Supplementary material for "A Monte Carlo approach for determining cluster evaporation rates from concentration measurements"

Oona Kupiainen-Määttä

University of Helsinki, Department of Physics, P.O. Box 64, FI-00014 University of Helsinki, Finland

---

This document contains a detailed description of the MCMC simulations (Sect. S1), a detailed discussion of the MCMC results for synthetic cluster distributions (Sect. S2), as well as supplementary figures related to the simulations discussed in the main text (Sect. S3).

S1. Details of the MCMC simulations

S1.1. Initialization

The DE-MC$_\mathrm{Z}$ algorithm (ter Braak and Vrugt, 2008) described in the main text requires an initial history chain $\boldsymbol{Z}_0$. As the step sizes depend mostly on the history, $\boldsymbol{Z}_0$ must correspond roughly to the posterior distribution, which however is not known beforehand. The distribution of each parameter in the chain $\boldsymbol{Z}_0$ must be wide enough to encompass all local maxima in order for the DE-MC$_\mathrm{Z}$ algorithm to be able to jump from one maximum to another. In this study, this was accomplished by running a large number of short MCMC chains starting from random points in the allowed parameter space. On the other hand, the distribution must not be too wide, since then the proposed steps will be too long leading to a low acceptance probability. This was effected by discarding the chains that fit very poorly to the data.

Each simulation was started by running 40 separate MCMC chains of 5000 steps. The collision coefficients were sampled using the Metropolis algorithm as described in the main text, with the steps drawn from a normal distribution with a width σ_step of 0.05 times the width of the allowed range of each parameter.

At any step, only a fraction of the coefficients were varied. Each coefficient separately was chosen with a probability of 10 % to be varied at a given step, but if no coefficients were picked, the selection process was repeated. For a total of 43 coefficients, the average number of coefficients that were varied per step was ∼4.3, but at 6 % of the steps 8 or more coefficients were varied.

The acceptance probability was computed with $\sigma_0 = 0.2$. The first half of each chain was discarded, and during the second half the lowest value of the square sum, $SS_{\mathrm{min},i}$, was recorded for each chain i. The second halves of the chains for which $SS_{\mathrm{min},i}$ was lower than the geometric mean $(\prod_{j=1}^{40} SS_{\mathrm{min},j})^{1/40}$ of all chains were combined and thinned to a total of 10000 points to form the initial history $\boldsymbol{Z}_0$.

The overall lowest square sum SS_min was also used for estimating the scatter in the measured data as $\sigma = (SS_\mathrm{min}/(n_\mathrm{out} - n_\mathrm{coefs}))^{1/2}$. This value was used for calculating the acceptance probability in the later stages of the simulation.

After combining the chains with lowest $SS_{\mathrm{min},i}$ but before thinning the resulting chain, the variance $\sigma^2_{k,\mathrm{ini}}$ of each parameter k was computed to give an estimate of the widths of the distributions. In the DE-MC$_\mathrm{Z}$ simulation, the term δ was drawn from a normal distribution with width $\sigma_{\delta,k} = 0.05 \times \sigma_{k,\mathrm{ini}}$ for each parameter.

S1.2. Main simulation and convergence

The history $\boldsymbol{Z}$ was initialized as the $10000 \times n_\mathrm{coefs}$ matrix $\boldsymbol{Z}_0$, and each of the five chains was started from a random point of the history. A varying number of coefficients were sampled simultaneously at each step as described above, now using the DE-MC$_\mathrm{Z}$ algorithm presented in the main text. At every 50th step, the parameter values of all chains were appended to the matrix $\boldsymbol{Z}$ as five new rows.

The first 20000 steps of each chain were discarded as burn-in. After that, every 50th step was printed out. Monitoring of the convergence was started after the burn-in period. The mean and variance of all parameters were calculated separately for each chain and updated at every step. These were used to compute for each parameter i the $\hat{R}$ statistic (Brooks and Gelman, 1998; Gelman and Rubin, 1992) $\hat{R}_i = \frac{n-1}{n} + \frac{m+1}{m}\frac{b_i}{W_i}$, where n is the number of steps in each chain, $m = 5$ is the number of chains, b_i is the variance of the means of parameter i in the different chains and W_i is the mean of the variances of parameter i in the different chains. A high value of $\hat{R}$ means that the chains are far away from each other compared to the width of the distributions within each chain. Typically $\hat{R}$ decreases with time as the chains cover the target distribution, but it can also increase again if different chains get stuck in different local maxima of the distribution. All simulations were run for a minimum of 10 000 000 steps per chain to ensure proper mixing and convergence. After that, the simulation was defined to have converged when $\hat{R} < 1.1$.

S2. Testing the method: analyzing synthetic cluster distributions

In order to test how reliably rate constants could be inferred from the MCMC simulation, artificial cluster distributions corresponding to known cluster evaporation rates were first produced using ACDC. The ion-molecule collision rates were computed according to the parameterization of Su and Chesnavich (1982), and evaporation rates were calculated based on quantum chemical cluster formation energies of dry clusters (Ortega et al., 2014). The synthetic data set consisted of a total of 22 cluster distributions corresponding to the same sulfuric acid and ammonia vapor concentrations as in the experimental cluster distributions.

However, for the 13 experiments with no added ammonia, the ammonia concentration was not known. No quantitative estimates were available for the fragmentation probabilities, and also the estimates of the ion production rate and wall loss may not have been very accurate. For all of these parameters, the values corresponding to the peaks in the posterior distribution of the MCMC simulation analyzing the experimental data were used as input for producing the synthetic cluster

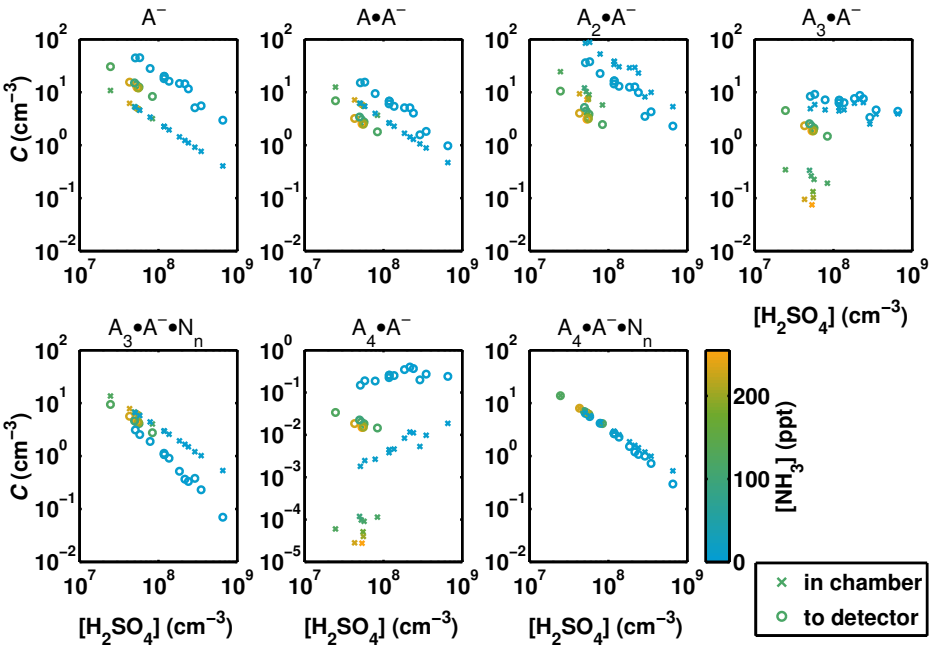

Figure S1: Synthetic cluster distributions before (x) and after (o) fragmentation assumed to happen in the APi-TOF. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

distribution. For consistency, all parameter values were taken from the same MCMC simulation where all possible parameters (evaporation rates, ion production rate, wall loss constant, fragmentation probabilities, background ammonia concentrations) were varied (Figs. 5, S11 and S12).

