Response to major revision comments for the paper "Identification of molecular cluster evaporation rates, enthalpies and entropies by Monte Carlo method"

September 9, 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 major revision stage of the review process for publication in Atmospheric Chemistry and Physics journal.

We wish to thank the Referee for their insightful comments which we feel substantially increased the quality of the manuscript. We believe that we have addressed all of the major and minor comments made by the reviewer and, in so doing, have produced a paper that is more rigorous in structure and more clear in presentation.

Next, in Section 2 we list the Referee’s comments. We also include our comment-by-comment responses. Each of the referee’s comments are denoted with “C” and our responses to the referee’s comments are denoted with “R”. At the end of the document we supply a marked-up version of the paper which contains a detailed comparison of the previous and revised versions of the manuscript.
2 Referee comments and our responses

Recommendation to the editor

1. Scientific significance
   Does the manuscript represent a substantial contribution to scientific progress within the scope of this journal (substantial new concepts, ideas, methods, or data)?
   Outstanding **Excellent** Good Fair Low

2. Scientific quality
   Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?
   Outstanding Excellent **Good** Fair Low

3. Presentation quality
   Are the scientific results and conclusions presented in a clear, concise, and well structured way (number and quality of figures/tables, appropriate use of English language)?
   Outstanding Excellent Good **Fair** Low

For final publication, the manuscript should be reconsidered after major revisions. I would be willing to review the revised paper, if the editor considers it necessary.

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

**Comment:** Like the way the authors wrote the paper in a casual way, they seem to respond to the referees’ comments in a similar way. It is hard to follow the response letter. There are so many errors, especially what were written in the letter are not the same as those in the revision. Although the authors addressed most of the comments and improvement has indeed seen in the revision, the authors will still need more efforts to improve the quality and the readability of the manuscript. There are still lot of errors/typos and those are really surprising. Below are some issues that need to be resolved before the paper can be publishable in ACP.

**Response:** Thank you for pointing out the typos, errors and inconsistencies that appear in the previous response letter. We strongly apologize for the inconvenience of reading and proof-checking our previous author responses and
tracking related manuscript changes. We further take your recommendations into account and thereby avoid confusions and inaccuracies here.

1. **Comment:** Section 2 is still hard to understand, although it is greatly improved after the revision. In addition, it is very lengthy and redundant. Would it be shortened to make it concise? Some of technical descriptions in my opinion can be moved to the Appendix or supplementary. In addition, the authors use a lot of very short paragraphs and the paper looks like a boring novel. Also, some languages used here are really awkward, given below are some examples:

**Response:** We have restructured and rewritten Section 2 to improve its quality and readability. We merged and reformulated the paragraphs to make the workflow more logical.

We first explain generation of synthetic data. Next, we place the section dedicated to Markov chain Monte-Carlo simulations which is subdivided into two parts: selection of minimum and maximum limits for unknown parameters, and overview of the MCMC runs. Both subsections have been made conciser. We have moved the technical details of Metropolis algorithm and its extended version (the DRAM method) to Appendix. These methods are given in A2 and A3, respectively. Additionally we reformulated the language in many of the sentences (see examples below). Here the lines from revised version of the manuscript are given in bold.

We have moved the first paragraph from “Discussions and future work” to Section 2.1 (“Generation of synthetic data”). Here we explain the sensitivity of MCMC parameter estimation to the quality and limitations of the synthetic data.

(a) **Comment:** Line 108, by the following method. You really mean by the following procedures or steps, right?

**Response:** Section 2.1 has been substantially rewritten. We removed this sentence from the text. Instead, we explain the origin of synthetic data as follows: ”We generated the birth-death equations using the ACDC code (McGrath et al., 2012), and then solved for the cluster concentrations using the Fortran ordinary differential equation solver VODE (N. Brown et al., 1989). ” See Lines 111-114.

(b) **Comment:** Line 118-119, what do you mean “each simulation was initialized with . . . and no sulphuric acid”? You mean “without
sulphuric acid”?
Response: We reformulated this part into ”...the initial sulfuric acid was set to zero in each simulation.” See Line 109.

(c) Comment: Line 122, what is “our particle system”? In particular, what is “particle” here?
Response: Thank you for highlighting this important issue which we believe makes our notions more consistent. Throughout the text, we have replaced the 'particle system' with 'the simulated system of clusters' or 'the set of molecular clusters considered here'. See, e.g., Lines 115, 122.

(d) Comment: Line 124, for time values less than the time at which..., do you mean “for time values before the system has attained the steady state”?
Response: Indeed, we intended to say “for time values before the system has attained the steady state”. We reformulated this part as follows: ”... measured at 1.5 min time intervals before the system reaches a steady state. This corresponded to a total of 41 time steps.” See Lines 123-124.

(e) Comment: Line 126-127, “with time resolution comprising 1.5 minutes”, do you mean “with a time interval of 1.5 minutes”?
Response: We reformulated this part, as mentioned above.

(f) Comment: Line 129, would it be “first...second”?
Response: Thank you for recommendation. We have changed the language accordingly. We explain two data sets generated for synthetic data (in Line 123). These are referred as ”...the fist set...” and ”In the second case...” (Line 125).

(g) Comment: Line 132, reached not reached to and the sentence “The measure of how close ...”is so complicated and awkward. Would it be modified for the sake of readers’ benefit?
Response: We modified this sentence as follows: ”Additionally, we include a convergence parameter for assessing the closeness of cluster concentrations to the steady state for every individual ACDC simulation.” See Lines 127-128.

(h) Comment: Line 228-229, please rearrange the sentence.
Response: As it was mentioned above, the discussion related to Metropolis algorithm was moved to Appendix. We changed the sentences which describe the initial assumptions for parameter values as follows: ”We first select the flat prior distribution from which we will initially sample unknown parameters, as we wish
to generate physically reasonable parameter estimates. Therefore, we generate unknown parameters within the chosen minimum and maximum bounds where all the points are equally likely to be sampled” (Lines 367-369).

**Comment:** It is strongly recommended that this section should be completely rewritten.

**Response:** The section had been restructured, rewritten and shortened in accordance with the advice of the Referee, as mentioned above.

2. **Comment:** The Results and Discussion section looks better than section 2. However, there are still some improvements need to be made. A lot of sentences are quite redundant and need to be modified for conciseness. For example, Line 293, adding “(Table 2) after concentrations will serve the purpose; you don’t need to say listed earlier in Table 1, “(Table 1) will be the same. Line 295, “the steady-state” steady-state here I believe is adjective. There are several throughout the manuscript. Line 300, An example of one of, is it necessary to include “one of” here?

**Response:** Following recommendations from the Referee, we restructured and rewrote Section 3. Below we summarize the main structural changes as well as some minor edits.

Initially, we reformulated Sections 3.1-3.3 and thus removed the redundancies. According to the Referee’s advise, we omit the references to Tabs. 1 and 2.

Next, we deleted Sections 3.4 and 3.5. The last section (i.e. ”Discussion and Future Work”) has been redistributed over the manuscript as follows. First, we have moved the text in Lines 436-449 (related to quality of the computer-simulated cluster concentrations) to Section 2.1 (”Generation of synthetic data”). See Lines 115-122.

Second, we inserted Lines 450-455 into Conclusion (Lines 311-317). This part explains two general principles of inverse problems/Bayesian parameter estimation applied in our study.

The final part of Section 3.5 (Lines 456-465), which explains the correlations between formation enthalpies and entropies is moved to Section 3.3 (Lines 288-297). Here we explain the results of parameter estimation after we expressed evaporation rates as parametrized functions of the temperature, with the cluster formation enthalpies and entropies as the unknown parameters. Naturally, we discuss why re-parametrization has improved the results, and the reasons why thermodynamic data display correlations.
Following advise from the Referee, we distinguish between the adjective "steady-state" and the noun "steady state" in the revised version of the manuscript.

The authors decided to keep the sentence in Line 300 as in the previous version ("An example of one of the sampled chains ...") We believe that the sentence is sufficiently comprehensive for the reader.

3. **Comment:** The ideas and the results are very interesting and the paper can benefit the community but the way the authors represent do really discouragement and the manuscript needs to be substantially improved in the next revision.

   **Response:** Thank you for acknowledging the advantages and benefits of modelling the type accomplished in this study. We made a considerable effort which we hope has helped to improve the quality of the presentation at the major revision stage. The changes are reflected in updated version of the manuscript.
Identification of molecular cluster evaporation rates, cluster formation enthalpies and entropies by Monte Carlo method

Anna Shcherbacheva¹, Tracey Balehowsky², Jakub Kubečka¹, Tinja Olenius³, Tapio Helin⁴, Heikki Haario⁴,⁵, Marko Laine⁵, Theo Kurtén⁶,¹, and Hanna Vehkamäki¹

¹Institute for Atmospheric and Earth System Research, P.O. Box 64 00014 University of Helsinki, Finland
²Department of Mathematics and Statistics, P.O. Box 64 00014 University of Helsinki, Finland
³Department of Environmental Science and Analytical Chemistry & Bolin Centre for Climate Research, Stockholm University, Svante Arrhenius väg 8, SE-11418 Stockholm, Sweden
⁴LUT School of Engineering Science, Lappeenranta-Lahti University of Technology, P.O.Box 20 FI-53851 Lappeenranta, Finland
⁵Finnish Meteorological Institute, P.O. Box 503, FI-00101 Helsinki, Finland
⁶Department 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 [..¹] synthetic cluster distributions generated by the Atmospheric Cluster Dynamic Code (ACDC) [..²] and based on gas kinetic collision rate coefficients and evaporation rates obtained using quantum chemical calculations [..³] and detailed balances. The studied system consisted of electrically neutral [..⁴] sulfuric acid and ammonia clusters with up to 5 of each type of molecules. We then treated [..⁵] the concentrations generated by ACDC as synthetic experimental data[..⁶]. With the assumption that the collision rates are known, we tested two approaches for estimating the evaporation rates [..⁷] from these data. First, we studied a scenario where [..⁸] time-dependent cluster distributions are measured at a single temperature before the system reaches a steady-state. In the second scenario, only steady-state cluster distributions are measured, but at several temperatures. [..⁹] Additionally, in the latter case the evaporation rates were represented in terms of cluster formation enthalpies and entropies[..¹⁰]. 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 previously published synthetic steady-state cluster concen-
tration data at one temperature [..11] and compared our two [..12] cases to this setting. Both the [..13] time-dependent and the two-temperature steady-state concentration data [..14] allowed us to estimate the evaporation rates with less variance than in the steady-state one temperature case.

