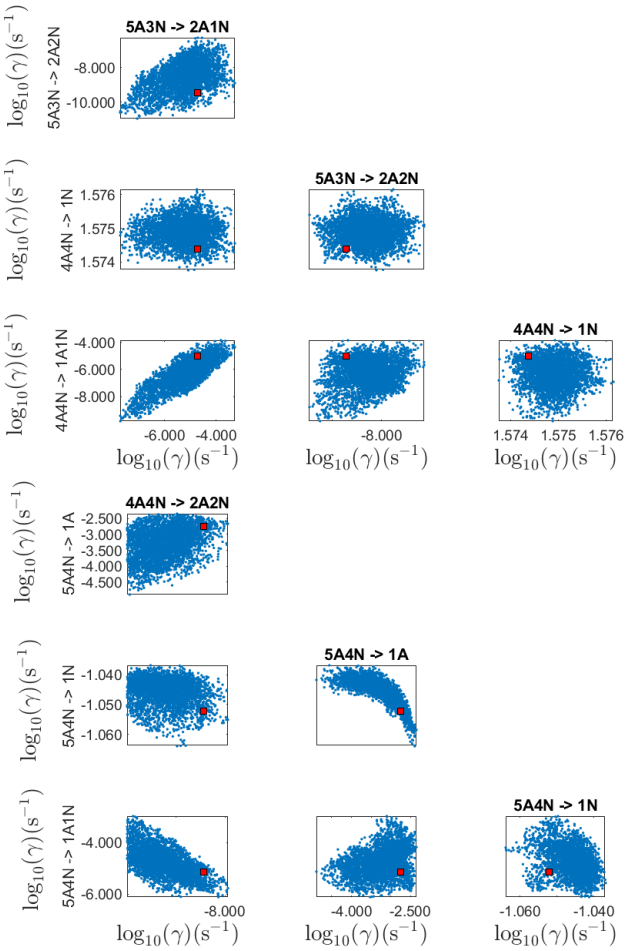


Figure C7. Pairwise marginal posterior distributions (for parameter indexes ranging from 25 to 32) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