Figure S1 shows the synthetic cluster distributions before and after fragmentation. The distributions before fragmentation correspond to the actual simulated cluster concentrations in the CLOUD chamber, while the distributions after fragmentation correspond to how these concentrations would ideally be observed with the APi-TOF if there was no noise in the measurement.

To mimic measurement errors, random noise was added to the cluster distribution as $\log_{10} C_{\text{meas.}} = \log_{10} C_{\text{exact}} + \epsilon$, where C_{exact} is the exact concentration after fragmentation, $C_{\text{meas.}}$ is the 'measured' value and ϵ is drawn from a normal distribution with $\sigma_\epsilon = 0.2$. The 'measured' distribution is shown as crosses in Fig. S2 and is used as input in the MCMC simulations.

Two sets of MCMC simulations were performed for the synthetic 'measured' data: first varying all unknown parameters (evaporation rates, ion production rate, wall loss constant, fragmentation probabilities, background ammonia concentrations), and then setting the background ammonia concentrations to 5 ppt and only varying the other parameters. The output cluster distributions from these two cases are presented in Figs. S2 and S3, respectively.

The medians of each concentration from the output of the MCMC simulations are presented as a horizontal line, and the vertical lines show the 2.5th and 97.5th percentiles. The true model concentrations are well reproduced by the MCMC simulations, and especially the trends with respect to sulfuric acid and ammonia concentrations are correct. In cases where the 'measured' concentration contains a large 'measurement error' and is therefore far from the exact value, the MCMC simulation still mainly captures the true value rather than the inaccurate 'measured value' used as input for the MCMC simulation.

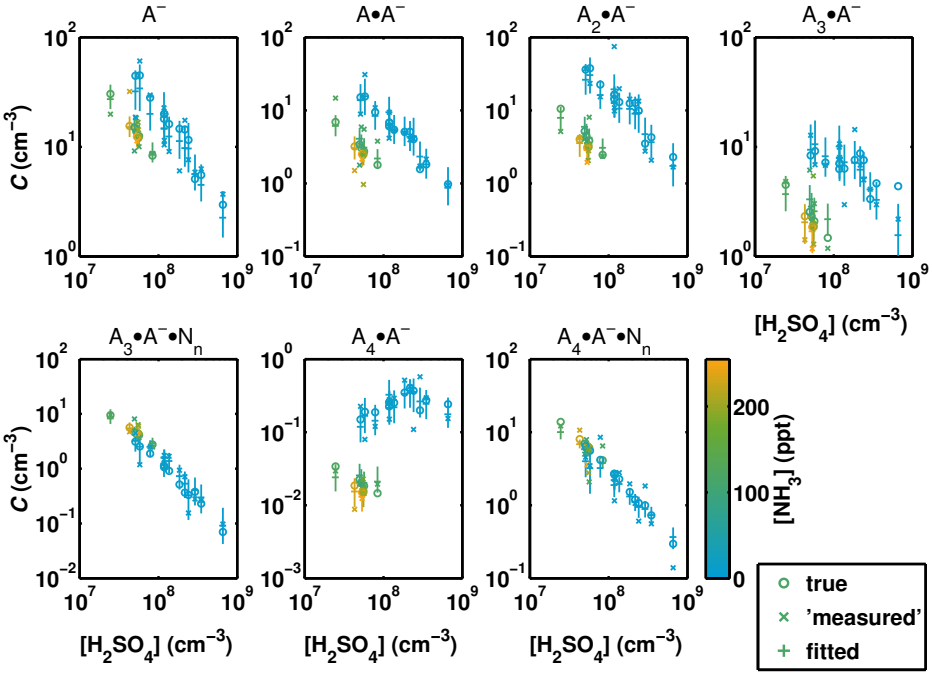

Figure S2: Synthetic cluster distributions after fragmentation (o) and with added noise (x), and the corresponding cluster concentrations from an MCMC simulation where evaporation rates, fragmentation probabilities, the ion production rate, the wall loss rate and the background ammonia concentrations are varied. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

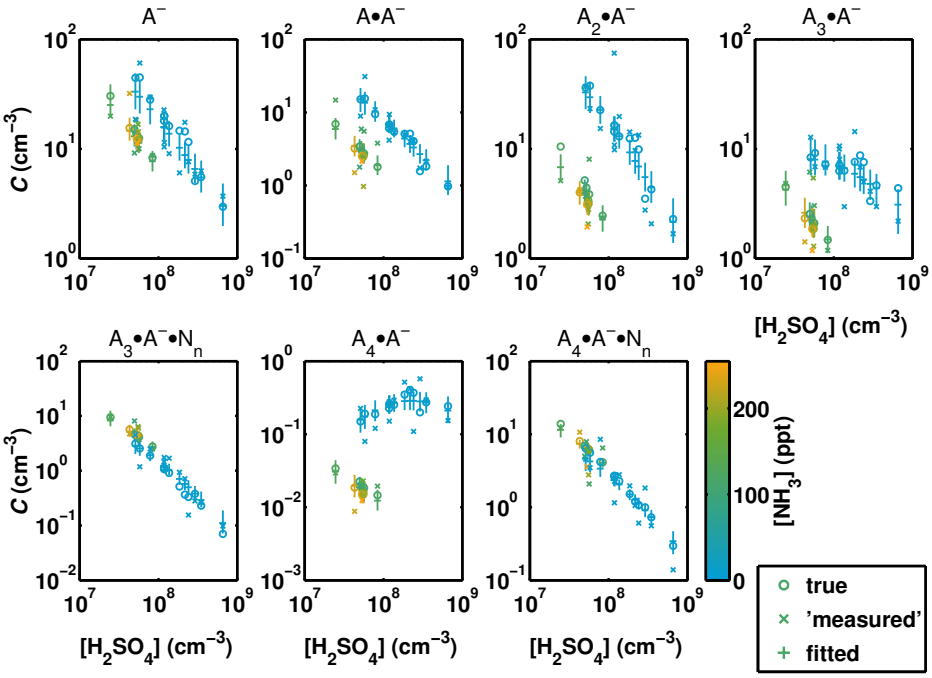

Figure S3: Same as Fig. S2, but all background ammonia concentrations are set to 5 ppt in the MCMC analysis instead of being varied.

S2.1. Extracting evaporation rates from the synthetic data

Figure S4 shows the posterior distributions of the coefficients corresponding to logarithms of the evaporation rates from the two sets of MCMC simulations. In order to interpret the results, it is useful first to consider how each evaporation rate affects the cluster distribution. In case of ammonia evaporation from clusters containing two or more ammonia molecules, both the evaporating cluster and the product contribute to the cluster distribution in the same way, as the clusters containing a given number of acid molecules are divided by ammonia content only into two groups, ammonia-free and ammonia-containing. Thus the value of the evaporation rate has no direct effect on the reported concentrations. On the other hand, ammonia evaporation from clusters containing one ammonia molecule converts the cluster from ammonia-containing to ammonia-free, and consequently alters the measured cluster concentration (unless the ammonia molecule would anyway be lost by fragmentation inside the instrument). Sulfuric acid evaporation rates also have a direct effect on the cluster distribution as clusters containing different numbers of acid molecules appear separately in the distribution. However, in case of acid evaporation from $HSO_4^- \cdot (H_2SO_4)_4 \cdot (NH_3)_{1-3}$ clusters, both the evaporating cluster and the product appear in the measured distribution grouped with other clusters.