We show that [..15] temperature-dependent steady-state data outperforms [..16] single-temperature time-dependent data for parameter estimation, even if only two temperatures are used. 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 17 [..18] . 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 [..19] acid-base clusters, negative charging tends to lead to a loss of base molecules, and positive charging to a loss of acid molecules (?). Modelling is thus needed to connect measured ion cluster distributions to the original neutral population.

Even when the atmospheric cluster [..20] distributions can be accurately deduced from experimental data, [..21] these distributions do not quantify the individual kinetic parameters, such as the cluster collision and evaporation rates (?). [..22] The 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 $\text{H}_2\text{SO}_4 – \text{H}_2\text{SO}_4$ 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[.23 ], and typically observed variations of up to several kcal/mol between [.24 ]the different applicable QC methods thus translate into orders of magnitude differences in evaporation rates [.25 ](??).

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 [.26 ]sulfuric acid concentration dependence of cluster concentrations (?), and even absolute particle formation rates (?) in [.27 ]sulfuric acid-ammonia and [.28 ]sulfuric 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 [.29 ]chain Monte Carlo (MCMC) methods. In a recent paper by ?, Differential Evolution (DE) MCMC (?) was applied to determine evaporation rates for negatively charged [.30 ]sulfuric acid and ammonia clusters (containing up to five of each type of molecules, with the HSO\(_\text{4}^-\) ion here defined as an "acid"). This study used steady-state cluster concentrations measured in the CLOUD Cosmics Leaving OUtdoor Droplets chamber experiment at constant temperature, with varying [.32 ]sulfuric acid and ammonia concentrations (we refer to ? for details relevant to the experimental data). [.33 ]The collision rates were computed from kinetic gas theory. ? concluded that these data were insufficient for estimation of all the evaporation rate coefficients. Another recent paper (?) reported thermodynamic data (cluster formation enthalpies and entropies) for 11 neutral [.34 ]sulfuric 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 [.35 ]sulfuric 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 (≥ 248 K).
In this study, we test which combinations of experimental data and fitted parameters lead to the best identification of 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. [..36] The use of synthetic data also allows us to know for sure if our inverse modelling actually produces the correct kinetic parameters[..37], 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 [..38]? 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 on the forward problem.

For simplicity, we consider the case of neutral [..39]sulfuric 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 [..40] three different scenarios for estimating evaporation rates. First, we use steady-state concentration measurements determined at a single temperature, similar to the approach used in ?. Next, 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 [..41] time-dependent data should provide additional information about the speed of the processes, which is missing from the steady-state data. [..42] Third, 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 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. Also, while ? used steady-state new-particle formation rates measured at different temperatures to fit their data, we use cluster concentrations.

---

36 removed: Use
37 removed: or not
38 removed: ?
39 removed: sulphuric
40 removed: two
41 removed: this transient
42 removed: Second
2 SIMULATION METHODS

2.1 Generation of synthetic data

We simulated the time evolution of cluster concentrations using collision rates computed from kinetic gas theory and evaporation rates computed from the Gibbs free energies reported by ?. To save computational time, we omitted clusters where the number of acid and base molecules differed by more than two. Based on both fundamental chemical principles and mass spectrometric data, these clusters are quite unstable, and thus have very high evaporation rates, leading to negligibly low concentrations. See Table 1 for a list of the considered clusters, 16 in total. We considered four different ammonia monomer mixing ratios between 5 and 200 ppt, corresponding to concentrations between $1.3 \times 10^8$ and $5.0 \times 10^9$ molecules per cm$^3$ for the temperature ranges studied here. In each individual case, the ammonia mixing ratio was kept constant throughout the simulation. The source rate of sulfuric acid monomer was kept constant at $Q = 6.3 \times 10^4$ cm$^{-3}$ s$^{-1}$. To reproduce experimental conditions in the CLOUD chamber as closely as possible, the initial sulfuric acid was set to zero in each simulation. See Table 2 for a summary of the concentration settings. Additionally, we considered the losses on the CLOUD chamber walls which depend on the cluster size and a dilution loss of $S = 9.6 \times 10^{-5}$ s$^{-1}$. For simplicity, we omitted the effect of relative humidity. We generated the birth-death equations using the ACDC code, and then solved for

removed: In this section we describe the methods used to create data sets for the synthetic cluster concentrations. We also explain parameter estimation by Markov Chain Monte Carlo which was used to obtain the evaporation rates and thermodynamic parameters from synthetic data.

removed: The 16 cluster types included in our study are summarized in Table 1.

removed: have excluded

removed: differs significantly from each other. Irrespective of the level of theory, quantum chemical data predict that these clusters will

removed: 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. The ratio is assumed to remain constant in each individual simulation, and varied

removed: (These correspond to concentrations of

removed: . (These correspond to concentrations of

removed: source rate is

removed: in all simulations

removed: the summary of ammonia mixing ratio and the source of sulphuric acid monomer used for the ACDC simulations.

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

removed: First, we computed the collision rates using the Eq. A3 from kinetic gas theory. Then, we used 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 ? into Eq. A5.

removed: consider

removed: computed with Eq. A2

removed: These values for the rates and losses were substituted into the ACDC algorithm, 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.
the cluster concentrations using the Fortran ordinary differential equation solver VODE (\textsuperscript{59}). These equations and all related parameters are explained in Appendix A1.

\textsuperscript{60} Our MCMC results are not specific to the set of molecular clusters considered here. This is supported by the fact that although the size of the system (the number of clusters, or more precisely the maximum size of the clusters, included in the simulations) has an impact on the particle formation rates at high temperatures (> 278 K), the particle formation rates and cluster concentrations produced using different cluster sets (e.g. 4x4, 5x5 and 6x6 sulfuric acid and ammonia molecules) are qualitatively similar (\textsuperscript{59}). Thus, minor changes of the ACDC outputs due to the difference in the sets of considered clusters should not change the MCMC parameter estimation results. Additionally, the boundary conditions for the outgrowing clusters (the choice of the clusters that are considered as formed particles) have only minor influence on the simulation results, as long as the simulated system of clusters is defined in a reasonable way (\textsuperscript{59}).

\textsuperscript{61} Two data sets were created. In the first set, we generated time-dependent concentrations for each cluster type, measured at 1.5 min time intervals before the system \textsuperscript{62} reaches a steady state.

\textsuperscript{63} This corresponded to a total of 41 time steps. The steady-state single-temperature data correspond to a subset of these data set. In the second case, we generated steady-state concentrations for all cluster types at two temperatures (278 \textsuperscript{66} and 292 K). In both cases, the steady-state cluster concentrations \textsuperscript{67} were calculated as the average of the concentrations \textsuperscript{68} at \( t_1 := 50 \text{ min} \) and \( t_2 := 60 \text{ min} \). \textsuperscript{69} Additionally, we include a convergence parameter for assessing the closeness of cluster concentrations to the steady state \textsuperscript{61} for every individual ACDC simulation. This is computed as a ratio of concentrations taken at times \( t_2 \) and \( t_1 \) \textsuperscript{70} in each case for the cluster for which this ratio deviated most from unity \textsuperscript{71}.

\textsuperscript{59} removed: A detailed description of this strategy for solving the forward-problem of finding the cluster concentration rates from Eq. A1 was published in \textsuperscript{59}. 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.

\textsuperscript{60} removed: The above method we used for producing synthetic concentration rates is similar to the one described in \textsuperscript{59}. We note that unlike \textsuperscript{59}, in this paper, our particle system is considered at various temperatures.

\textsuperscript{61} removed: Using the above algorithm, model configuration and parameters, we generated two data sets. First, time evolution of the concentrations \( Y_i (t) \) 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

\textsuperscript{62} removed: concentration data for each of the cluster types \( i \) measured from beginning to the end of each simulation,

\textsuperscript{63} removed: has attained

\textsuperscript{64} removed: Secondly, we solve for time-independent

\textsuperscript{65} removed: the cluster types for two temperatures comprising

\textsuperscript{66} removed: K

\textsuperscript{67} removed: data configurations

\textsuperscript{68} removed: are

\textsuperscript{69} removed: determined for time instances

\textsuperscript{70} removed: The measure of how close the system has reached

\textsuperscript{71} removed: is monitored by a convergence parameter, which is the ratio of the concentrations

\textsuperscript{72} removed: , taken

\textsuperscript{73} removed: ,
Finally, we added measurement error (noise) to the cluster concentrations in both data sets. We call the resulting noisy cluster concentrations synthetic data. Our measurement error was sampled from a multivariate Gaussian distribution, with the variance depending on cluster type $i$, temperature $T$, and time instance $t$. We assume that the standard deviation of the measurement error is 0.001% of the original concentration.

Table 1. Neutral molecular clusters included in the model system (16 in total). The first column indicates the number of sulfuric acid molecules, the second column stands for the number of ammonia in the cluster.