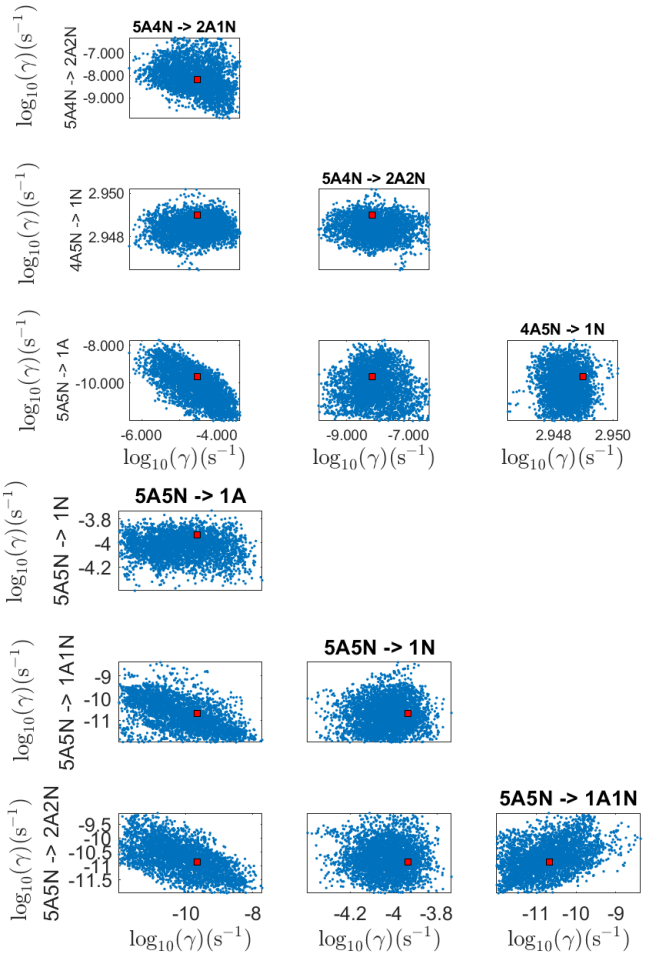


Figure C8. Pairwise marginal posterior distributions (for parameter indexes ranging from 33 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) from transient measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Symbol	Steady-state data (s^{-1})	Transient data (s^{-1})	QC (s^{-1})
1: 2A → 1A	$\mathbf{8.16 \times 10^{2}}$ (8.05×10^{2},8.31×10^{2})	$\mathbf{8.23 \times 10^{2}}$	8.23×10^{2}
2: 1A1N → 1N	$\mathbf{4.75 \times 10^{3}}$ (4.69×10^{3},4.87×10^{3})	$\mathbf{4.74 \times 10^{3}}$	4.74×10^{3}
3: 2A1N → 1A	$\mathbf{4.22 \times 10^{-4}}$ (5.92×10^{-11},7.27×10^{-4})	$\mathbf{3.30 \times 10^{-4}}$ (1.75×10^{-4},5.37×10^{-4})	3.64×10^{-4}
4: 2A1N → 1N	$\mathbf{1.56 \times 10^{-3}}$ (8.78×10^{-4},1.67×10^{-3})	$\mathbf{1.33 \times 10^{-3}}$ (1.04×10^{-3},1.4×10^{-3})	1.21×10^{-3}
5: 3A1N → 1A	$\mathbf{2.99 \times 10^{1}}$ (2.94×10^{1},3.08×10^{1}	$\mathbf{3.02 \times 10^{1}}$ (3.01×10^{1},3.02×10^{1})	3.02×10^{1}
6: 3A1N → 2A	– 1.50×10^{-1}	$\mathbf{2.81 \times 10^{-6}}$ (2.86×10^{-9},2.76×10^{-3})	6.09×10^{-6}
7: 2A2N → 1N	$\mathbf{1.74 \times 10^{2}}$ (1.71×10^{2},1.79×10^{2})	$\mathbf{1.76 \times 10^{2}}$	1.76×10^{2}
8: 2A2N → 1A1N	$\mathbf{5.52 \times 10^{-4}}$ $< 5.16 \times 10^{-3}$	$\mathbf{2.11 \times 10^{-6}}$ (2.95×10^{-10},3.59×10^{-4})	5.33×10^{-6}
9: 3A2N → 1A	$\mathbf{3.30 \times 10^{-4}}$ $< 2.91 \times 10^{-3}$	$\mathbf{7.51 \times 10^{-4}}$ (3.18×10^{-7},1.78×10^{-3})	6.07×10^{-4}
10: 3A2N → 1N	$\mathbf{4.47 \times 10^{-3}}$ (5.85×10^{-4},5.60×10^{-3})	$\mathbf{4.16 \times 10^{-3}}$ (2.86×10^{-3},4.66×10^{-3})	3.84×10^{-3}
11: 3A2N → 1A1N	$\mathbf{9.79 \times 10^{-5}}$ $< 3.88 \times 10^{-3}$	$\mathbf{1.00 \times 10^{-5}}$ (4.68×10^{-10},7.22×10^{-4})	1.64×10^{-5}
12: $4A2N \rightarrow 1A$	$\mathbf{5.50 \times 10^{0}}$ (4.50×10^{0},5.72×10^{0})	$\mathbf{5.46 \times 10^{0}}$ (5.39×10^{0},5.51×10^{0})	5.43×10^{0}
13: 4A2N → 2A	$\mathbf{5.24 \times 10^{-7}}$ $< 2.74 \times 10^{-1}$	$\mathbf{1.03 \times 10^{-6}}$ (5.66×10^{-11},1.88×10^{-2})	1.48×10^{-6}
14: 4A2N → 1A1N	$\mathbf{2.79 \times 10^{-1}}$ $< 6.92 \times 10^{-1}$	$\mathbf{2.78 \times 10^{-6}}$ (6.50×10^{-10},1.66×10^{-3})	2.80×10^{-6}
15: 4A2N → 2A1N	$\mathbf{6.49 \times 10^{-2}}$ $< 1.02 \times 10^{0}$	$\mathbf{9.04 \times 10^{-2}}$ (3.66×10^{-2},1.33×10^{-1})	9.94×10^{-2}
16: 3A3N → 1N	$\mathbf{4.62 \times 10^{-2}}$ (4.50×10^{-2},4.78×10^{-2})	$\mathbf{4.61 \times 10^{-2}}$ (4.58×10^{-2},4.62×10^{-2})	4.60×10^{-2}
17: 3A3N → 1A1N	$\mathbf{1.37 \times 10^{-9}}$ $< 3.58 \times 10^{-4}$	$\mathbf{6.32 \times 10^{-9}}$ (1.05×10^{-12},4.91×10^{-6})	3.74×10^{-9}
18: 4A3N → 1A	$\mathbf{2.08 \times 10^{-3}}$ (1.79×10^{-3},2.27×10^{-3})	$\mathbf{2.10 \times 10^{-3}}$ (2.07×10^{-3},2.12×10^{-3})	2.10×10^{-3}
19: 4A3N → 1N	$\mathbf{1.19 \times 10^{-5}}$ $< 7.29 \times 10^{-5}$	$\mathbf{1.96 \times 10^{-5}}$ (1.11×10^{-5},2.50×10^{-5})	1.88×10^{-5}
20: 4A3N → 1A1N	$\mathbf{9.29 \times 10^{-11}}$ $< 2.65 \times 10^{-4}$	– (1.81×10^{-12},1.96×10^{-5})	1.23×10^{-8}