In case of the synthetic cluster distributions, the coefficients presented in Fig. S4 can roughly be divided into four categories according to the shape of the posterior distributions. A similar approach is also used in the main text for interpreting the MCMC simulations performed on the experimental data.

Coefficients number 3 and 5 have one or more clear peaks located near the input value of the parameter (shown as a vertical green line), and practically zero probability density elsewhere. Both of these evaporation coefficients have an input value close to 1 s^{-1}.

The second category is formed by the coefficients number 1, 2, 7 and, in the fixed background ammonia case, coefficient number 9. The posterior distributions of these coefficients have a constant non-zero probability density at low values, a constant practically zero probability density at high values, and between the two regions, close to the value 1 s^{-1}, either a sharp drop or a peak followed

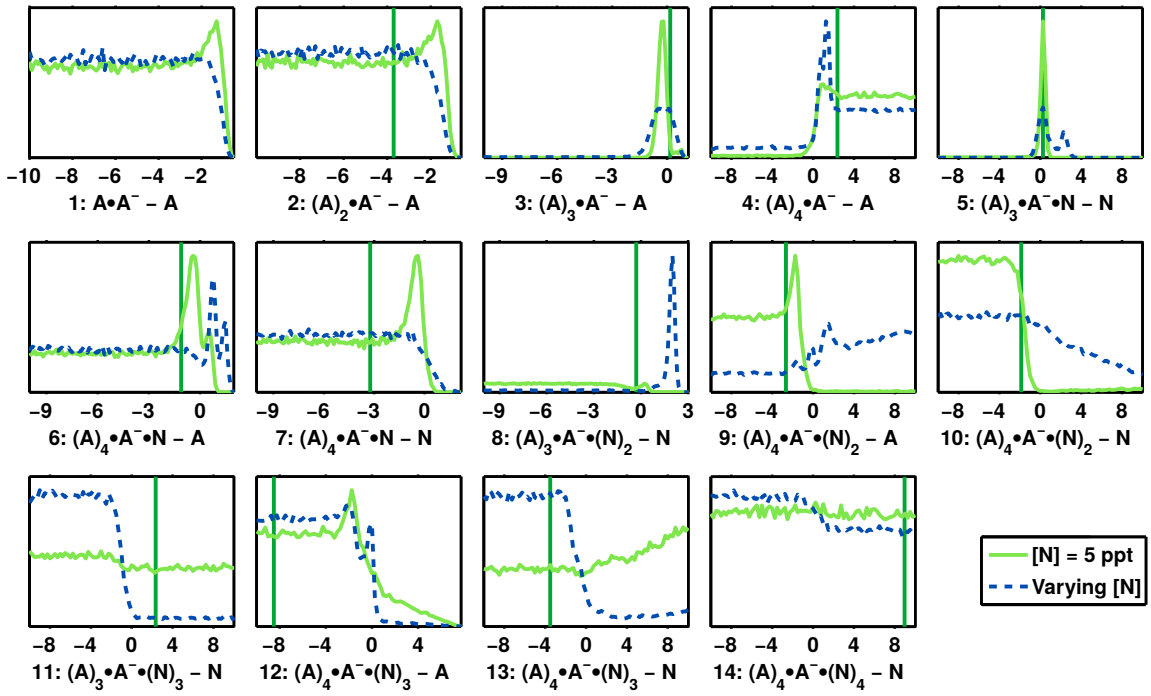

Figure S4: Posterior distributions of the base 10 logarithm of the evaporation rates (in units of s^{-1}) corresponding to the synthetic cluster distributions. The background ammonia concentrations have either been set to 5 ppt (green) or varied with MCMC (blue). The vertical green lines corresponds to the quantum chemistry–based estimates used as input parameter values for producing the synthetic cluster distributions. (For coefficient number 1, the input value -17.8 is outside the range sampled by MCMC.) A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

by a sharp drop. The input values of these evaporation rates are well below 1 s^{-1}, but the precise values cannot be captured from the MCMC analysis. The uniform shape of the distributions at low evaporation rates is explained by the fact that if a cluster has a long enough lifetime with respect to evaporation, it will collide and grow before evaporating, and a further decrease of its evaporation rate does not change the situation. The peaks in the posterior distribution, when present, are mostly far from the corresponding input value of the coefficient, and the existence and prominence of the peaks also varies with the exact values of the random 'measurement errors' added to the cluster distributions. Therefore, the significance and interpretation of the peaks remain unclear.

Coefficients number 6, 8 and 12 also have a uniform non-zero probability density at low values and a zero probability at high values, but the distribution goes to zero at a higher evaporation rate value than in the previous category. In some cases there are also more than one peak. Similarly as in the second category, these evaporation rates have input values lower than 1 s^{-1}, but the exact value cannot be captured from the MCMC analysis.

The remaining coefficients form the fourth category. These have posterior distributions with a distinctly non-zero probability density over the whole range. These rate constants most probably do not have a strong impact on the cluster distribution, and can therefore have whatever value without interfering with the goodness of the fit. Some of the corresponding evaporation processes occur between clusters that are mutually indistinguishable in the cluster distribution, while others are perhaps not on the main formation pathway or at least do not correspond to rate-limiting processes.

S2.2. Extracting the ion production rate and wall loss constant from the synthetic data

The posterior distribution of the ion production rate (Fig. S5) peaks very close to the input value. However, in case of the synthetic data, the same value for the charging rate of sulfuric acid

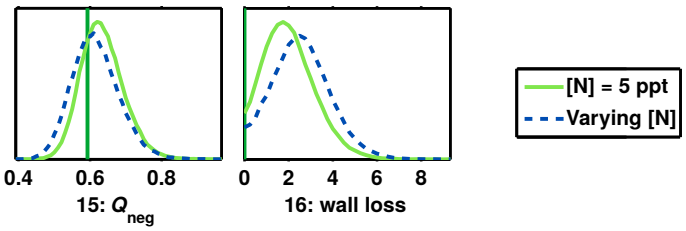

Figure S5: Posterior distributions of the ion production rate rates (in units of $\mathrm{cm}^{-3}\mathrm{s}^{-1}$) and wall loss constant (in units of $10^{-3}\mathrm{s}^{-1}$) corresponding to the synthetic cluster distributions. The background ammonia concentrations have either been set to 5 ppt (green) or varied with MCMC (blue). The vertical green lines corresponds to the MCMC estimates from the experimental data, which are used as input parameter values for producing the synthetic cluster distributions.

molecules by the produced charger ions is used both when producing the cluster distributions and analyzing them. For the experimental data, on the other hand, the true value is not known and the rate coefficient used in the MCMC simulation might not be accurate. Therefore, the promising result of Fig. S5, alone, does not prove how reliably the ion production rate can be determined from the experimental data.

The posterior distribution of the wall loss, on the other hand, does not peak at its input value. In fact, the MCMC result would seem to suggest a non-zero wall loss, as opposed to the input value of exactly zero. This serves as a reminder that the peak of the probability density cannot be directly interpreted as giving the true value of the corresponding parameter. Instead, the whole region of non-zero probability density should be interpreted as possible (and even likely) values for the parameter.