<table>
<thead>
<tr>
<th>Number of H$_2$SO$_4$ molecules</th>
<th>Number of NH$_3$ molecules</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0-1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0-2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1-3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2-5</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3-5</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. Monomer concentrations used in simulations

<table>
<thead>
<tr>
<th>[H$_2$SO$_4$] monomer source</th>
<th>[NH$_3$] concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$6.3 \times 10^4$ cm$^{-3}$ s$^{-1}$</td>
<td>5 ppt</td>
</tr>
<tr>
<td>$6.3 \times 10^4$ cm$^{-3}$ s$^{-1}$</td>
<td>35 ppt</td>
</tr>
<tr>
<td>$6.3 \times 10^4$ cm$^{-3}$ s$^{-1}$</td>
<td>100 ppt</td>
</tr>
<tr>
<td>$6.3 \times 10^4$ cm$^{-3}$ s$^{-1}$</td>
<td>200 ppt</td>
</tr>
</tbody>
</table>

2.2 Markov chain Monte-Carlo simulations

removed: In both data settings, the simulation outputs are amended with the measurement errors
removed: non-correlated,
removed: where the variance of the distribution depends
removed: $T$
removed: $t$. While a simplification of noise characteristics of the real data obtained from a mass spectrometer, we impose
removed: noise comprises
removed: Note that apart from generation of synthetic data, we apply the ACDC as a kinetic model of cluster population in the MCMC simulations. The ACDC outputs are compared to the synthetic measurements and explained in Section 2.2.
We used a Markov Chain Monte Carlo (MCMC) [83] based approach to estimate the evaporation rates which reproduce the synthetic cluster concentration data. Unlike optimization algorithms [86] which compute a single optimal parameter set, MCMC methods sample from a target distribution which contains the most-likely combinations of parameter values [87] for the given data. Multiple samples of possible parameter sets are taken along a random walk in the target distribution, and are saved as a parameter chain. As the length of the chain increases, the sampled sets converge to a probability (posterior) distribution of parameters, which estimates the likelihood of those parameters giving rise to the data. [88]

2.2.1

The evaporation rate coefficients $\gamma_{i,j} \rightarrow \gamma_{i,j}$ appearing in the ACDC simulation of
removed: 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 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
removed: sampling.

The objective of MCMC in parameter estimation is to identify possible parameter values which yield the best fit with the experimental
removed: that produce one best combination of parameter values, in the MCMC procedure all the most-probable
removed: are estimated given the
removed: 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 which gives the likelihood of observing each of the parameters given the concentration data.
removed: The Metropolis algorithm
removed: First, a prior distribution for the parameter values $\theta$ (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 $\theta_0$ or $\theta_{old}$) is selected from the prior distribution.

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

removed: Then the candidate point $\theta_{new}$ is either accepted or rejected, according to the least-squares fit of the output to the data, which measures the difference between the modelled $Y_{mod}$ and measured $Y_{exp}$ cluster concentrations:
removed: 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 $\sigma_i^2$ to avoid overfitting to the larger concentration values. The error variance $\sigma_i^2$ is matched depending on cluster type, time instance and temperature. See A2 for more details.

96 removed: At each iteration of the Metropolis algorithm, the value $F(\theta_{\text{new}})$ is compared to the least-square sum from the previous step $F(\theta_{\text{old}})$. 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 $F(\theta_{\text{new}}) > F(\theta_{\text{old}})$, the point will be accepted with the probability

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

99 removed: 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 

100 removed: introduced in ?. We briefly explain our approach below.

101 removed: The DRAM algorithm

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

103 removed: We make our initial guess $\theta = \theta_{\text{old}}$, where $\theta_{\text{old}}$ 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.

104 removed: Once initialized, the following iterative steps take place. From the likelihood probability distribution for $\theta_{\text{old}}$, a new candidate for the unknown parameter values, $\theta_{\text{new}}$, 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 $\theta_{\text{new}}$. In the first stage of DRAM, we chose to accept the new proposed values $\theta_{\text{new}}$ with probability

108 removed: where $Y_{\text{exp}}$ is the array of synthetic cluster concentration data, and $p(Y_{\text{exp}}|\theta_{\text{old}})$, $p(Y_{\text{exp}}|\theta_{\text{new}})$ denote the likelihood (conditional) probabilities for the old and new parameter values, respectively. These likelihood probabilities quantify how closely the kinetic model with parameters $\theta$ reproduce the data, as they depend on the sum of squared residuals (see Eqs. A8 and A7) between the given data and the concentrations obtained from the ACDC and VODE simulations with parameters $\theta_{\text{old}}$ and $\theta_{\text{new}}$, respectively. This relationship is explained further in Appendix A1.

109 removed: In DRAM we allow for partial modification of the proposed parameters (the "delayed rejection" component of DRAM). This second stage of sampling improves the computational time needed to obtain an estimate for $\theta$; it is performed as follows. If the proposed $\theta_{\text{new}}$ is rejected, a nearby proposal is created, $\theta_{\text{new2}}$. We accept this second proposal keeping in mind the rejection probability of the first, according to
which is an extended variant of the classical Metropolis algorithm \(^?\). We chose the DRAM algorithm as it is more efficient than the Metropolis regime at parameter estimation when the parameter space is large \(^?\). The two algorithms and their application to our cases are described in the Appendix.

### 2.2.1 Selection of minimum and maximum limits for unknown parameters

Simultaneously, the second-stage proposal covariance is computed as a scaled version of the first-stage proposal covariance: \(\gamma = 5\) borrowed from \(^?\). This value was chosen to increase the acceptance at the second stage.

Then, if both \(\theta_{old}\) and \(\theta_{new}\) 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.

Parameter estimation is conducted using the

123 removed: 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. ??-3).

\(^{109}\) removed: 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 \(p(Y_{\text{exp}} | \cdot)\) and \(p(Y_{\text{exp}} | \cdot, \cdot)\) are Gaussian. They are updated at each successive iteration of the MCMC algorithm to improve the mixing of the chains.

\(^{110}\) removed: first-stage proposal covariance is recomputed via the Adaptive Metropolis (AM) procedure \(^?\). Let \(d\) be the dimension of the parameter space, and \(\{X_0, \ldots, X_n\} \subset \mathbb{R}^d\) 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 \(X_n\), whereas the corresponding proposal covariance \(C_{1n}\) is updated using the path of the previously sampled MCMC chain:

\[ s_d = 2.4/d \] is the scaling parameter, and \(\text{Cov}(X_0, \ldots, X_{n-1})\) is the empirical covariance matrix for the vectors \(X_0, \ldots, X_{n-1}\):

\[^{112}\] removed: 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,

\[^{114}\] removed: where \(\overline{X}_{n-1}^T = \frac{1}{n} \sum_{i=0}^{n-1} X_i\) and \(X_i \in \mathbb{R}^d\) are column vectors. In our study and all runs therein, we set \(n_0\) to be 100 iterations.

\[^{115}\] removed: Simultaneously, the second-stage proposal covariance is computed as a scaled version of the first-stage proposal covariance:

\[^{117}\] removed: with the scaling factor \(\gamma = 5\) borrowed from \(^?\). This value was chosen to increase the acceptance at the second stage.

\[^{118}\] removed: Then, if both \(\theta_{old}\) and \(\theta_{new}\) 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.

\[^{119}\] removed: Parameter estimation is conducted using the

\[^{120}\] removed: 'mcmcestat'

\[^{121}\] removed: toolbox implemented for FORTRAN \(^?\). See the description and the examples of usage on the web page

\[^{122}\] removed: .

\[^{123}\] removed: Overview of the MCMC runs

\[^{124}\] removed: 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. ??-3).
We emphasize that there are currently no theoretical principles or experimental results which set sound restrictions for even the order of magnitude of the evaporation rates. However, [\[125\] \text{evaporation rates much lower than } 10^{-10} \text{s}^{-1} \text{ are irrelevant in}\] practice, since the timescale for evaporation is then much longer than the cluster lifetime with respect to further growth. Similarly, when the evaporation rate is [\[128\] \text{much greater than } 10^{+10} \text{s}^{-1}\] the cluster will certainly evaporate before it has a chance to grow further. [\[131\] \text{The base 10 logarithm of the evaporation rates was } \text{therefore sampled in the interval of } -12 \text{ to } 12.}

For the cluster formation enthalpies, we chose an upper limit of 0 kcal/mol, as a positive $\Delta H$ would mean an absence of attractive interactions in the molecular cluster, which is physically incorrect for polar, H-bonding molecules such as $\text{H}_2\text{SO}_4$ and $\text{NH}_3$. [\[134\] \text{This same argument also applies for each individual molecule, which gives rise to the requirement that the formation enthalpy of each cluster must be lower (more negative) than that of clusters with less acid and/or base molecules. See Table 3 for the full list of restrictions arising from this requirement. As a lower limit for the overall cluster formation enthalpies, we used } \Delta H = -400 \text{ kcal/mol}. [\[136\] \text{As our largest clusters contain 10 molecules, this would imply that, on average, each } \text{H}_2\text{SO}_4 \text{ in all the studied clusters is bound substantially stronger than in the exceptionally strongly bound } \text{HSO}_4^- \text{H}_2\text{SO}_4 \text{ cluster}, [\[137\] \text{(for which recent high-level computational studies indicate a binding enthalpy roughly around } -40 \text{ kcal/mol, (??)}\] ]. \text{This in turn implies that the evaporation rate is zero for all } \text{(practical purposes).}

\[140\] \text{The upper limit for the formation entropies was set to 0 cal/K/mol, as clustering must have a negative } \Delta S, \text{ since the number of gas molecules is reduced (and translational and rotational degrees of freedom are thus converted into much more constrained vibrational degrees of freedom).} \text{[\[139\] \text{The lower limit of } -400 \text{ cal/K/mol} \text{ can be justified by}\]}

---

\[125\] removed: indicate possible
\[126\] removed: we assume that the evaporation rates with orders of magnitude less
\[127\] removed: practise, since such an evaporation event is highly improbable, and it is very likely that instead the cluster will grow further by collisions
\[128\] removed: of the order of magnitude more
\[129\] removed: it is reasonable to expect that
\[130\] removed: most
\[131\] removed: With these assumptions, the prior distribution of the evaporation rates spans over several orders of magnitude, and the
\[132\] removed: sampled from the range
\[133\] removed: Next, we justify the limits selected for data setting 2, where we sample thermodynamic parameters. For the formation enthalpies
\[134\] removed: is chosen by the fact that
\[135\] removed: For the lower limit (\)
\[136\] removed: ) we mean that on average
\[137\] removed: for which the most recent
\[138\] removed: . 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 $\text{H}_2\text{SO}_4$ and 5 $\text{NH}_3$, 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
\[139\] removed: purposes of measurement (??)
\[140\] removed: Next, we set the
\[141\] removed: since molecule
\[142\] removed: as
\[143\] removed: For the
\[144\] removed: , we state
noting that the typical per-molecule $\Delta S$ for clustering is around $-30$ cal/K/mol, with a typical variation of up to $\pm 10$ cal/K/mol \cite{145}. For a 10-molecule cluster this would imply a lower bound to $\Delta S$ of around $-400$ cal/K/mol. \cite{147}.