Table C1. Part 1. Evaporation rates (units given in s^{-1}) determined from the steady-state and the transient data presented in Figure 5-6 and ~~Figures~~ Figs. 16-17, respectively. For parameters that have a posterior distribution with the clear peak and practically zero probability density elsewhere, the mode of the distribution (bold face) is given together with the range of possible values in the parenthesis. In some of the cases only the limits can be determined. The last column presents the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Symbol	Steady-state data (s^{-1})	Transient data (s^{-1})	QC (s^{-1})
21: 4A3N → 2A1N	– $< 2.14 \times 10^{-4}$	$\mathbf{4.83 \times 10^{-9}}$ $(3.36 \times 10^{-12}, 6.93 \times 10^{-6})$	1.66×10^{-8}
22: 5A3N → 1A	$\mathbf{7.88 \times 10^{-1}}$ $(7.56 \times 10^{-1}, 8.20 \times 10^{-1})$	$\mathbf{7.81 \times 10^{-1}}$ $(7.77 \times 10^{-1}, 7.86 \times 10^{-1})$	7.83×10^{-1}
23: 5A3N → 2A	$\mathbf{2.35 \times 10^{-8}}$ ($< 1.21 \times 10^{-2}$)	$\mathbf{6.34 \times 10^{-7}}$ $(1.26 \times 10^{-11}, 3.35 \times 10^{-4})$	6.37×10^{-7}
24: 5A3N → 1A1N	$\mathbf{9.12 \times 10^{-12}}$ $< 3.39 \times 10^{-3}$	$\mathbf{1.50 \times 10^{-9}}$ $(1.02 \times 10^{-12}, 2.22 \times 10^{-6})$	1.70×10^{-9}
25: 5A3N → 2A1N	$\mathbf{7.22 \times 10^{-4}}$ $< 6.95 \times 10^{-3}$	$\mathbf{1.24 \times 10^{-5}}$ $(1.86 \times 10^{-8}, 5.33 \times 10^{-4})$	1.85×10^{-5}
26: 5A3N → 2A2N	$\mathbf{1.52 \times 10^{-8}}$ $< 4.49 \times 10^{-3}$	– $< 1.25 \times 10^{-4}$	3.52×10^{-10}
27: 4A4N → 1N	$\mathbf{3.79 \times 10^{1}}$ $(3.70 \times 10^{1}, 3.88 \times 10^{1})$	$\mathbf{3.76 \times 10^{1}}$ $(3.75 \times 10^{1}, 3.77 \times 10^{1})$	3.75×10^{1}
28: 4A4N → 1A1N	– $< 5.38 \times 10^{-3}$	$\mathbf{9.05 \times 10^{-6}}$ $(1.52 \times 10^{-10}, 2.57 \times 10^{-4})$	9.06×10^{-6}
29: 4A4N → 2A2N	$\mathbf{2.07 \times 10^{-12}}$ $< 2.43 \times 10^{-3}$	$\mathbf{8.55 \times 10^{-11}}$ $< 1.90 \times 10^{-4}$	1.33×10^{-9}
30: 5A4N → 1A	$\mathbf{3.87 \times 10^{-6}}$ $< 2.52 \times 10^{-2}$	$\mathbf{2.51 \times 10^{-3}}$ $(1.20 \times 10^{-6}, 5.86 \times 10^{-3})$	1.77×10^{-3}
31: 5A4N → 1N	$\mathbf{8.92 \times 10^{-2}}$ $(6.68 \times 10^{-2}, 9.74 \times 10^{-2})$	$\mathbf{9.03 \times 10^{-2}}$ $(8.52 \times 10^{-2}, 9.19 \times 10^{-2})$	8.87×10^{-2}
32: 5A4N → 1A1N	– $< 1.55 \times 10^{-2}$	$\mathbf{3.60 \times 10^{-6}}$ $(6.48 \times 10^{-12}, 1.04 \times 10^{-3})$	7.33×10^{-6}
33: 5A4N → 2A1N	$\mathbf{2.28 \times 10^{-4}}$ $< 1.06 \times 10^{-2}$	$\mathbf{1.32 \times 10^{-4}}$ $(6.46 \times 10^{-10}, 1.53 \times 10^{-3})$	2.97×10^{-5}
34: 5A4N → 2A2N	– $< 1.08 \times 10^{-2}$	$\mathbf{7.30 \times 10^{-9}}$ $(1.51 \times 10^{-11}, 3.17 \times 10^{-4})$	6.42×10^{-9}
35: 4A5N → 1N	$\mathbf{8.75 \times 10^{2}}$ $(8.59 \times 10^{2}, 9.03 \times 10^{2})$	$\mathbf{8.88 \times 10^{2}}$ $(8.85 \times 10^{2}, 8.92 \times 10^{2})$	8.89×10^{2}
36: 5A5N → 1A	– $< 2.32 \times 10^{-4}$	– $< 1.14 \times 10^{-6}$	2.23×10^{-10}
37: 5A5N → 1N	$\mathbf{4.96 \times 10^{-4}}$ $< 9.89 \times 10^{-4}$	$\mathbf{1.00 \times 10^{-4}}$ $(3.48 \times 10^{-5}, 1.85 \times 10^{-4})$	1.17×10^{-4}
38: 5A5N → 1A1N	$\mathbf{5.93 \times 10^{-9}}$ $< 5.06 \times 10^{-4}$	$\mathbf{1.48 \times 10^{-11}}$ $< 1.06 \times 10^{-5}$	2.11×10^{-11}
39: 5A5N → 2A2N	– $< 3.09 \times 10^{-4}$	$\mathbf{2.06 \times 10^{-11}}$ $< 4.11 \times 10^{-7}$	1.31×10^{-11}

Table C2. Part 2. Evaporation rates (units given in s^{-1}) determined from the steady-state and the transient data presented in Figure 5-6 and ~~Figures~~ Figs. 16-17, respectively. For parameters that have a posterior distribution with the clear peak and practically zero probability density elsewhere, the mode of the distribution (bold face) is given together with the range of possible values in the parenthesis. In some of the cases only the limits can be determined. The last column presents the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

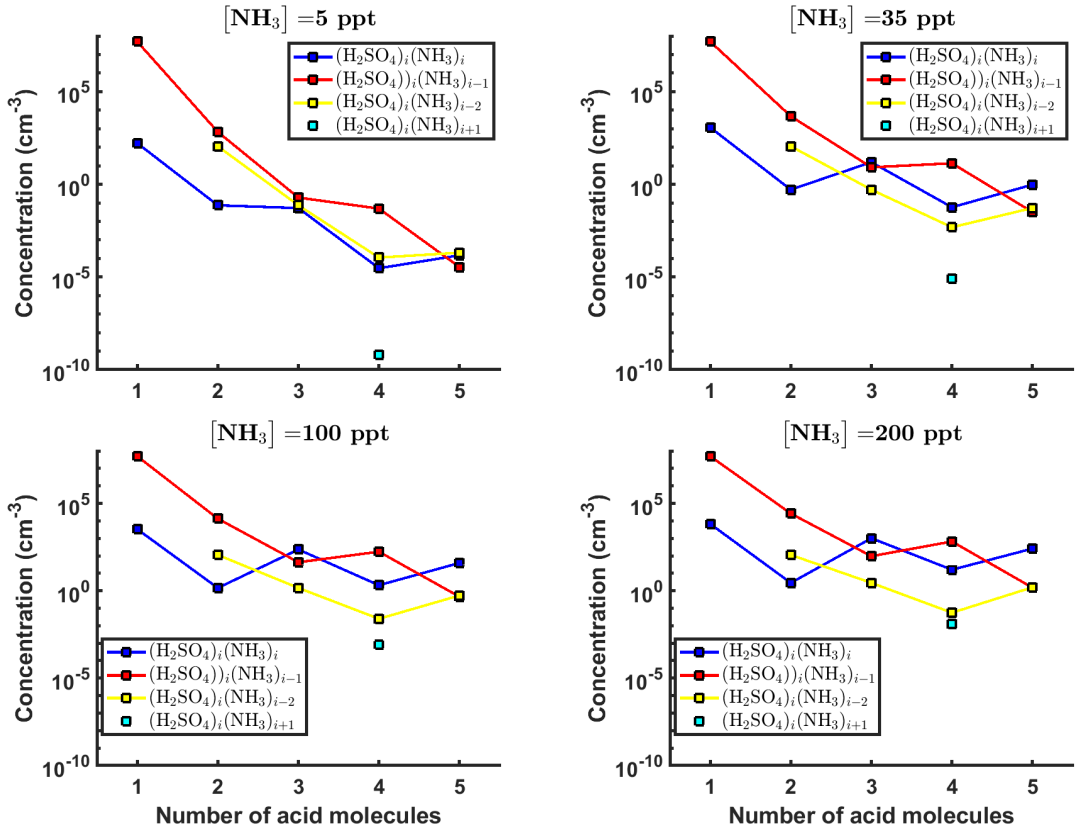


Figure D1. Steady-state cluster concentrations for the clusters containing sulphuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for $[NH_3]$ concentrations comprising (a) 5 ppt, (b) 35 ppt, (c) 100 ppt and (d) 200 ppt at temperature T=292 K amended with multivariate non-correlated Gaussian noise with standard deviation comprising 0.001% of the original cluster concentration. The source of sulphuric acid monomer comprises $[H_2SO_4] = 6.3 \times 10^4$ s^{-1}in all the simulations. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