S2.3. Extracting the fragmentation probabilities from the synthetic data

The posterior distribution of the fragmentation probabilities are presented in (Fig. S6). For all parameters and both sets of simulations (either varying or not varying the background ammonia concentrations), the posterior distribution has a non-zero probability at the input parameter value. For the narrower distributions, the peak is mostly close to the input value, while especially for the simulations where the ammonia concentration is varied, some of the wider distributions peak quite far from the input value.

S2.4. Extracting the background ammonia concentrations from the synthetic data

Figure S7 shows the posterior distribution of the ammonia mixing ratios for the thirteen experiments where no ammonia was added intentionally and its concentration was below the detection limit. For each of these mixing ratios, the posterior distribution peaks at a value higher than the input mixing ratio. In most cases, the input value is at the very tail of the probability distribution. It must therefore be concluded that background ammonia concentrations cannot be reliably determined based on this kind of MCMC study.

S2.5. Options for treating the background ammonia concentrations

As the determination of the low ammonia concentrations was seen not to be reliable, the analysis of the experimental data in the main text concentrates mostly on MCMC simulations where the background ammonia concentration is not treated as a free parameter. In the case where the ammonia concentrations are varied with MCMC, the peak locations of the posterior distribution range between 6.4 and 18.3 ppt. Assuming that these values somewhat overestimate the background concentrations, a better estimate for the background concentration might be around 5 ppt or lower. In the MCMC simulations where the ammonia concentration is not allowed to vary, two fixed values are tested: either all background ammonia concentrations are set to 5 ppt, or in an other MCMC simulation they are all set to 1 ppt.

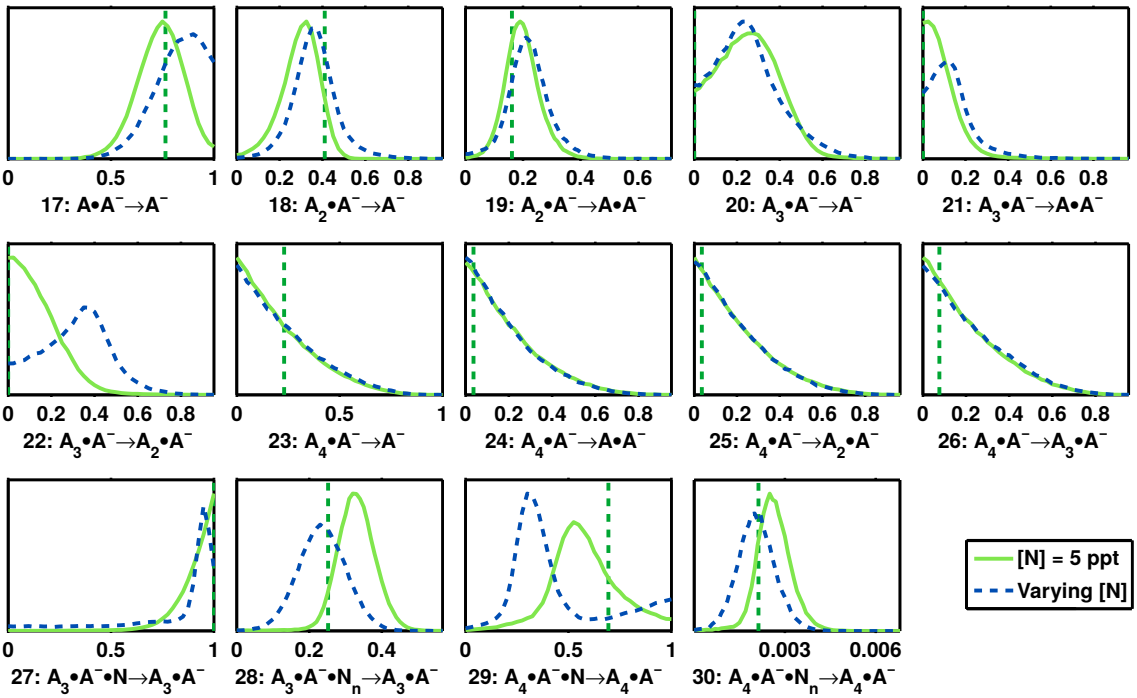

Figure S6: Posterior distributions of the fragmentation probabilities of clusters in the mass spectrometer inlet corresponding to the synthetic cluster distributions. The background ammonia concentrations have either been set to 5 ppt (green) or varied with MCMC (blue). The vertical green lines corresponds to the MCMC estimates from the experimental data, which are used as input parameter values for producing the synthetic cluster distributions. A stands for H_2SO_4, A^- for $HSO_4{}^-$ and N for NH_3.

S2.6. Separating the posterior distributions into alternative scenarios

Ideally, the parameters varied in an MCMC simulation should all be uncorrelated. As a worst-case example of problems caused by correlated parameters, consider a case were the observed quantity depends on the ratio of two of the varied parameters. These might both have a uniform probability density over the whole allowed range, while the posterior distribution of their ratio might have a narrow peak. In this example case, the problem could easily be resolved by using as free parameters the ratio and product of the two parameters, instead of using the original parameters themselves.

With a large number of fitted parameters, also more complicated correlations involving several parameters can occur, and it may not, at least in practice, be possible to choose the parameters so as to eliminate all correlations. For a complicated enough system, there may even be several very different solutions that give an equally good fit to the data. In order to find and study the correlations between different parameters, the output from the MCMC simulations must be considered as a collection of *sets of parameter values* instead of seeing it as separate probability distribution for each parameter. The correlations can be found by grouping the sets of parameter values according to the value of one specific parameter, and comparing the posterior distributions of the other parameters for these groups.

Figure S8 demonstrates the separation of the MCMC results into three scenarios for the synthetic cluster concentration data and a fixed background ammonia concentration in the MCMC simulation. For clarity, the distributions of the scenarios are normalized to have the same overall probability.

First, it can be seen in Fig. S4 that the posterior distribution of coefficient number 3 has two peaks. The points in the smaller peak, that is the sets of parameter values with coefficient number 3 being higher than 0.3, are chosen to form first group of points, denoted (a). The remaining points are further divided into two more groups according to the value of coefficient number 5. The points