Table 3. Additional restrictions on the cluster formation enthalpies arising from the requirement that each individual molecule is bound The cluster formation enthalpy of the $i$-th cluster is denoted by $\Delta H_i$. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.

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

\footnotesize
\cite{145}removed: So for the largest clusters the upper limit corresponds to a per-molecule
\cite{146}removed: -40
\cite{147}removed: In this situation, all the new vibrational degrees of freedom formed in the product clusters are quite rigid, i.e. have very low entropy (?).
\cite{148}removed: h!
\cite{149}removed: Domain limitations for two data settings under consideration imposed to exclude non-physical parameters in parameter estimation procedure.
\cite{150}removed: Data settings
\cite{151}removed: Estimated parameters
\cite{152}removed: Minimal value
\cite{153}removed: Maximal value
\cite{154}removed: Data setting 1
\cite{155}removed: Base 10 logarithms of
\cite{156}removed: -12
\cite{157}removed: 12
\cite{158}removed: evaporation rates (in s$^{-1}$)
\cite{159}removed: Data setting 2
\cite{156}removed: Cluster formation
\cite{151}removed: enthalpies (kcal mol$^{-1}$) and
\cite{152}removed: -400
\cite{153}removed: 0
\cite{154}removed: entropies (cal K$^{-1}$ mol$^{-1}$)
\cite{156}removed: -400
\cite{157}removed: 0
2.2.2 Overview of the MCMC runs

We first performed DRAM parameter estimation from both steady-state and time-dependent cluster concentrations at 278 K, treating evaporation rates as the unknown parameters \( \theta \). For the time-dependent synthetic data, the number of output coefficients was \( n_{\text{out}} = N_C \times N_t + 1 \), where \( N_C = 16 \) is the number of cluster types included into simulations, and \( N_t = 41 \) is the number of time-step measurements available for each of the cluster types.

Next, we performed parameter estimation based on steady-state cluster concentrations at two temperatures, 278 K and 292 K. The number of output coefficients in this case was \( n_{\text{out}} = (N_C + 1) \times N_T \), where \( N_T = 2 \) denotes the number of experiments conducted at different temperatures. We use Eq. A4 and A5 to express the evaporation rates as functions of formation enthalpies, entropies and temperature:

\[
\gamma_{i+j \rightarrow i,j} = f(T, \{ \Delta H_k, \Delta S_k \}_{k \in \{i+j,i,j\}}).
\]

In Eq. 1, we set \( T = 278 \) K or \( T = 292 \) K. We emphasize that the rates \( \gamma_{i+j \rightarrow i,j} \) now depend on temperature and six other parameters: the formation enthalpy \( \Delta H_{i+j} \) and entropy \( \Delta S_{i+j} \) of the evaporating/fragmenting cluster \( i+j \), and the formation enthalpies \( \Delta H_i, \Delta H_j \) and entropies \( \Delta S_i, \Delta S_j \) of the product clusters \( i \) and \( j \) respectively. In this setting \( \theta \) represents the array of quantities \( \Delta H_{i+j}, \Delta S_{i+j}, \Delta H_i, \Delta H_j, \Delta S_i, \Delta S_j \) with \( i+j \in \{1,2,\ldots,16\} \). Similar approaches were applied for the inverse problem of chemical kinetics modelled by the Arrhenius equation, where chemical reaction rates are temperature-dependent (?).

Many evaporation/fragmentation reactions have the same clusters as products, and thus several of the pairs \( \Delta H_i, \Delta S_i \) appear in Eq. 1 for the evaporation rates of multiple different reactant clusters. The formation enthalpies and entropies of monomers are defined in the context of molecular clustering to be zero. The number

---

\(^{172}\)removed: An outline of the sampling procedure is illustrated in Figure 1 below.
\(^{173}\)removed: Schematic representation of the study methods.
\(^{174}\)removed: next explicitly describe what synthetic data (\( Y_{\text{exp}} \)) and parameters ( \( \theta \) which give the acceptance probability in
\(^{175}\)removed: represent in the two study cases.
\(^{176}\)removed: In the first study, the free parameters \( \theta \) represent the evaporation rates. The data \( Y_{\text{exp}} \) is either the time-independent
\(^{177}\)removed: or transient cluster concentrations measured at temperature
\(^{178}\)removed: .
\(^{179}\)removed: In the second study, we
\(^{180}\)removed: thermodynamic data, parametrized by
\(^{181}\)removed: cluster
\(^{182}\)removed: At either temperature \( T = 278 \) K or \( T = 292 \) K, the smaller clusters for certain combinations of ammonia and sulphuric acid may arise from the evaporation of several larger clusters. This implies that
\(^{183}\)removed: expression
\(^{184}\)removed: different cluster types. Additionally, the Gibbs formation free energies
\(^{185}\)removed: fixed
\(^{186}\)removed: , and their associated
of distinct unknown formation enthalpies and entropies \([..^{188}]\), is thus only 28\([..^{189}]\), compared to 39 unknown evaporation rates. Furthermore, the cluster formation entropy and enthalpy values all lie within two orders of magnitude\([..^{190}]\)\([..^{191}]\)\([..^{192}]\)\([..^{193}]\)\([..^{194}]\), compared to the evaporation rates which span 24 orders of magnitude. This makes the MCMC method more efficient.

\([..^{195}]\) To create a reliable sample from the underlying parameter distribution, the length of the MCMC chain must be \([..^{196}]\)“large enough”\([..^{197}]\); that is, many different parameter combinations must be tested. \([..^{197}]\) In our simulations, the MCMC chain length typically comprised \([..^{198}]\)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 \([..^{199}]\)forward ACDC model (in which the evaporation and collision rates are known) is deterministic.

In \([..^{200}]\) the MCMC simulations, all sets of parameters which produce cluster concentrations within the allotted noise level of the data \((0.001\%)\) are kept in the chain. \([..^{201}]\) The sampling procedure is outlined in Figure 1 below. We tested that the MCMC chains converge to the ‘true’ values (i.e., the reference parameter values from \([..^{202}]\)) when we start sampling the chain from randomly selected initial guess.

3 Results and Discussion

3.1 Identification of \([..^{203}]\) evaporation rate coefficients from steady-state data at a single temperature

\(^{188}\) removed: do not vary in our simulations. This imposes additional constraints on possible parameter values. One can calculate that of the 39 evaporation rates that are involved in the dynamics of the neutral cluster system under consideration, \(^{189}\) removed: 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 \(^{190}\) removed: This feature of the cluster formation entropies and enthalpies has the effect of reducing the \(^{191}\) removed: stiffness

\(^{192}\) removed: of the differential system in

\(^{193}\) removed: (computed via ACDC) which allows for easier integration via VODE.

\(^{194}\) removed: For the setting above, the data \(Y_{\text{exp}}\) 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 \([..^{195}]\). 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 \([..^{196}]\)

\(^{195}\) removed: Note that to

\(^{196}\) removed: “large enough” in an appropriate sense \([..^{197}]\).

\(^{197}\) removed: We remark here that in both our studies

\(^{198}\) removed: of

\(^{199}\) removed: “forward”

\(^{200}\) removed: all simulations of the algorithm given in the previous section,

\(^{201}\) removed: 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

\(^{202}\) removed: the
Figure 1. Schematic representation of the study methods.

A graphical representation of the steady-state cluster concentration data at 278 K, as a function of the number of acid molecules in the clusters, is given in Figure 2.

Next, we determine the base 10 logarithms of the evaporation rate coefficients from the synthetic data. Since the noise added to the 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 this random bias, we generated 3 sets of synthetic data by adding random increments to the 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 Figs. B1-B2. We omit the initial one million samples, and plot the stationary parts of the chains. As we observe from the plots in Figs. B1-B2, all the parameter chains 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)}
Figure 2. Steady-state cluster concentrations for the clusters containing sulfuric acid and a varying number of ammonia molecules, as a function of the number of acid molecules, for [NH$_3$] mixing ratios of (a) 5 ppt, (b) 35 ppt, (c) 100 ppt and (d) 200 ppt at the 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 rate of sulfuric acid monomers is $[\text{H}_2\text{SO}_4] = 6.3 \times 10^4$ s$^{-1}$. 

Notably, all monomer evaporation rates are bounded from below, except for some of the rates from the largest clusters: $\text{H}_2\text{SO}_4$ from $(\text{H}_2\text{SO}_4)_5(\text{NH}_3)_4$ and $(\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5$, and NH$_3$ from $(\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5$. For each evaporation rate, 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 Figs. 3-4. 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 removed: This subset of evaporation parameters is comprised of the evaporation rates of monomers, with the exception of removed: for removed: the evaporation rate of removed: These excluded parameters correspond to the evaporation 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.
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 Figs. 3-4), as their estimated variances are well within our accepted error range of less than 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, most of the marginal posterior distributions are non-uniform, except for the evaporation rate of \((\text{H}_2\text{SO}_4)\)\(_2\)\((\text{NH}_3)\)\(_2\) and \((\text{NH}_3)\)\(_5\). In five cases (refer to subfigures labelled 6, 21, 28, 32 and 36 in Figs. 3-4), the estimated parameter values are not unique: the marginal posterior distributions feature multiple modes. The results of our parameter estimation are summarized in Tabs. C1- C2 and in subfigures labelled (a) and (b) in Figure 5.