Appendix D: ~~Identification~~ Estimation of the cluster formation enthalpies and entropies from steady-state concentration measurements

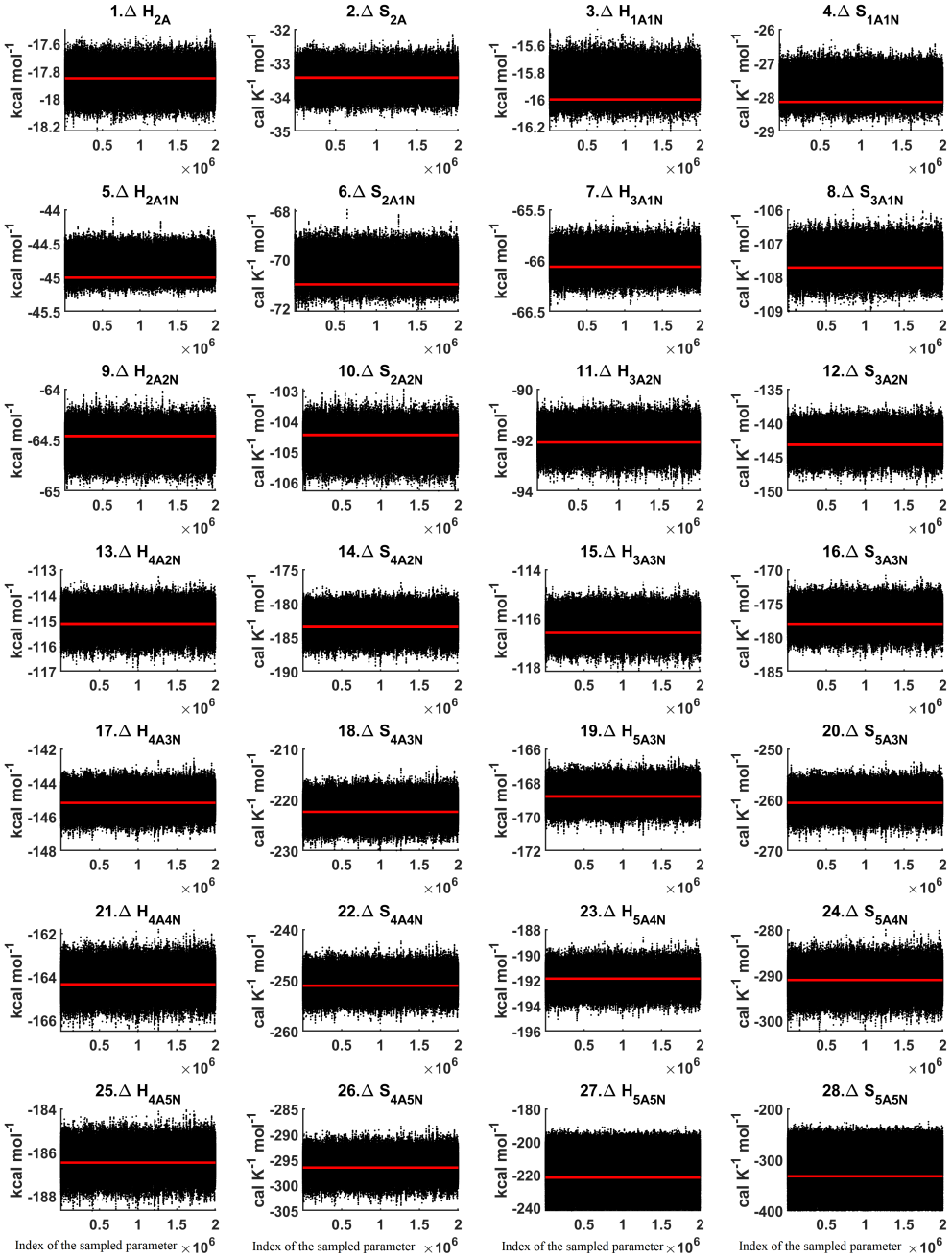


Figure D2. Parameter chains of the cluster formation enthalpies (units given in ~~kkal~~kcal/mol) and entropies (units given in $\mathrm{cal\ K^{-1}\ mol^{-1}}$) determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red lines denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote $\mathrm{H_2SO_4}$ and "$\mathrm{NH_3}$", correspondingly.