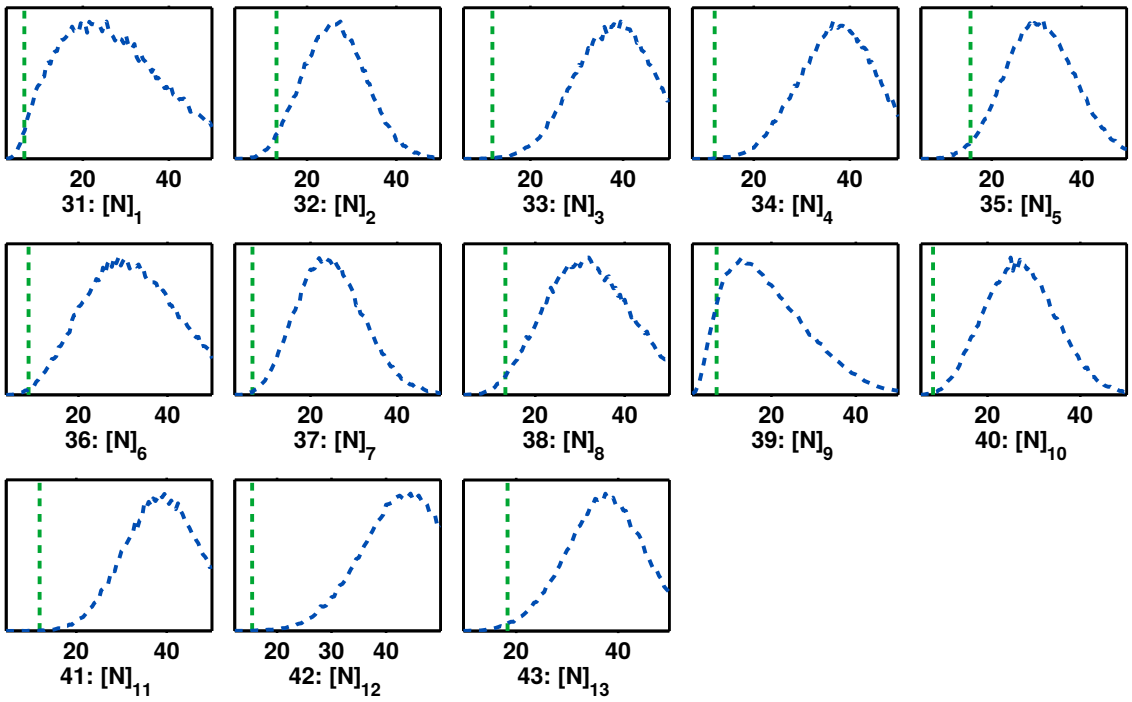

Figure S7: Posterior distributions of the background ammonia concentrations (as mixing ratios in ppt) corresponding to the synthetic cluster distributions and the MCMC simulation where the background ammonia concentrations were varied. The vertical green lines corresponds to the MCMC estimates from the experimental data, which are used as input parameter values for producing the synthetic cluster distributions.

with coefficient number 3 below 0.3 and coefficient number 5 below 2 form group (b), and the points with coefficient number 3 below 0.3 and coefficient number 5 above 2 form group (c).

The grouping of the evaporation rates presented in Sect. S2.1 can now be performed separately for each of the scenarios (a)–(c). A vast majority, about 90%, of the points fall into category (b). For this scenario, the conclusions from Sect. S2.1 do not change much, except for coefficients number 6 and 8 now belonging to group 2 (evaporation rates below 1 s^{-1}).

Scenarios (a) and (c) correspond to two very different solutions that nevertheless give an equally good fit to the synthetic data. In all three scenarios, the pure dimer and trimer are very stable, the $HSO_4^- \cdot (H_2SO_4)_4 \cdot NH_3$ cluster is stable with respect to ammonia evaporation, and coefficients number 10, 11 and 14 have distinctly non-zero probability density over the whole allowed range and their values cannot be determined. For the remaining coefficients, the parameters have different values (well below 1 s^{-1}, close to 1 s^{-1}, undefined) in the three different scenarios.

As the 'correct values' of the parameters are in this case known, it can be seen that scenario (b) is the 'right answer' that predicts the correct values for those parameters for which an estimate is possible. Scenario (b) also happens to have more points than the other two scenarios, but based on only one example, it is not clear that this would be a sufficient criterion for finding the correct scenario.

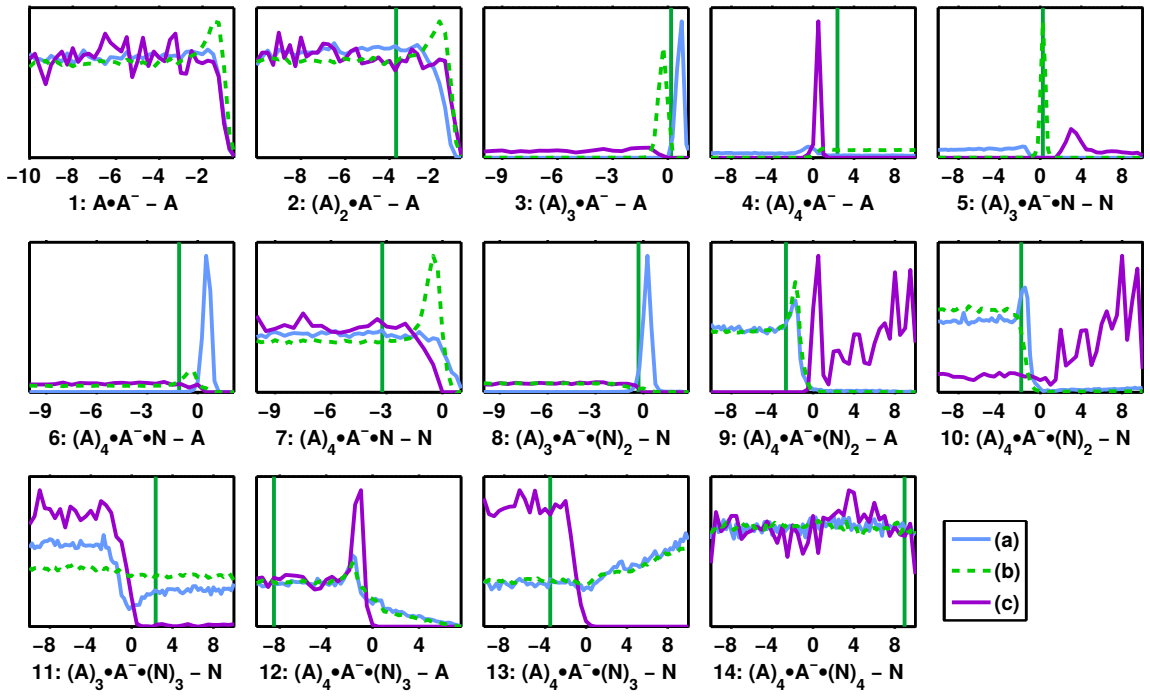

Figure S8: Posterior distributions of the base 10 logarithm of the evaporation rates (in units of s^{-1}) corresponding to the synthetic cluster distributions, separated into three scenarios. The background ammonia concentrations have been set to 5 ppt in the MCMC simulation. The vertical green lines corresponds to the quantum chemistry–based estimates used as input parameter values for producing the synthetic cluster distributions. (For coefficient number 1, the input value -17.8 is outside the range sampled by MCMC.) A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

S3. Additional plots related to the analysis of the experimental data

S3.1. Output cluster distributions for different background ammonia options

The output cluster distributions from the MCMC simulation are presented in the main text for the case with an assumed background ammonia concentration of 5 ppt. Figs. S9 and S10 show similar plots for an assumed background ammonia concentration of 1 ppt and varying background ammonia concentrations, respectively. In all of these simulations, the output concentrations represent very well the main trends with respect to both ammonia and sulfuric acid vapor concentrations.

S3.2. Posterior distributions for the ion production rate, wall loss coefficient and background ammonia concentration

Figure S11 presents the posterior distributions of the ion production rate and wall loss coefficient from the MCMC simulations using the experimental cluster distributions and the three different options for treating the low ammonia concentrations. For the MCMC simulation where the background ammonia concentrations were varied, the corresponding posterior distributions of each ammonia concentration are shown in Fig. S12.

S3.3. Posterior distributions separated into different scenarios

Figures S13-S18 show the posterior distributions of Figs. 4, 5 and 6 of the main article corresponding to MCMC simulations with a fixed background ammonia concentration separated into the alternative solutions (A)–(E) of Table 1. The posterior distributions were split into the different solutions similarly as described in Sect. S2.6, for the simulation with 1 ppt based on coefficient number 6 and for the case with 5 ppt ammonia based on coefficient number 5.