The majority of the parameters are not correlated. However, the evaporation of monomers from \((\text{H}_2\text{SO}_4)\)\(_3\)(\text{NH}_3), \((\text{H}_2\text{SO}_4)\)\(_2\)(\text{NH}_3)\(_2\) and \((\text{H}_2\text{SO}_4)\)\(_5\)(\text{NH}_3)\(_4\) display non-linear inverse correlations. This implies that either \(\text{H}_2\text{SO}_4\) rarely evaporates (at a rate less than \(10^{-4}\) s\(^{-1}\) and that \(\text{NH}_3\) evaporates often, or that the evaporation rates of \(\text{H}_2\text{SO}_4\) and \(\text{NH}_3\) are of comparable magnitude. Additionally, it can be seen from the pairwise posteriors that most of the estimated parameters are highly uncertain.

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. The general solution to this issue is to regularize the problem, either by adding more data or information to the model, or by reducing the number of possible estimates.

Based on parameter estimation results, we conclude that a single-temperature steady-state cluster concentrations are not enough to estimate the evaporation rates with a reasonable accuracy (i.e., to obtain an upper and lower limits for the rates that reasonably restrict the cluster kinetics involved in the molecular-level process).
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 [228] used to generate the synthetic data. [229] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..231] used to generate the synthetic data. [..232] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.

3.2 Identification of [..234] evaporation rate coefficients from [..235] time dependent data at a single temperature [..236]

The data set for [..237] time-dependent cluster concentrations is much larger than the data set for steady-state cluster concentrations[..238], as it contains the concentration values at multiple [..239] time instances. The time-dependent data also contain information about the [..240] time derivatives of the concentrations[..241], (see C1.), which [..242] should contribute to quantification of [..243] kinetic parameters (in this case evaporation rates). Our time-dependent cluster concentration data

---

235 removed: transient
236 removed: First, we extend the synthetic measurement data from steady state concentrations to transient concentrations.
237 removed: transient cluster concentrations at one temperature is
238 removed: at one temperature, as the transient data
239 removed: times instances. Also the transient data
240 removed: slope
241 removed: changing with time
242 removed: contributes
243 removed: 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
sets contain in total 656 concentration measurements (corresponding to 16 cluster types and 41 timesteps), for each of the four ammonia mixing ratios.

From this transient 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 Figs. C2-C3. Again, we omit the first one million samples and merge the stationary parts of the sampled chains to obtain the posterior distributions.

As seen in Figs. C2-C3, all the chains have upper limits. Most of the chains are also bounded from below, with five exceptions. These exceptions, with arbitrarily large magnitudes, are the evaporation rates of \((\text{H}_2\text{SO}_4)_{2}(\text{NH}_3)_{2}\) from \((\text{H}_2\text{SO}_4)_{4}(\text{NH}_3)_{4}\) and \((\text{H}_2\text{SO}_4)_{5}(\text{NH}_3)_{5}\), and the evaporation rates of \(\text{H}_2\text{SO}_4\), \((\text{H}_2\text{SO}_4)(\text{NH}_3)\) and \((\text{H}_2\text{SO}_4)_{2}(\text{NH}_3)_{2}\) from \((\text{H}_2\text{SO}_4)_{5}(\text{NH}_3)_{5}\).

The one-dimensional marginal posterior distributions for the estimated parameters are shown in Figs. 6-7. Most of the estimates are close to the “true” values used in the generation of the synthetic data. However, the estimated evaporation rates 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 Figs. 6-7).

The evaporation rate of \((\text{H}_2\text{SO}_4)_{2}(\text{NH}_3)_{2}\) from \((\text{H}_2\text{SO}_4)_{5}(\text{NH}_3)_{5}\) (which corresponds to subfigure 26) has a uniform posterior distribution, corresponding to an enormous uncertainty. Further, for the evaporation rates depicted in subfigures 20 and 36, we can only determine upper limits of less than \(1.96 \times 10^{-5} \text{ s}^{-1}\). However, the time-dependent data allows us conclude that the evaporation processes \((\text{H}_2\text{SO}_4)_{4}(\text{NH}_3)_{4} \rightarrow (\text{H}_2\text{SO}_4)_{4}(\text{NH}_3)_{2} + \text{NH}_3\) and \((\text{H}_2\text{SO}_4)_{5}(\text{NH}_3)_{5} \rightarrow (\text{H}_2\text{SO}_4)_{4}(\text{NH}_3)_{5} + \text{H}_2\text{SO}_4\) can be neglected, as they are relatively slow compared with competing evaporation processes.

\[^{244}\text{removed: set using the method in Section 2.1. The time resolution of our new synthetic data set is 1.5 minutes, which results in 656 total concentration measurements for all the cluster type measured for four different ammonia concentrations. These data sets are illustrated in C1.}\]

\[^{245}\text{removed: transient}\]

\[^{246}\text{removed: analogous MCMC runs (}\]

\[^{247}\text{removed: )}\]

\[^{248}\text{removed: , which are the samples before the chains have obtained their stationary}\]

\[^{249}\text{removed: It is shown}\]

\[^{250}\text{removed: that}\]

\[^{251}\text{removed: the}\]

\[^{252}\text{removed: Specifically,}\]

\[^{253}\text{removed: H}_2\text{SO}_4\text{NH}_3\]

\[^{254}\text{removed: have arbitrarily large magnitude}\]

\[^{255}\text{removed: We examine the}\]

\[^{256}\text{removed: From these plots, one sees that most}\]

\[^{257}\text{removed: baseline values used for}\]

\[^{258}\text{removed: parameters}\]

\[^{259}\text{removed: Three parameters (subfigures 20, 29 and 36 in Figs. 6-7) have multimodal marginal posterior distributions. We also note that the}\]

\[^{260}\text{removed: we can only specify that the upper limits}\]

\[^{261}\text{removed: are}\]

\[^{262}\text{removed: given the reliable upper estimates,}\]

\[^{263}\text{removed: when compared with the other competing}\]
Pairwise marginal posterior distributions for the evaporation rates are plotted in Figs. C4-C8. 

Most of the evaporation rates do not display substantial correlations. However, the evaporation rates of monomers from the cluster \((\text{H}_2\text{SO}_4)_2\text{NH}_3\) display a strong inverse linear relationship, indicated by the pairwise marginal posterior distribution of the coefficients

\[
(\text{H}_2\text{SO}_4)_2\text{NH}_3 \rightarrow (\text{H}_2\text{SO}_4)_2 + \text{NH}_3 \quad \text{and} \quad (\text{H}_2\text{SO}_4)_2\text{NH}_3 \rightarrow \text{H}_2\text{SO}_4\text{NH}_3 + \text{H}_2\text{SO}_4,
\]

(see Figure C4). Also, the estimated rate coefficients \((\text{H}_2\text{SO}_4)_2 \rightarrow \text{H}_2\text{SO}_4 + \text{H}_2\text{SO}_4\) and \(\text{H}_2\text{SO}_4\text{NH}_3 \rightarrow \text{H}_2\text{SO}_4 + \text{NH}_3\) exhibit linear correlation. Additionally, the uncertainties in all the correlated parameters are relatively small (less than an order of magnitude).

In Tabs. C1-C2 we summarize the results of parameter estimation for the two data settings (steady-state and time-dependent) at a single temperature. Note that the estimated upper limits for some of the small evaporation rates (less than \(10^{-5} \text{ s}^{-1}\)) determined from the steady-state data can be as large as \(1.55 \times 10^{-2} \text{ s}^{-1}\). 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 7. In these cases, the identification is improved when we extend the data set with time-dependent measurements. Overall, the time-dependent data enabled us to determine the lower bounds for most of the parameters, with the exception of the 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 5).

\[ \]
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. [277] Here blue asterisks denote the baseline values used for creating the synthetic data (borrowed from [277]). Black circle and horizontal line markers indicate the mode and the mean value of the distribution, respectively. The notation $x$-$y$-$N$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 time-dependent 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. The notation $xA_yN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..\textsuperscript{282}] time-dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from [..\textsuperscript{283}] used to generate the synthetic data. [..\textsuperscript{284}] The notation \( xA_yN \) corresponds to a cluster with \( x \) sulfuric acid and [..\textsuperscript{285}] \( y \) ammonia molecules.

3.3 Estimating [..\textsuperscript{286}] formation enthalpies and entropies from steady-state concentration measurements at multiple temperatures [..\textsuperscript{287}] We determined [..\textsuperscript{288}] cluster formation enthalpies and entropies [..\textsuperscript{289}] based on two sets of [..\textsuperscript{289}] steady-state cluster concentrations, [..\textsuperscript{290}] corresponding to two temperatures: 278 and 292 K. [..\textsuperscript{291}] These data sets are plotted in Figs. 2 and D1 for 278 K and 292 K, respectively.

\[286\text{ removed: thermodynamic data} \]
\[287\text{ removed: In this section we describe another method for regularizing our problem of estimating evaporation rates from steady-state concentration data.} \]
\[288\text{ removed: from} \]
\[289\text{ removed: synthetic,} \]
\[290\text{ removed: now measured at} \]
\[291\text{ removed: This data set is} \]
\[292\text{ removed: We will demonstrate that reparameterization (in terms of thermodynamic data) plus the extended data set transforms our parameter estimation problem from an ill-posed problem to a well-posed one. We use synthetic steady-state cluster concentrations generated for two temperatures to recover the} \]
As in the previous sections, three MCMC runs were conducted to average the bias attributed to random noise. An example of the sampled chains is shown in Figure D2. It can be seen that all the chains are bounded, with the exception of the formation enthalpy and entropy of the largest cluster \((\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5\).

The one-dimensional marginal posterior distributions of the formation enthalpies and entropies, built from the stationary parts of the three sampled chains merged together, are shown in Figure 9. For all the clusters except \((\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5\), the variances of 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\(^{-1}\). The estimated free parameters together with the "true" quantum chemistry-based values from ? used for generation of the synthetic data are summarized in Table D1.