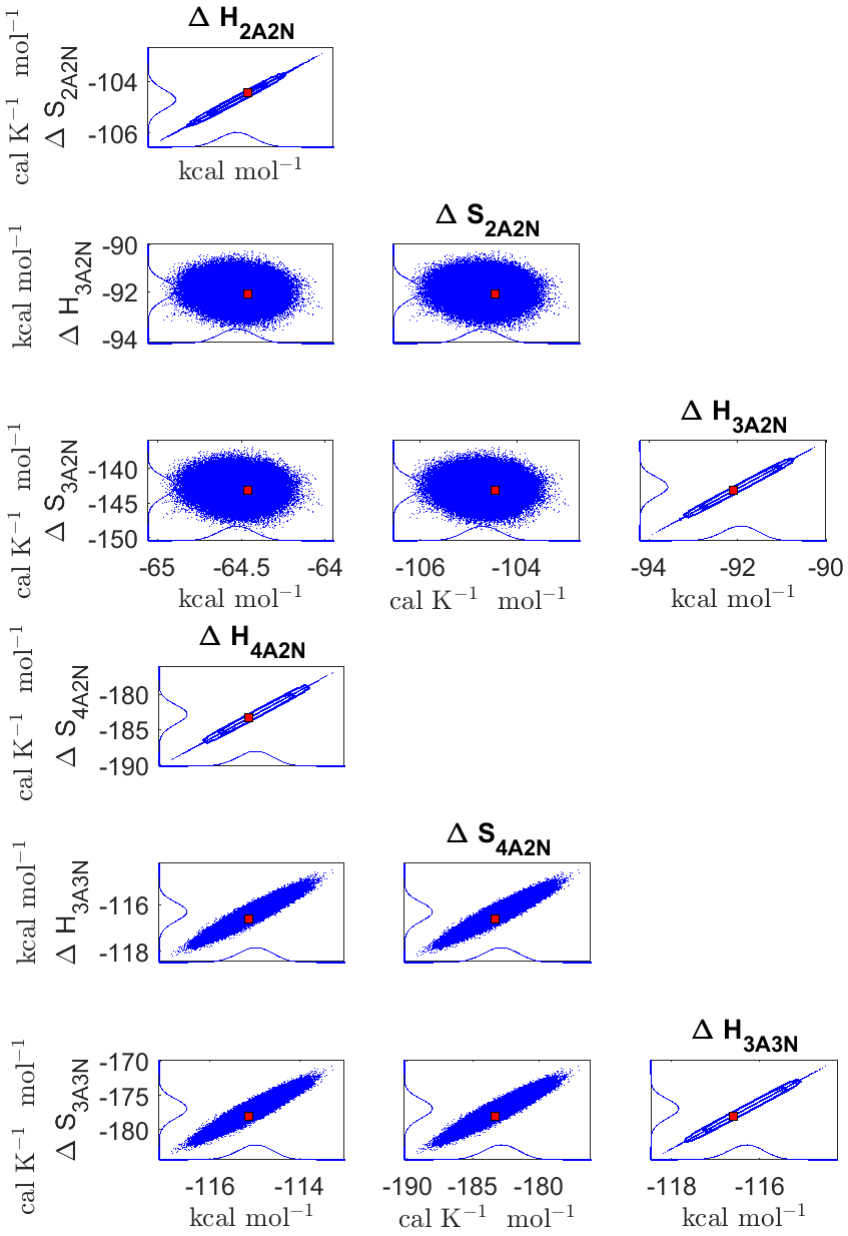


Figure D3. Pairwise marginal posterior distributions (for parameter indexes ranging from 9 to 16) of the cluster formation enthalpies and entropies determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

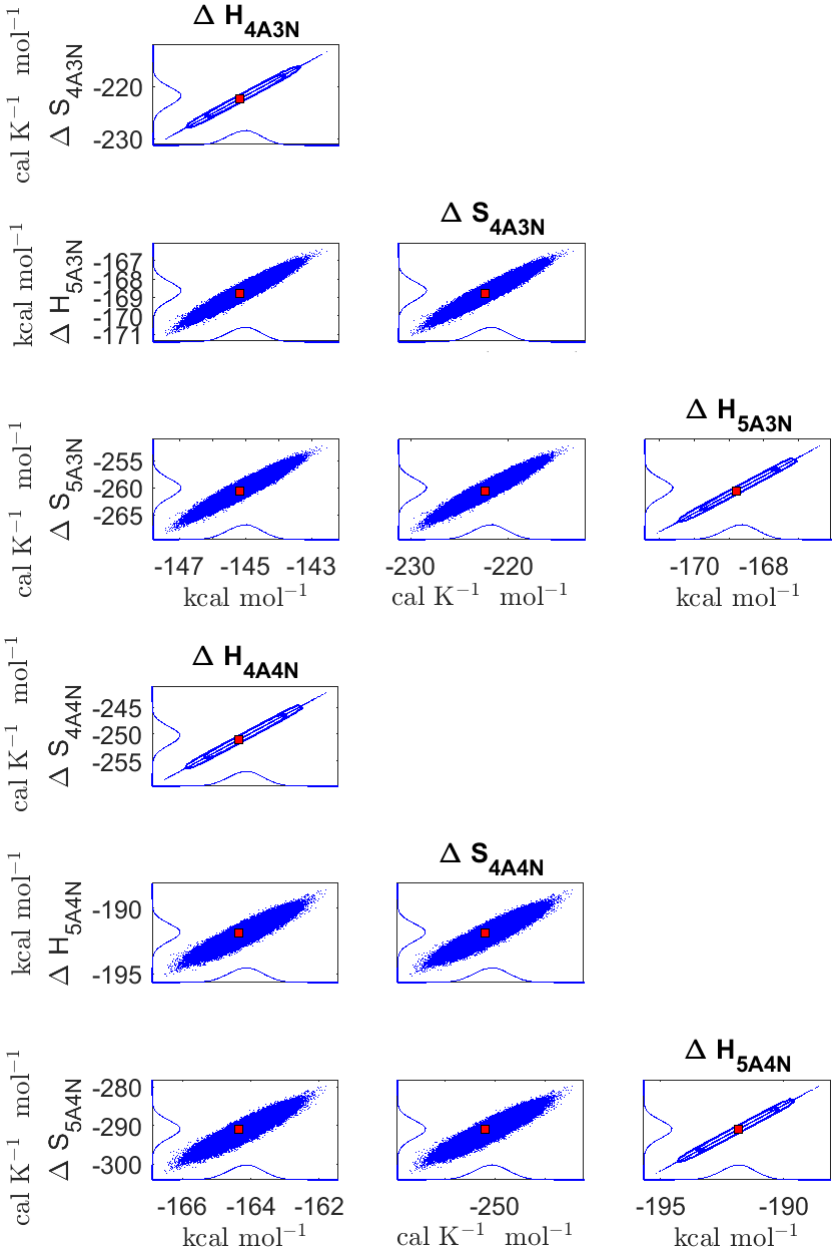


Figure D4. Pairwise marginal posterior distributions (for parameter indexes ranging from 17 to 24) of the cluster formation enthalpies and entropies determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