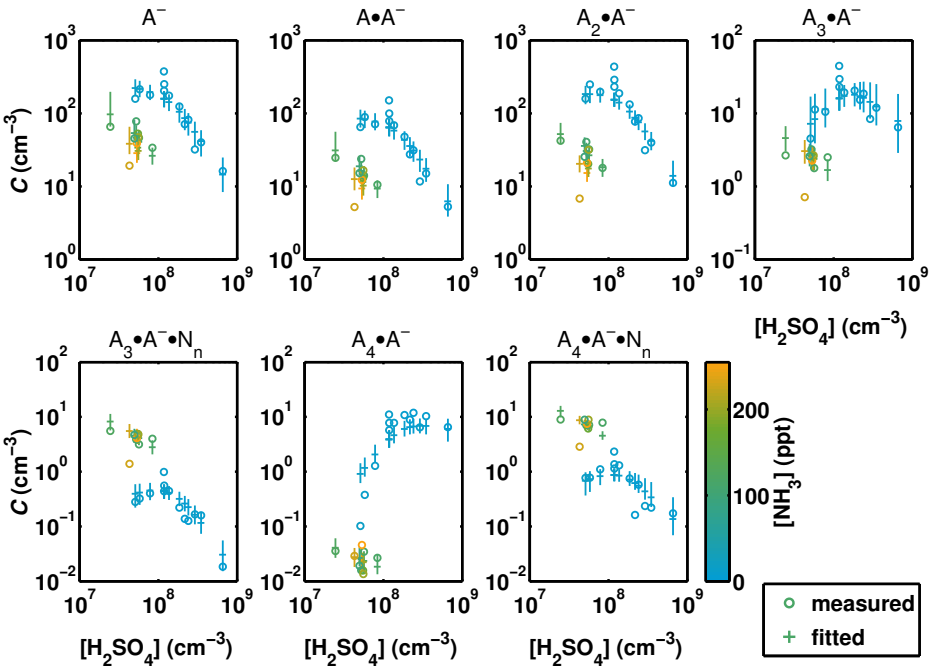

Figure S9: Cluster distributions measured at CLOUD and the corresponding modeled cluster concentrations from an MCMC simulation where evaporation rates, fragmentation probabilities, the ion production rate and the wall loss rate are varied, and the background ammonia concentration is set to 1 ppt. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

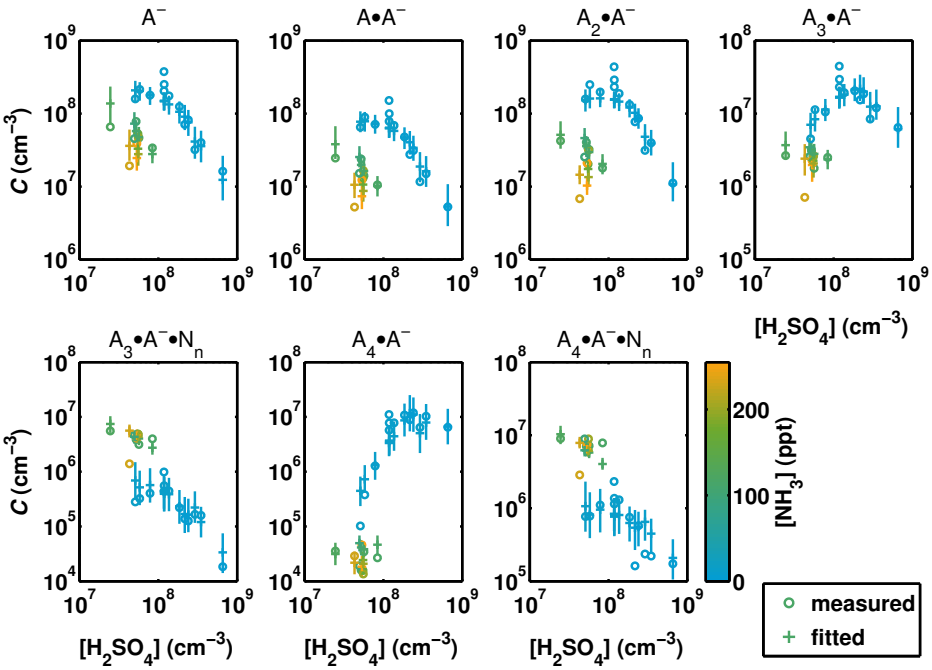

Figure S10: Cluster distributions measured at CLOUD and the corresponding modeled cluster concentrations from an MCMC simulation where evaporation rates, fragmentation probabilities, the ion production rate, the wall loss rate, and the all below-detection-limit ammonia concentrations are varied. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

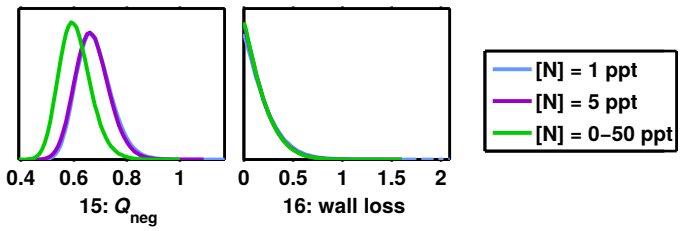

Figure S11: Posterior distributions of the ion production rate rates (in units of $cm^{-3}s^{-1}$) and wall loss constant (in units of $10^{-3}s^{-1}$) corresponding to the experimental cluster distributions and different options for treating the background ammonia concentration in the experiments where it was below the detection limit and therefore unknown. A stands for H_2SO_4, A^- for $HSO_4{}^-$ and N for NH_3.

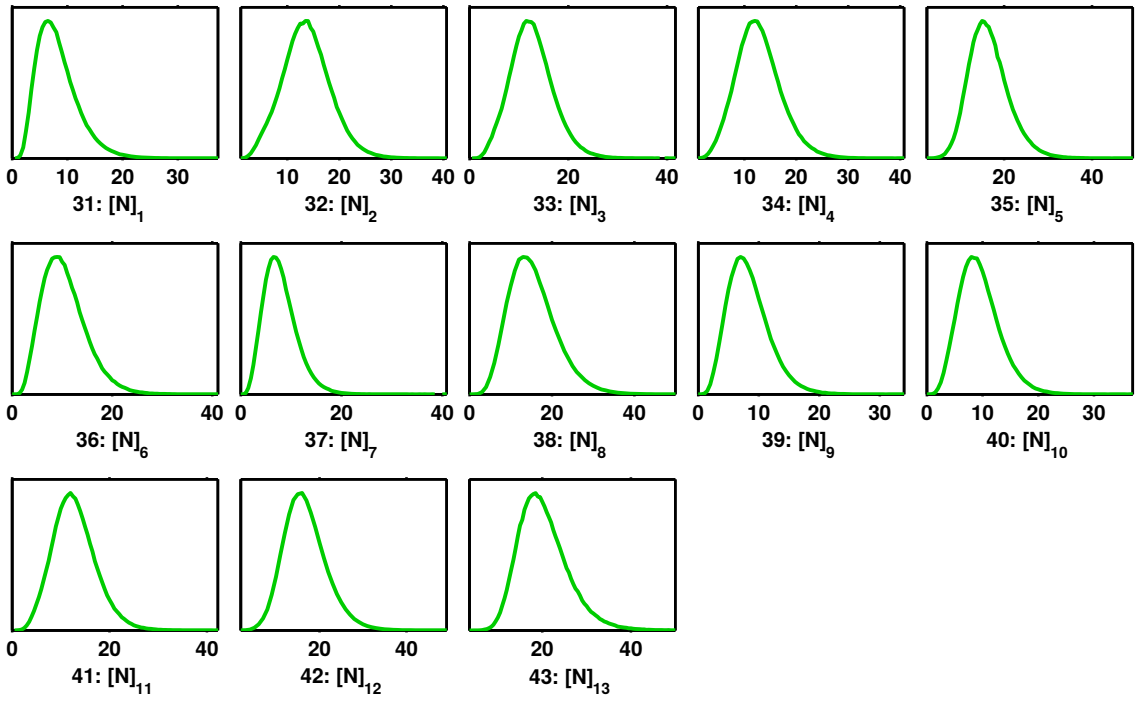

Figure S12: Posterior distributions of the background ammonia concentrations (as mixing ratios in ppt) corresponding to the experimental cluster distributions and and the MCMC simulation where the background ammonia concentrations were varied. A stands for H_2SO_4, A^- for $HSO_4{}^-$ and N for NH_3.