Although the posterior distributions of the formation enthalpies and entropies of \((\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5\) feature higher uncertainties in comparison to those of the smaller clusters, the evaporation rates from \((\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5\) as calculated from the aforementioned posterior distributions, have low variances, see Table D3.

Additionally, strong correlations are observed between formation enthalpies and entropies of clusters containing the same number \(n\) of ammonia molecules when \(n > 2\), except the case of \((\text{H}_2\text{SO}_4)_5(\text{NH}_3)_5\). thermodynamic parameters. This is done to improve the identification by using the temperature dependence of the Gibbs free energies (and the evaporation rates).
3.4  

used to generate the synthetic data. Here the symbols $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote $H_2SO_4$ and "NH$_3$", correspondingly.

removed: One-dimensional marginal posterior distributions of the cluster formation enthalpies (units given in kcal/mol) and entropies (units given in 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 $\tau$ used to generate the synthetic data. Here the symbols $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. Symbols "A", "N" denote $H_2SO_4$ and "NH$_3$", correspondingly.

removed: Comparison to previous evaporation rate determinations

removed: 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.

removed: 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.

removed: 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.

removed: 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.

removed: Discussion and future work

removed: 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 (?).

removed: 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
These strong correlations are consistent with general principles of clustering thermodynamics. If a cluster has very strong bonds between its constituent molecules, then the formation enthalpy is very negative, and also the intermolecular vibrational frequencies corresponding in a broad sense to vibrations involving those bonds are fairly high, meaning that the entropy loss in forming the cluster is large. These intermolecular frequencies dominate the "variable part" of the formation entropy, as the entropy change from the loss of translational and rotational degrees of freedom is almost a constant factor. Thus, if the formation enthalpy of a cluster 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) formation entropy.

Evaporation rates for all the molecular clusters calculated from a posterior distribution of sampled formation enthalpies and entropies are close to the "true" values used for generation of the synthetic data at both temperatures (278 K and 292K) and their variances are less than one order of magnitude, see Figs. D6-D7. Thus, reparametrization of evaporation rates in terms of formation enthalpies and entropies, and use of data at two different temperatures, thus transforms our parameter estimation problem from an ill-posed to a well-posed one.

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.

removed: 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.

removed: 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

removed: the
removed: that means
removed: (note that these
removed: effect
removed: ) are fairly high, meaning that the entropy loss in forming the cluster is large. So
removed: of course
removed: (?).

removed: 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.
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 $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
Figure 9. One-dimensional marginal posterior distributions of the cluster formation enthalpies (units given in kcal/mol) and entropies (units given in 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 $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. The notation $xAyN$ corresponds to a cluster with x sulfuric acid and y ammonia molecules.
4 Conclusions

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

First, we sought to determine the cluster evaporation rates from both steady-state and time-dependent cluster concentration data at one temperature. We were only able to identify a subset of the free parameters (evaporation rates) from the available data using either of these approaches.

Next, we used steady-state concentration data corresponding to two different temperatures. We introduced a reparametrization which expressed the evaporation rates in terms of temperature and cluster formation enthalpies and entropies. Using steady-state concentrations at 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 formation enthalpies and entropies), which reduced the number of unknown parameters. This reduced the number of parameters we sought to identify.

Second, it also lessened the stiffness of the system, as the cluster formation enthalpies and entropies for our system span a much smaller range compared to the evaporation rates. 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.

In general, the accuracy of the MCMC results naturally 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, as well as different sulfuric acid source rates, will be considered in future work.
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 instrumental and data-analysis-related 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. Moreover, reliable steady-state concentrations of clusters at various temperatures are generally easier to obtain experimentally (e.g. in chamber experiments) compared to time-dependent concentrations. This finding demonstrates the more general feature of modelling of the type performed here: it can be used to optimize planning of experiments, and thus save both time and resources. Determining very low (below $10^{-5} \text{s}^{-1}$) evaporation rates may also require additional measurements at low vapor concentrations, which naturally require longer timescales to reach a steady state. Treating the uncertainties inherent in experimental data will be the topic of our future studies.

*Code availability.* The code is available via GitHub repository: http://doi.org/10.5281/zenodo.3766925
A Supplementary mathematical material

A1 Cluster kinetic

The kinetics of cluster formation is described by Becker-Döring equations (??), 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 \( i \)th cluster concentration \( Y_i \) is governed by

\[
\frac{dY_i}{dt} = \frac{1}{2} \sum_{j<i} \beta_{i,j} Y_i Y_{i-j} + \sum_j \gamma_{i+j\rightarrow i,j} Y_{i+j} - \sum_j \beta_{i,j} Y_i Y_j - \frac{1}{2} \sum_{j<i} \gamma_{j\rightarrow i,j-i} Y_i + Q_i - S_i, \quad (A1)
\]

where \( \beta_{i,j} \) is the collision coefficient of clusters \( i \) with \( j \), and \( \gamma_{i+j\rightarrow i,j} \) is the evaporation coefficient of cluster \( i+j \) into clusters \( i \) and \( j \), \( Q_i \) is an external source term of \( i \), and \( S_i \) 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 sulfuric acid monomers are supplied to the system at a constant rate comprising \( Q = 6.3 \times 10^4 \text{ cm}^{-3}\text{s}^{-1} \). This settings are selected to imitate the conditions inside of the CLOUD chamber, (??). Further, we include wall losses arising from clusters sticking on the walls of the experimental chamber, (??). These wall losses are parametrized by the size of the cluster

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

where \( r_i \) is the mass radius of the cluster (in cm). From Eq. A2, 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 \( S_{\text{dil},i} = 9.6 \times 10^{-5} \text{ s}^{-1} \), which had previously been determined in the CLOUD chamber, (??).

Let \( T \) denote the temperature of the system of molecular clusters. Using classical kinetic gas theory, the collision rates \( \beta_{i,j} \) in Eq. A1 obey

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

where \( m_i \) and \( V_i \) are respectively the mass and volume of cluster \( i \), and \( k_B \) is Boltzmann’s constant. In this paper, we assume that the masses and volumes are temperature-independent.

The cluster evaporation rates \( \gamma_{i+j\rightarrow i,j} \) in Eq. A1 are given by the expression

\[
\gamma_{i+j\rightarrow i,j} = \beta_{i,j} \frac{P_{\text{ref}}}{k_B T} \exp \left( \frac{\Delta G_{i+j} - \Delta G_i - \Delta G_j}{k_B T} \right), \quad (A4)
\]

---

346removed: kinematics
347removed: sulphuric
where $P_{\text{ref}}$ is the reference pressure and $\Delta G_i$ is the Gibbs free energy of formation for cluster $i$. We may further describe the $i$th Gibbs free energy in terms of the cluster formation enthalpy $\Delta H_i$ and entropy $\Delta S_i$:

$$\Delta G_i = \Delta H_i - T\Delta S_i.$$  \hfill (A5)

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

### A1.1 The Metropolis algorithm

We first select the flat prior distribution from which we will initially sample unknown parameters, as we wish to generate physically reasonable parameter estimates. Therefore, we generate unknown parameters within the chosen minimum and maximum bounds where all the points are equally likely to be sampled. Please see Section 2.2.3 and Tabs. 3-4 for more details. From the prior distribution, a starting guess for the parameters $\theta_{\text{old}} \in \mathbb{R}^{n_{\text{coef}}}$ is chosen (here $n_{\text{coef}}$ is the total number of parameters).

A2 [..348 ]

The Metropolis algorithm then requires us to specify how to sample new parameter values $\theta_{\text{new}}$. This is done by choosing a proposal distribution. We chose a multivariate Gaussian proposal density $q$, defined by:

$$q(\theta_{\text{old}}, \theta_{\text{new}}) \approx \exp \left( -\frac{1}{2} \Sigma^{-1} (\theta_{\text{new}} - \theta_{\text{old}}) \right),$$  \hfill (A6)

where $\Sigma$ is a covariance matrix (of dimensions $n_{\text{coef}} \times n_{\text{coef}}$) which specifies the scaling and spatial orientation of the Gaussian proposal distribution. As the normalization constants are cancelled out in Eq. A9, we do not take them into consideration.

Next, we run the ACDC and Fortran simulations with the parameter values $\theta_{\text{new}}$. We collect the cluster concentration outputs in the column-vector $y_{\text{mod}}(\theta_{\text{new}}) \in \mathbb{R}^{n_{\text{out}}}$, where $n_{\text{out}}$ is the number of elements. The candidate vector of parameters $\theta_{\text{new}}$ is either accepted or rejected according to the least-squares fit of $y_{\text{mod}}(\theta_{\text{new}})$ to the synthetic cluster concentrations $y_{\text{exp}}$:

$$SS(\theta_{\text{new}}) = \sum_{i=1}^{n_{\text{out}}} \frac{(y_{\text{exp},i} - y_{\text{mod},i}(\theta_{\text{new}}))^2}{\sigma_i^2},$$  \hfill (A7)

where $n_{\text{out}}$ is the number of elements.

---

348 removed: Likelihood, data and cost function
349 removed: The likelihood of observing the data $Y_{\text{exp}}$ given the parameter values $\theta$ is
350 removed: $n_{\text{out}}$ is the number of measurements and $F(\theta)$ 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 $n_{\text{out}} = 4 \times (N_c \times N_t + 1)$, where $N_c = 16$
352 removed: cluster types whose concentrations are measured and $N_t = 41$
358 removed: 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 concentration to ensure that the system has attained the steady-state.
measurements in the synthetic concentrations. 