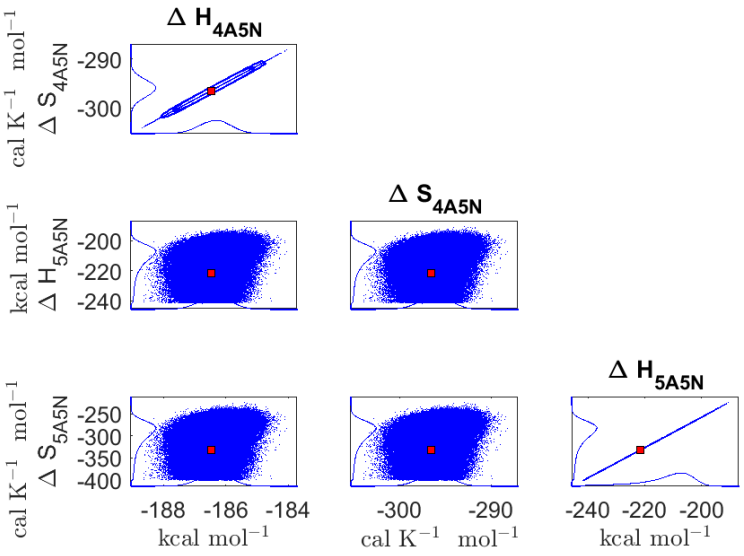


Figure D5. Pairwise marginal posterior distributions (for parameter indexes ranging from 25 to 28) of the cluster formation enthalpies and entropies determined from steady-state cluster concentration measurements at two temperatures T=278 K and T = 292 K. Red rectangles denote the baseline values from **?** used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

Symbol	Mode value	95% confidence interval	QC	Units
1: ΔH_{2A}	-17.8891	(-18.1913,-17.4941)	-17.85	kcal mol^{-1}
2: ΔS_{2A}	-33.5475	(-34.6104,-32.1575)	-33.42	cal K^{-1} mol^{-1}
3: ΔH_{1A1N}	-15.8751	(-16.2344,-15.5158)	-16	kcal mol^{-1}
4: ΔS_{1A1N}	-27.6984	(-28.9594,-26.4374)	-28.14	cal K^{-1} mol^{-1}
5: ΔH_{2A1N}	-44.8076	(-45.2922,-44.174)	-45	kcal mol^{-1}
6: ΔS_{2A1N}	-70.3501	(-72.029,-68.1545)	-71.02	cal K^{-1} mol^{-1}
7: ΔH_{3A1N}	-66.0006	(-66.428,-65.5732)	-66.06	kcal mol^{-1}
8: ΔS_{3A1N}	-107.5233	(-109.0059,-106.0407)	-107.72	cal K^{-1} mol^{-1}
9: ΔH_{2A2N}	-64.5005	(-64.9799,-64.021)	-64.46	kcal mol^{-1}
10: ΔS_{2A2N}	-104.6181	(-106.2857,-102.9505)	-104.45	cal K^{-1} mol^{-1}
11: ΔH_{3A2N}	-91.8512	(-93.9174,-90.2712)	-92.09	kcal mol^{-1}
12: ΔS_{3A2N}	-142.3625	(-149.4438,-136.9474)	-143.18	cal K^{-1} mol^{-1}
13: ΔH_{4A2N}	-115.0105	(-116.7515,-113.2696)	-115.13	kcal mol^{-1}
14: ΔS_{4A2N}	-182.938	(-188.9067,-176.9693)	-183.34	cal K^{-1} mol^{-1}
15: ΔH_{3A3N}	-116.3273	(-118.1437,-114.5108)	-116.6	kcal mol^{-1}
16: ΔS_{3A3N}	-177.0462	(-183.2768,-170.8156)	-177.99	cal K^{-1} mol^{-1}
17: ΔH_{4A3N}	-144.9757	(-147.3975,-142.554)	-145.17	kcal mol^{-1}
18: ΔS_{4A3N}	-221.6575	(-229.9554,-213.3595)	-222.33	cal K^{-1} mol^{-1}
19: ΔH_{5A3N}	-168.7305	(-171.0579,-166.4031)	-168.79	kcal mol^{-1}
20: ΔS_{5A3N}	-260.3509	(-268.3225,-252.3794)	-260.55	cal K^{-1} mol^{-1}
21: ΔH_{4A4N}	-164.1272	(-166.4394,-161.815)	-164.35	kcal mol^{-1}
22: ΔS_{4A4N}	-250.2634	(-258.1819,-242.3449)	-251.03	cal K^{-1} mol^{-1}
23: ΔH_{5A4N}	-191.7779	(-194.9426,-188.6133)	-191.86	kcal mol^{-1}
24: ΔS_{5A4N}	-290.7782	(-301.6196,-279.9369)	-291.05	cal K^{-1} mol^{-1}
25: ΔH_{4A5N}	-186.3473	(-188.639,-184.0557)	-186.47	kcal mol^{-1}
26: ΔS_{4A5N}	-296.0839	(-303.9359,-288.2319)	-296.51	cal K^{-1} mol^{-1}
27: ΔH_{5A5N}	-205.943	(-241.6193,-190.6532)	-221.65	kcal mol^{-1}
28: ΔS_{5A5N}	-277.4	(-,-224.8575)	-332.49	cal K^{-1} mol^{-1}

Table D1. Thermodynamic parameters identified from steady-state data measured at two temperatures (278 and 292 K). The last column presents the quantum-chemistry based values from ~~?~~ (?) used to generate the synthetic data. Here the symbols ΔH and ΔS stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote H_2SO_4 and "NH_3", correspondingly.

Figure D6. One-dimensional marginal distributions (for parameter indexes ranging from 1 to 28) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) at temperature 278 K obtained from a posterior distribution of thermodynamic parameters (cluster formation enthalpies and entropies) determined from steady-state cluster concentration measured at temperatures 278 K and 292 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