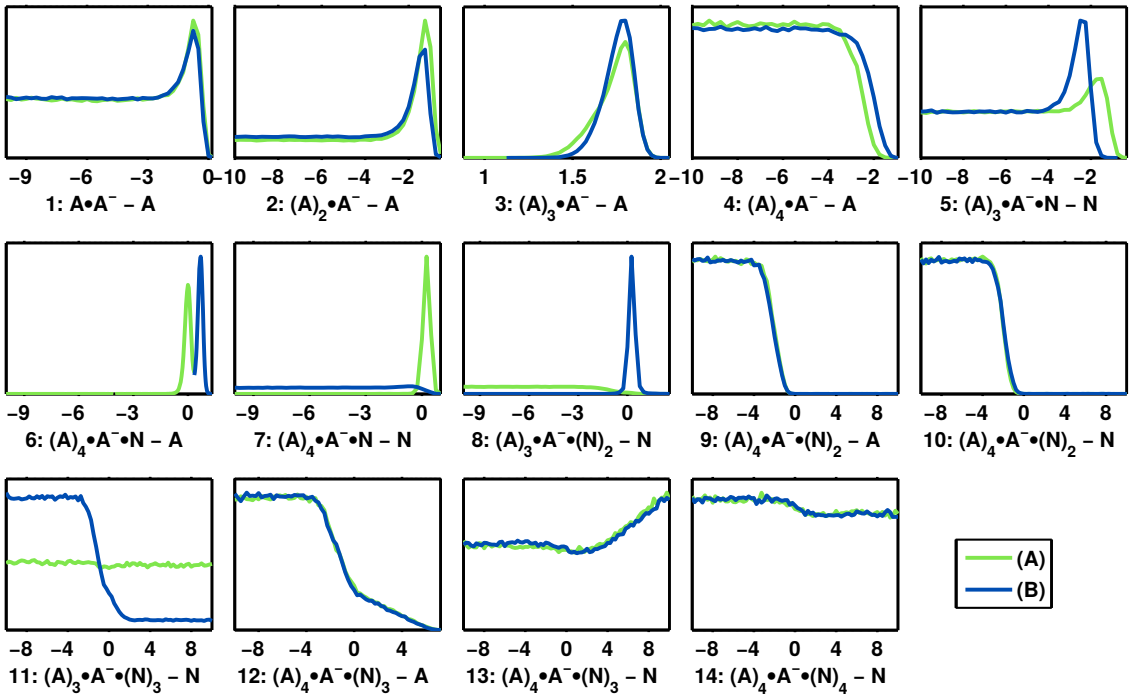

Figure S13: Posterior distributions of the base 10 logarithm of the evaporation rates (in units of s^{-1}) corresponding to the experimental cluster distributions and an assumed background ammonia concentration of 1 ppt, separated into two scenarios. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

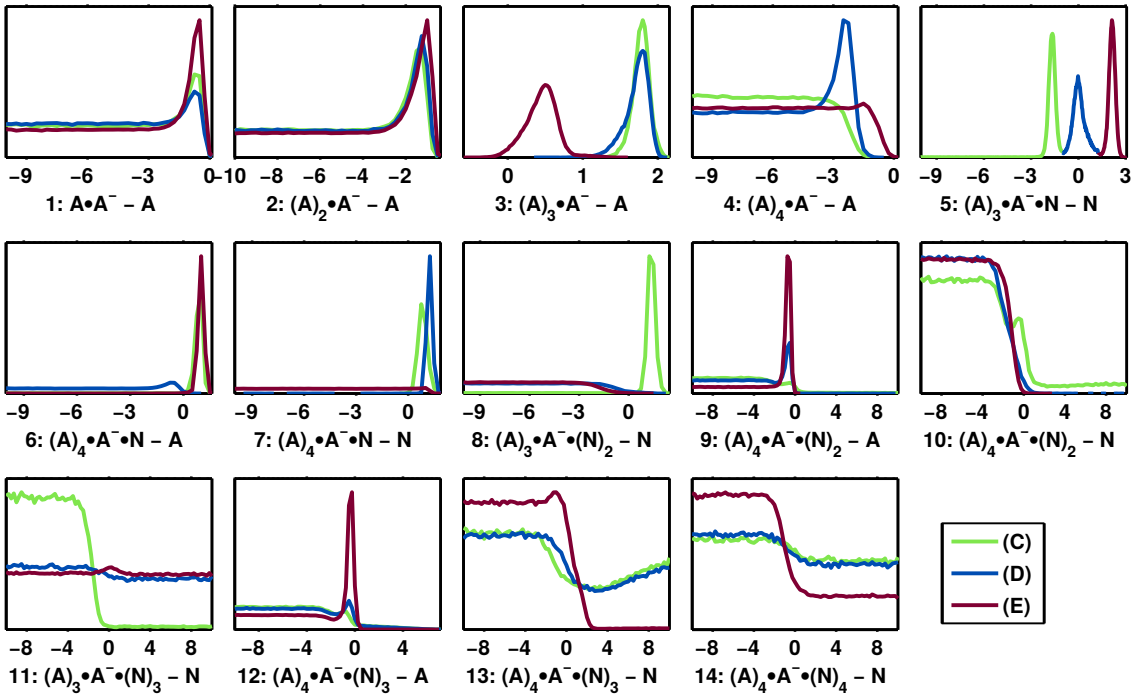

Figure S14: Posterior distributions of the base 10 logarithm of the evaporation rates (in units of s^{-1}) corresponding to the experimental cluster distributions an assumed background ammonia concentration of 5 ppt, separated into three scenarios. A stands for H_2SO_4, A^- for HSO_4^- and N for NH_3.