By construction our synthetic data contains uncorrelated Gaussian measurement error, hence the likelihood of observing the data \( y_{\text{exp}} \) given some parameter values \( \theta \) is

\[
p(y_{\text{exp}}|\theta) \simeq \exp\left(-\frac{1}{2}SS(\theta)\right).
\]

(A8)

The value \( SS(\theta_{\text{new}}) \) is then compared to the least-square sum from the previous step \( SS(\theta_{\text{old}}) \) and accepted with the probability

\[
p_{\text{acc}}(\theta_{\text{old}},\theta_{\text{new}}) = \min\left\{ 1, \frac{p(y_{\text{exp}}|\theta_{\text{new}})}{p(y_{\text{exp}}|\theta_{\text{old}})} \right\} = \min\left\{ 1, \exp\left[-\frac{1}{2}(SS(\theta_{\text{new}}) - SS(\theta_{\text{old}}))\right] \right\}.
\]

(A9)

If \( \theta_{\text{new}} \) is accepted, this parameter combination is added as the next element in the chain; else the old value is replicated in the chain. Finally, the value \( SS(\theta_{\text{old}}) \) is replaced with \( SS(\theta_{\text{new}}) \) and saved. This completes an iteration of the Metropolis algorithm.

We remark here that the likelihoods \( p(y_{\text{exp}}|\theta_{\text{old}}) \) and \( p(y_{\text{exp}}|\theta_{\text{new}}) \) in Eq. A9 characterize how closely the outputs of the ACDC simulations with the parameters \( \theta_{\text{old}} \) and \( \theta_{\text{new}} \) respectively fit the synthetic data. By definition of the acceptance probability \( p_{\text{acc}}(\theta_{\text{old}},\theta_{\text{new}}) \) in Eq. A9, the candidate step is always accepted if the new parameters fit the data at least as good as the old values \( (SS(\theta_{\text{new}}) \leq SS(\theta_{\text{old}})) \).

A1.1 The DRAM algorithm for sampling from large parameter space

Our implementation of the Delayed Rejection Adaptive Metropolis (DRAM) approach to MCMC parameter estimation modifies the above Metropolis algorithm in the following way.

First, we use the Adaptive Metropolis (AM) method for updating the covariance matrix \( \Sigma \) of the proposal distribution \( q(\theta_{\text{old}},\theta_{\text{new}}) \) in Eq. A6. That is, if we have generated samples \( (\theta_0, \theta_1, \ldots, \theta_{n-1}) \), the next candidate set \( \theta_{\text{new}} \) is proposed from \( q(\theta_{n-1},\theta_{\text{new}}) \) using the empirical covariance \( \Sigma = \text{Cov}(\theta_0, \theta_1, \ldots, \theta_{n-1}) \). Therefore the next candidate set is generated by taking a step with direction and size determined from the values of parameters previously sampled in the MCMC.
chain. This procedure is carried out after every 100 successive accept/reject iterations. To ensure computational stability, we also apply additional scaling and regularization for the proposal covariance (see ??); please see ? for a detailed explanation.

Second, we carry out local adaptation of the proposal distribution using the Delayed Rejection (DR) algorithm (?). It is implemented as follows: given \( n \) parameter sets \((\theta_0, \theta_1, \ldots, \theta_n)\) generated by the AM method above, a candidate \( \theta_{\text{new}} \) is proposed from the distribution \( q(\theta_n, \theta_{\text{new}}) \) in Eq. A6 and accepted with probability as in Eq. A9, as discussed before. However, if the proposed \( \theta_{\text{new}} \) is rejected, instead of replicating the previous values in the MCMC chain (i.e., \( \theta_{n+1} = \theta_n \)), the algorithm tests a new candidate move \( \theta_{\text{new,2}} \) which is close to the current estimate \( \theta_n \). Then the second-stage proposal \( \theta_{\text{new,2}} \) is accepted with appropriately adjusted acceptance probability (see ?).

In summary, our application of the DRAM algorithm combines the AM procedure with a two-stage DR modification. In the first stage, our algorithm carries out the Metropolis regime with both AM adaptation. The proposal covariance at the initialization of DR (denoted as \( \Sigma \)) is computed as by AM method above, no matter at which stage of DR these points have been accepted in the sampling process. The covariance of the proposal for the second stage (denoted as \( \Sigma^2 \)) is always computed as the scaled version of the first-stage proposal covariance:

\[
\Sigma^2 = \gamma \Sigma, \tag{A10}
\]

with the scaling coefficient \( \gamma = 1/5 \) that was chosen to increase the number of accepted candidate steps at the second stage (?).

This DRAM parameter estimation was conducted using the 'mcmcstat' toolbox implemented for FORTRAN (??). See the description and the examples of usage on the web page helios.fmi.fi/~lainema/.

\(^{370}\)removed: Now \( N_T = 2 \) denotes the number of steady state configurations at different temperatures.
\(^{371}\)removed: \( T \) stands for the measured temperature.
\(^{372}\)removed: In this study, the number of measurements for likelihood given by Eq. ? is \( n_{\text{out}} = 4 \times (N_c \times N_T + 1) \) (again \( N_c = 16 \) cluster types).
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 \cite{376} used to generate the synthetic data. \cite{377} The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..379] used to generate the synthetic data. [..380] The notation \(x:AyN\) corresponds to a cluster with \(x\) sulfuric acid and \(y\) ammonia molecules.
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 [..$^{382}$] used to generate the synthetic data [..$^{383}$]. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [...]. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [.388] used to generate the synthetic data. [.389] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 $^{[391]}$ used to generate the synthetic data $^{[392]}$. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [394] used to generate the synthetic data. [395] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
Figure C1. Time-dependent cluster concentrations. 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 sulfuric acid monomer is [H$_2$SO$_4$] = 6.3 × 10$^4$ s$^{-1}$ in all simulations. The notation $x_Ay_N$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.