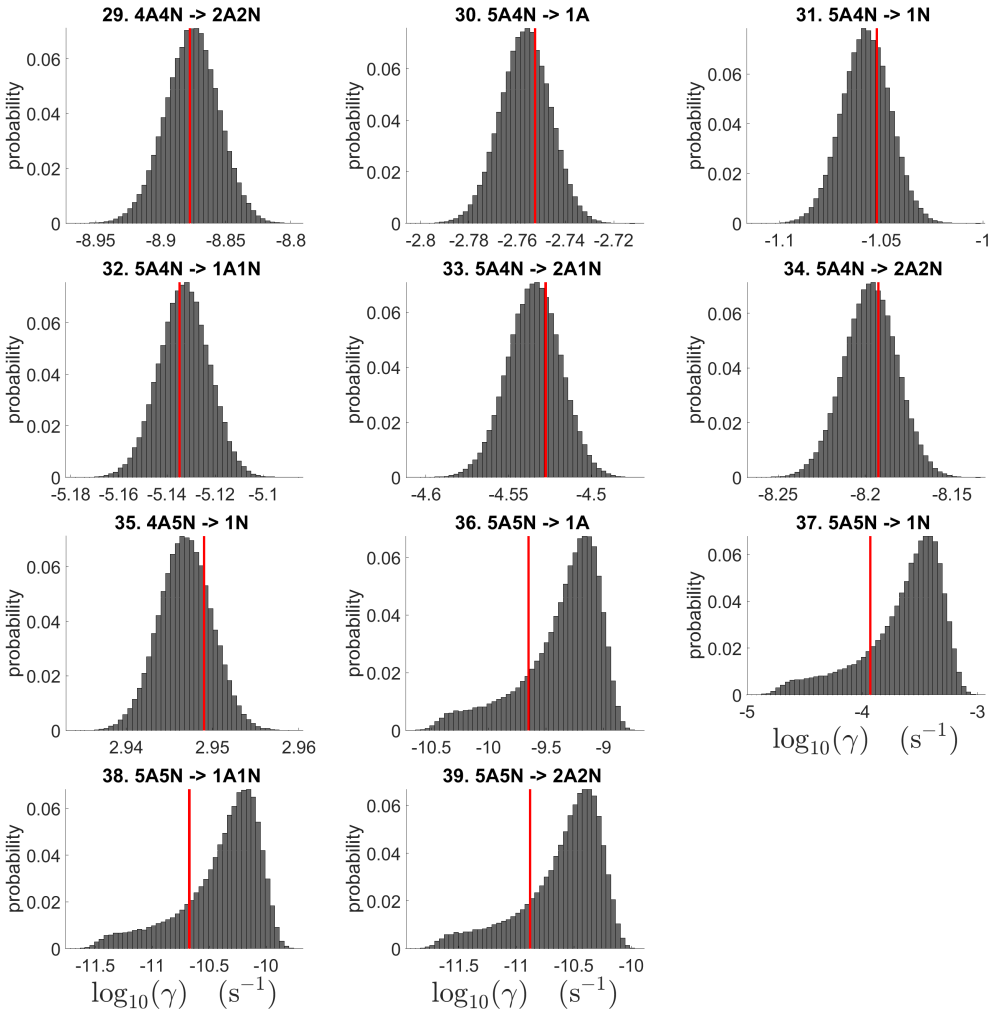


Figure D7. One-dimensional marginal distributions (for parameter indexes ranging from 29 to 39) of the base 10 logarithm of the evaporation rates (units given in s^{-1}) at temperature 278 K obtained from a posterior distribution of thermodynamic parameters (cluster formation enthalpies and entropies) determined from steady-state cluster concentration measured at temperatures 278 K and 292 K. Red lines denote the baseline values from **?** used to generate the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Symbol	Steady-state data for 278 K and 292 K (s^{-1})	QC (s^{-1})
1: 2A $\rightarrow$ 1A	$\mathbf{8.17 \times 10^{2}}$ (8.03×10^{2},8.36×10^{2})	8.23×10^{2}
2: 1A1N $\rightarrow$ 1N	$\mathbf{4.76 \times 10^{3}}$ (4.66×10^{3},4.87×10^{3})	4.74×10^{3}
3: 2A1N $\rightarrow$ 1A	$\mathbf{3.64 \times 10^{-4}}$ (3.48×10^{-4},3.84×10^{-4})	3.64×10^{-4}
4: 2A1N $\rightarrow$ 1N	$\mathbf{1.23 \times 10^{-3}}$ (1.16×10^{-3},1.29×10^{-3})	1.21×10^{-3}
5: 3A1N $\rightarrow$ 1A	$\mathbf{3.01 \times 10^{1}}$ (2.93×10^{1},3.09×10^{1})	3.02×10^{1}
6: 3A1N $\rightarrow$ 2A	$\mathbf{6.12 \times 10^{-6}}$ (5.77×10^{-6},6.47×10^{-6})	6.09×10^{-6}
7: 2A2N $\rightarrow$ 1N	$\mathbf{1.77 \times 10^{2}}$ (1.71×10^{2},1.82×10^{2})	1.76×10^{2}
8: 2A2N $\rightarrow$ 1A1N	$\mathbf{5.33 \times 10^{-6}}$ (5.02×10^{-6},5.64×10^{-6})	5.33×10^{-6}
9: 3A2N $\rightarrow$ 1A	$\mathbf{6.09 \times 10^{-4}}$ (5.14×10^{-4},7.05×10^{-4})	6.07×10^{-4}
10: 3A2N $\rightarrow$ 1N	$\mathbf{3.89 \times 10^{-3}}$ (3.27×10^{-3},4.50×10^{-3})	3.84×10^{-3}
11: 3A2N $\rightarrow$ 1A1N	$\mathbf{1.65 \times 10^{-5}}$ (1.40×10^{-5},1.90×10^{-5})	1.64×10^{-5}
12: $4A2N \rightarrow 1A$	$\mathbf{5.45 \times 10^{0}}$ (5.25×10^{0},5.65×10^{0})	5.43×10^{0}
13: 4A2N $\rightarrow$ 2A	$\mathbf{1.49 \times 10^{-6}}$ (1.27×10^{-6},1.72×10^{-6})	1.48×10^{-6}
14: 4A2N $\rightarrow$ 1A1N	$\mathbf{2.82 \times 10^{-6}}$ (2.37×10^{-6},3.26×10^{-6})	2.80×10^{-6}
15: 4A2N $\rightarrow$ 2A1N	$\mathbf{1.01 \times 10^{-1}}$ (8.35×10^{-2},1.18×10^{-1})	9.94×10^{-2}
16: 3A3N $\rightarrow$ 1N	$\mathbf{4.64 \times 10^{-2}}$ (4.47×10^{-2},4.81×10^{-2})	4.60×10^{-2}
17: 3A3N $\rightarrow$ 1A1N	$\mathbf{3.77 \times 10^{-9}}$ (3.19×10^{-9},4.36×10^{-9})	3.74×10^{-9}
18: 4A3N $\rightarrow$ 1A	$\mathbf{2.08 \times 10^{-3}}$ (1.86×10^{-3},2.29×10^{-3})	2.10×10^{-3}
19: 4A3N $\rightarrow$ 1N	$\mathbf{1.87 \times 10^{-5}}$ (1.69×10^{-5},2.05×10^{-5})	1.88×10^{-5}
20: 4A3N $\rightarrow$ 1A1N	$\mathbf{1.21 \times 10^{-8}}$ (1.09×10^{-8},1.33×10^{-8})	1.23×10^{-8}