\footnote{removed: transient}
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 time-dependent 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. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..403] time-dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red lines denote the baseline values from [..404] used to generate the synthetic data [..405]. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and [..406] $y$ ammonia molecules.
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 \cite{407} time-dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from \cite{408} used to generate the synthetic data. \cite{409} The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 \cite{411} time-dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from \cite{412} used to generate the synthetic data \cite{413}. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 \[..^{415}\] time-dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from \[..^{416}\] used to generate the synthetic data \[..^{417}\] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and [$..^{418}$] $y$ ammonia molecules.
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 time dependent 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. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..423] Time dependent measurements of the cluster concentrations with time resolution comprising 1.5 minutes at the temperature 278 K. Red rectangles denote the baseline values from [..424] used to generate the synthetic data. [..425] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
Table C1. Part 1. Evaporation rates (units given in $s^{-1}$) determined from the steady-state and the time dependent data presented in Figure 5-6 and 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 $^{429}$. The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
Figure D1. Steady-state cluster concentrations for the clusters containing sulfuric acid and a varying number of ammonia molecules as a function of the number of acid molecules for \([\text{NH}_3]\) 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 sulfuric acid monomer comprises \([\text{H}_2\text{SO}_4]=6.3 \times 10^4\) s\(^{-1}\) in all the simulations. Here the symbols \(\Delta H\) and \(\Delta S\) stand for cluster formation enthalpies and entropies, respectively. The notation \(x:yzw\) corresponds to a cluster with \(x\) sulfuric acid and \(y\) ammonia molecules.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Steady-state data (s⁻¹)</th>
<th>Time dependent data (s⁻¹)</th>
<th>QC (s⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21: 4A3N → 2A1N</td>
<td>$&lt; 2.14 \times 10^{-4}$</td>
<td>$4.83 \times 10^{-9}$</td>
<td>$1.66 \times 10^{-8}$</td>
</tr>
<tr>
<td></td>
<td>(3.36 × 10⁻¹², 6.93 × 10⁻⁶)</td>
<td>(7.77 × 10⁻¹¹, 7.86 × 10⁻¹)</td>
<td></td>
</tr>
<tr>
<td>22: 5A3N → 1A</td>
<td>$7.88 \times 10^{-1}$</td>
<td>$7.81 \times 10^{-1}$</td>
<td>$7.83 \times 10^{-1}$</td>
</tr>
<tr>
<td></td>
<td>$(5.56 \times 10^{-1}, 8.20 \times 10^{-1})$</td>
<td>$(4.57 \times 10^{-1}, 5.33 \times 10^{-1})$</td>
<td></td>
</tr>
<tr>
<td>23: 5A3N → 2A</td>
<td>$2.35 \times 10^{-8}$</td>
<td>$6.34 \times 10^{-7}$</td>
<td>$6.37 \times 10^{-7}$</td>
</tr>
<tr>
<td></td>
<td>$(&lt; 1.21 \times 10^{-2})$</td>
<td>$(1.26 \times 10^{-11}, 3.35 \times 10^{-4})$</td>
<td></td>
</tr>
<tr>
<td>24: 5A3N → 1A1N</td>
<td>$9.12 \times 10^{-12}$</td>
<td>$1.50 \times 10^{-9}$</td>
<td>$1.70 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 3.39 \times 10^{-3}$</td>
<td>$&lt; 1.02 \times 10^{-12}, 2.22 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>25: 5A3N → 2A1N</td>
<td>$7.22 \times 10^{-4}$</td>
<td>$1.24 \times 10^{-5}$</td>
<td>$1.85 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 6.95 \times 10^{-3}$</td>
<td>$(1.86 \times 10^{-8}, 5.33 \times 10^{-4})$</td>
<td></td>
</tr>
<tr>
<td>26: 5A3N → 2A2N</td>
<td>$1.52 \times 10^{-8}$</td>
<td>$&lt; 1.25 \times 10^{-4}$</td>
<td>$3.52 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 4.49 \times 10^{-3}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27: 4A4N → 1N</td>
<td>$3.79 \times 10^{-1}$</td>
<td>$3.76 \times 10^{-1}$</td>
<td>$3.75 \times 10^{-1}$</td>
</tr>
<tr>
<td></td>
<td>$(3.70 \times 10^{-1}, 3.88 \times 10^{-1})$</td>
<td>$(3.75 \times 10^{-1}, 3.77 \times 10^{-1})$</td>
<td></td>
</tr>
<tr>
<td>28: 4A4N → 1A1N</td>
<td>$9.05 \times 10^{-6}$</td>
<td></td>
<td>$9.06 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 5.38 \times 10^{-3}$</td>
<td>$(1.52 \times 10^{-10}, 2.57 \times 10^{-4})$</td>
<td></td>
</tr>
<tr>
<td>29: 4A4N → 2A2N</td>
<td>$2.07 \times 10^{-12}$</td>
<td>$8.55 \times 10^{-11}$</td>
<td>$1.33 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 2.43 \times 10^{-3}$</td>
<td>$&lt; 1.90 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>30: 5A4N → 1A</td>
<td>$3.87 \times 10^{-6}$</td>
<td>$2.51 \times 10^{-3}$</td>
<td>$1.77 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 2.52 \times 10^{-2}$</td>
<td>$(1.20 \times 10^{-6}, 5.86 \times 10^{-3})$</td>
<td></td>
</tr>
<tr>
<td>31: 5A4N → 1N</td>
<td>$8.92 \times 10^{-2}$</td>
<td>$9.03 \times 10^{-2}$</td>
<td>$8.87 \times 10^{-2}$</td>
</tr>
<tr>
<td></td>
<td>$(6.68 \times 10^{-2}, 9.74 \times 10^{-2})$</td>
<td>$(8.52 \times 10^{-2}, 9.19 \times 10^{-2})$</td>
<td></td>
</tr>
<tr>
<td>32: 5A4N → 1A1N</td>
<td>$3.60 \times 10^{-6}$</td>
<td></td>
<td>$7.33 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 1.55 \times 10^{-2}$</td>
<td>$(6.48 \times 10^{-12}, 1.04 \times 10^{-3})$</td>
<td></td>
</tr>
<tr>
<td>33: 5A4N → 2A1N</td>
<td>$2.28 \times 10^{-4}$</td>
<td>$1.32 \times 10^{-4}$</td>
<td>$2.97 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 1.06 \times 10^{-2}$</td>
<td>$(6.46 \times 10^{-10}, 1.53 \times 10^{-3})$</td>
<td></td>
</tr>
<tr>
<td>34: 5A4N → 2A2N</td>
<td>$&lt; 1.08 \times 10^{-2}$</td>
<td>$7.30 \times 10^{-9}$</td>
<td>$6.42 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$(1.51 \times 10^{-11}, 3.17 \times 10^{-4})$</td>
<td></td>
</tr>
<tr>
<td>35: 4A5N → 1N</td>
<td>$8.75 \times 10^{2}$</td>
<td>$8.88 \times 10^{2}$</td>
<td>$8.89 \times 10^{2}$</td>
</tr>
<tr>
<td></td>
<td>$(8.59 \times 10^{2}, 9.03 \times 10^{2})$</td>
<td>$(8.85 \times 10^{2}, 8.92 \times 10^{2})$</td>
<td></td>
</tr>
<tr>
<td>36: 5A5N → 1A</td>
<td>$&lt; 2.32 \times 10^{-4}$</td>
<td></td>
<td>$2.23 \times 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$&lt; 1.14 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>37: 5A5N → 1N</td>
<td>$4.96 \times 10^{-4}$</td>
<td>$1.00 \times 10^{-4}$</td>
<td>$1.17 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 9.89 \times 10^{-4}$</td>
<td>$(3.48 \times 10^{-5}, 1.85 \times 10^{-4})$</td>
<td></td>
</tr>
<tr>
<td>38: 5A5N → 1A1N</td>
<td>$5.93 \times 10^{-9}$</td>
<td>$1.48 \times 10^{-11}$</td>
<td>$2.11 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 5.06 \times 10^{-4}$</td>
<td>$&lt; 1.06 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>39: 5A5N → 2A2N</td>
<td>$&lt; 3.09 \times 10^{-4}$</td>
<td>$2.06 \times 10^{-11}$</td>
<td>$1.31 \times 10^{-11}$</td>
</tr>
<tr>
<td></td>
<td>$&lt; 4.11 \times 10^{-7}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table C2. Part 2. Evaporation rates (units given in s⁻¹) determined from the steady-state and the time dependent data presented in Figure 5-6 and 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 \cite{432} used to generate the synthetic data. \cite{433} The notation xAyN corresponds to a cluster with x sulfuric acid and y ammonia molecules.
Figure D2. Parameter chains of the cluster formation enthalpies (units given in kcal/mol) and entropies (units given in 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 $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. [..] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [..442] used to generate the synthetic data. Here the symbols $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. [..443] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 $^{[445]}$ used to generate the synthetic data. Here the symbols $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. $^{[446]}$ The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
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 [448] used to generate the synthetic data. Here the symbols $\Delta H$ and $\Delta S$ stand for cluster formation enthalpies and entropies, respectively. [449] The notation $xAyN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Mode value</th>
<th>95% confidence interval</th>
<th>QC</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: (\Delta H_{2A})</td>
<td>-17.8891</td>
<td>(-18.1913,-17.4941)</td>
<td>-17.85</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>2: (\Delta S_{2A})</td>
<td>-33.5475</td>
<td>(-34.6104,-32.1575)</td>
<td>-33.42</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>3: (\Delta H_{1A1N})</td>
<td>-15.8751</td>
<td>(-16.2344,-15.5158)</td>
<td>-16</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>4: (\Delta S_{1A1N})</td>
<td>-27.6984</td>
<td>(-28.9594,-26.3474)</td>
<td>-28.14</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>5: (\Delta H_{2A1N})</td>
<td>-44.8076</td>
<td>(-45.2922,-44.174)</td>
<td>-45</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>6: (\Delta S_{2A1N})</td>
<td>-70.3501</td>
<td>(-72.029,-68.1545)</td>
<td>-71.02</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>7: (\Delta H_{3A1N})</td>
<td>-66.0006</td>
<td>(-66.428,-65.5732)</td>
<td>-66.06</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>8: (\Delta S_{3A1N})</td>
<td>-107.5233</td>
<td>(-109.0059,-106.0407)</td>
<td>-107.72</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>9: (\Delta H_{2A2N})</td>
<td>-15.8751</td>
<td>(-16.2344,-15.5158)</td>
<td>-16</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>10: (\Delta S_{2A2N})</td>
<td>-104.6181</td>
<td>(-106.2857,-102.9505)</td>
<td>-104.45</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>11: (\Delta H_{3A2N})</td>
<td>-121.6575</td>
<td>(-123.9554,-121.3595)</td>
<td>-122.33</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>12: (\Delta S_{3A2N})</td>
<td>-115.0105</td>
<td>(-116.7515,-113.2696)</td>
<td>-115.13</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>13: (\Delta H_{4A2N})</td>
<td>-44.8076</td>
<td>(-45.2922,-44.174)</td>
<td>-45</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>14: (\Delta S_{4A2N})</td>
<td>-70.3501</td>
<td>(-72.029,-68.1545)</td>
<td>-71.02</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>15: (\Delta H_{3A3N})</td>
<td>-91.8512</td>
<td>(-93.9174,-90.2712)</td>
<td>-92.09</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>16: (\Delta S_{3A3N})</td>
<td>-121.6575</td>
<td>(-123.9554,-121.3595)</td>
<td>-122.33</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>17: (\Delta H_{4A3N})</td>
<td>-144.9757</td>
<td>(-147.3975,-142.554)</td>
<td>-145.17</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>18: (\Delta S_{4A3N})</td>
<td>-115.0105</td>
<td>(-116.7515,-113.2696)</td>
<td>-115.13</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>19: (\Delta H_{5A3N})</td>
<td>-168.7305</td>
<td>(-171.0579,-166.4031)</td>
<td>-168.79</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>20: (\Delta S_{5A3N})</td>
<td>-260.3509</td>
<td>(-268.3225,-252.3794)</td>
<td>-260.55</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>21: (\Delta H_{4A4N})</td>
<td>-164.1272</td>
<td>(-166.4394,-161.815)</td>
<td>-164.35</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>22: (\Delta S_{4A4N})</td>
<td>-250.2634</td>
<td>(-258.1819,-242.3449)</td>
<td>-251.03</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>23: (\Delta H_{5A4N})</td>
<td>-250.2634</td>
<td>(-258.1819,-242.3449)</td>
<td>-251.03</td>
<td>cal K(^{-1}) mol(^{-1})</td>
</tr>
<tr>
<td>24: (\Delta S_{5A4N})</td>
<td>-191.7779</td>
<td>(-194.9426,-188.6133)</td>
<td>-191.86</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>25: (\Delta H_{4A5N})</td>
<td>-290.7782</td>
<td>(-301.6196,-279.9369)</td>
<td>-291.05</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>26: (\Delta S_{4A5N})</td>
<td>-290.7782</td>
<td>(-301.6196,-279.9369)</td>
<td>-291.05</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>27: (\Delta H_{5A5N})</td>
<td>-290.7782</td>
<td>(-301.6196,-279.9369)</td>
<td>-291.05</td>
<td>kcal mol(^{-1})</td>
</tr>
<tr>
<td>28: (\Delta S_{5A5N})</td>
<td>-290.7782</td>
<td>(-301.6196,-279.9369)</td>
<td>-291.05</td>
<td>kcal mol(^{-1})</td>
</tr>
</tbody>
</table>

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 [..451] used to generate the synthetic data. Here the symbols \(\Delta H\) and \(\Delta S\) stand for cluster formation enthalpies and entropies, respectively. [..452] The notation \(xAyN\) corresponds to a cluster with \(x\) sulfuric acid and \(y\) ammonia molecules.

[451] [452] [453]
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 [454] used to generate the synthetic data. [455] The notation \(xAyN\) corresponds to a cluster with \(x\) sulfuric acid and \(y\) ammonia molecules.
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 [457] used to generate the synthetic data [458]. The notation $x_{A}y_{N}$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.
### 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 (borrowed from ?). The notation \( xA_y N \) corresponds to a cluster with \( x \) sulfuric acid and \( y \) ammonia molecules. [..\textsuperscript{460} ]
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 (borrowed from ?). The notation $xA_yN$ corresponds to a cluster with $x$ sulfuric acid and $y$ ammonia molecules.$^{[..461]}$

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., Kurtén T. and Vehkamäki H. and Haario H. are responsible for writing the manuscript. 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. 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.