Table D2. Part 1. Evaporation rates (units given in s^{-1}) computed from a posterior distribution of the thermodynamic parameters (cluster formation enthalpies and entropies) which had previously been determined from the steady-state concentration measurements at temperatures 278 and 292 K. Here the mode of distribution (bold face) is given together with the range of possible values in the parenthesis. The last column presents the quantum-chemistry-based evaporation rates used for creating the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

Symbol	Steady-state data for 278 K and 292 K (s^{-1})	QC (s^{-1})
21: 4A3N → 2A1N	$\mathbf{1.65 \times 10^{-8}}$ (1.30×10^{-8},1.99×10^{-8})	1.66×10^{-8}
22: 5A3N → 1A	$\mathbf{7.98 \times 10^{-1}}$ (7.63×10^{-1},8.43×10^{-1})	7.83×10^{-1}
23: 5A3N → 2A	$\mathbf{6.40 \times 10^{-7}}$ (5.76×10^{-7},7.24×10^{-7})	6.37×10^{-7}
24: 5A3N → 1A1N	$\mathbf{1.71 \times 10^{-9}}$ (1.54×10^{-9},1.88×10^{-9})	1.70×10^{-9}
25: 5A3N → 2A1N	$\mathbf{1.87 \times 10^{-5}}$ (1.66×10^{-5},2.07×10^{-5})	1.85×10^{-5}
26: 5A3N → 2A2N	$\mathbf{3.56 \times 10^{-10}}$ (2.83×10^{-10},4.30×10^{-10})	3.52×10^{-10}
27: 4A4N → 1N	$\mathbf{3.82 \times 10^{1}}$ (3.69×10^{1},3.95×10^{1})	3.75×10^{1}
28: 4A4N → 1A1N	$\mathbf{8.97 \times 10^{-6}}$ (8.13×10^{-6},1.01×10^{-5})	9.06×10^{-6}
29: 4A4N → 2A2N	$\mathbf{1.34 \times 10^{-9}}$ (1.07×10^{-9},1.62×10^{-9})	1.33×10^{-9}
30: 5A4N → 1A	$\mathbf{1.76 \times 10^{-3}}$ (1.56×10^{-3},1.96×10^{-3})	1.77×10^{-3}
31: 5A4N → 1N	$\mathbf{8.70 \times 10^{-2}}$ (7.68×10^{-2},1.00×10^{-1})	8.87×10^{-2}
32: 5A4N → 1A1N	$\mathbf{7.42 \times 10^{-6}}$ (6.59×10^{-6},8.24×10^{-6})	7.33×10^{-6}
33: 5A4N → 2A1N	$\mathbf{2.92 \times 10^{-5}}$ (2.45×10^{-5},3.40×10^{-5})	2.97×10^{-5}
34: 5A4N → 2A2N	$\mathbf{6.40 \times 10^{-9}}$ (5.40×10^{-9},7.40×10^{-9})	6.42×10^{-9}
35: 4A5N → 1N	$\mathbf{8.85 \times 10^{2}}$ (8.58×10^{2},9.12×10^{2})	8.89×10^{2}
36: 5A5N → 1A	$\mathbf{5.38 \times 10^{-10}}$ (2.01×10^{-11},2.24×10^{-9})	2.23×10^{-10}
37: 5A5N → 1N	$\mathbf{2.77 \times 10^{-4}}$ (1.09×10^{-5},1.15×10^{-3})	1.17×10^{-4}
38: 5A5N → 1A1N	$\mathbf{5.05 \times 10^{-11}}$ (1.87×10^{-12},2.10×10^{-10})	2.11×10^{-11}
39: 5A5N → 2A2N	$\mathbf{3.07 \times 10^{-11}}$ (1.16×10^{-12},1.28×10^{-10})	1.31×10^{-11}

Table D3. Part 2. Evaporation rates (units given in s^{-1}) computed from a posterior distribution of the thermodynamic parameters (cluster formation enthalpies and entropies) which had previously been determined from the steady-state concentration measurements at temperatures 278 and 292 K. Here the mode of distribution (bold face) is given together with the range of possible values in the parenthesis. The last column presents the quantum-chemistry-based evaporation rates used for creating the synthetic data. In reactions "A" stands for H_2SO_4 and "N" for NH_3.

590 *Author contributions.* Author Shcherbacheva A. produced the codes and conducted all the computational experiments for generation of the synthetic data and the MCMC parameter estimation, prepared all the plots presented in the manuscripts. Authors Balehowsky T. and Shcherbacheva A. are responsible for writing the Abstract, Methods and Results sections, and partly the Conclusion section. Author Olenius

T. assisted with generation of the synthetic data, preformed sanity check of the results, gave valuable comments regarding the manuscript. Authors Helin T. and Balehowsky T. actively participated in development of the methodological approach. Author Laine M. provided technical assistance with the 'mcmcstat' toolbox which was used for MCMC simulations. Author Kubečka J. assisted with the code compilation and debug. Author Haario H. assisted with interpretation of the MCMC results and proper usage of the DRAM computational method. Authors Kurtén T. and Vehkamäki H. wrote the Introduction and partly the conclusion, verified the text of the manuscript and helped to interpret the results. The latter two authors verified and edited the manuscript and helped to interpret the outcomes of the study.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. We thank the European Research Council project 692891-DAMOCLES, Academy of Finland (project number 307331), and University of Helsinki: Faculty of Science ATMATH project, for funding, and the CSC-IT Centre for Science in Espoo, Finland, for computational resources. We also thank Olli Pakarinen (Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland) for advise in plotting the synthetic data used in the present study.