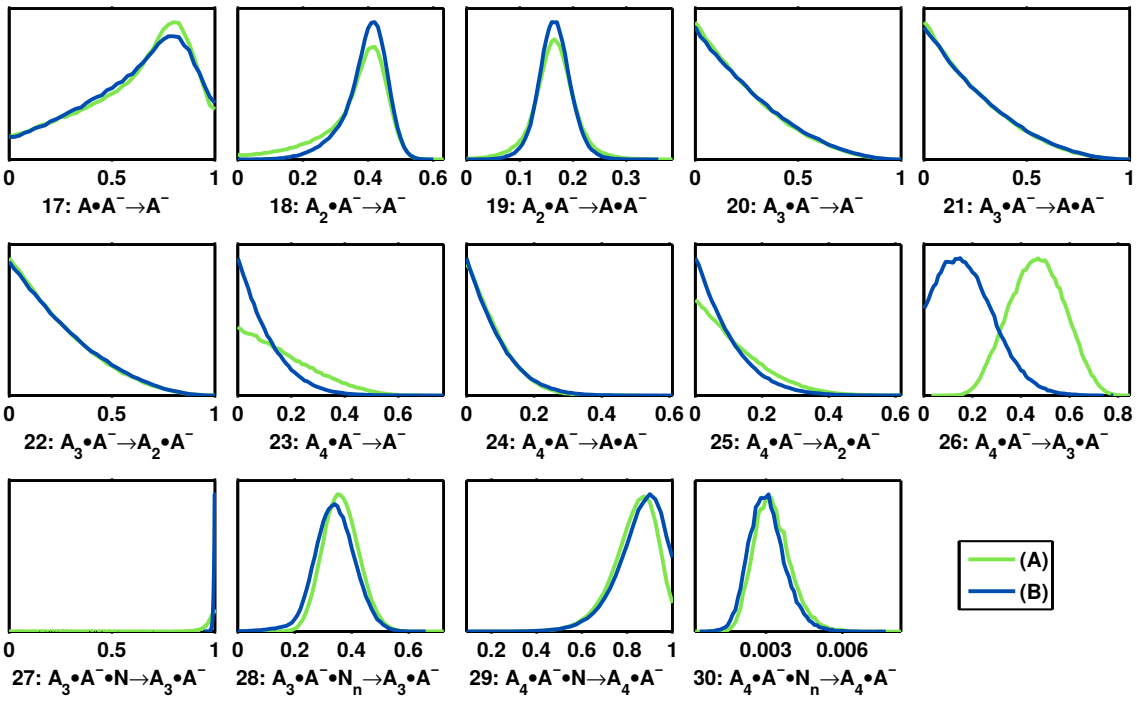

Figure S15: Posterior distributions of the fragmentation probabilities of clusters in the mass spectrometer inlet corresponding to the experimental cluster distributions and an assumed background ammonia concentration of 1 ppt, separated into two scenarios. A stands for H_2SO_4, A^- for $HSO_4{}^-$ and N for NH_3.

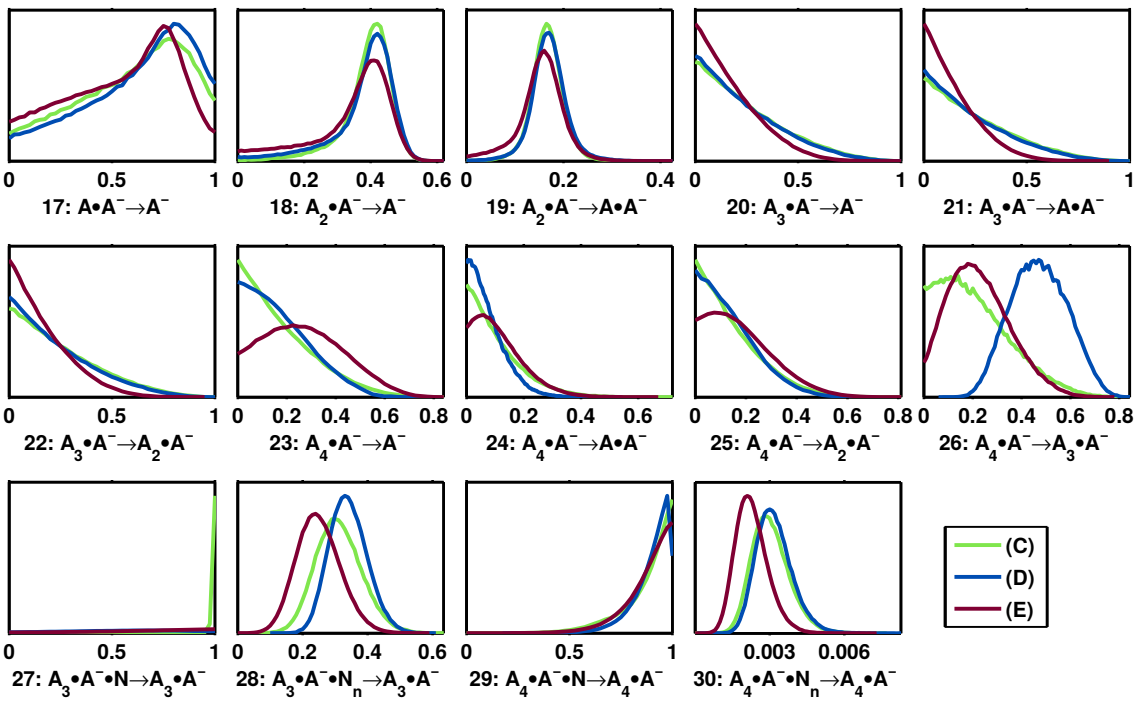

Figure S16: Posterior distributions of the fragmentation probabilities of clusters in the mass spectrometer inlet corresponding to the experimental cluster distributions and an assumed background ammonia concentration of 5 ppt, separated into three scenarios. A stands for H_2SO_4, A^- for $HSO_4{}^-$ and N for NH_3.

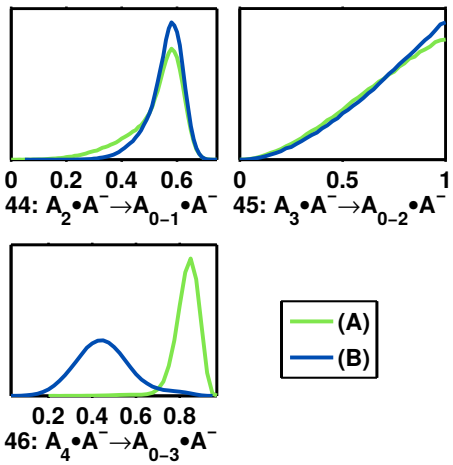

Figure S17: Posterior distributions of the total fragmentation probabilities of the $HSO_4^- \cdot (H_2SO_4)_{2\text{-}4}$ clusters corresponding to the experimental cluster distributions and an assumed background ammonia concentration of 1 ppt, separated into two scenarios. A stands for H_2SO_4 and A^- for HSO_4^-.

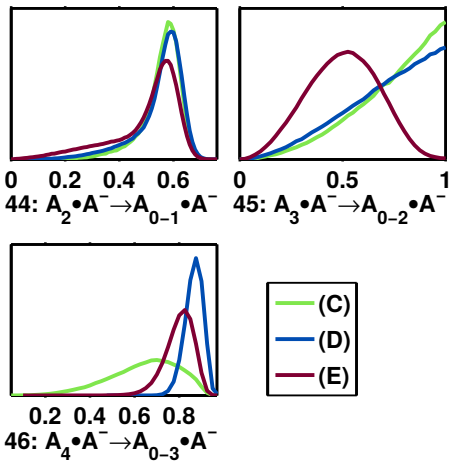

Figure S18: Posterior distributions of the total fragmentation probabilities of the $HSO_4^- \cdot (H_2SO_4)_{2\text{-}4}$ clusters corresponding to the experimental cluster distributions and an assumed background ammonia concentration of 5 ppt, separated into three scenarios. A stands for H_2SO_4 and A^- for HSO_4^-.

References

Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *J. Comp. Graph. Stat.*, 7(4):434 – 455.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472.

Ortega, I. K., Olenius, T., Kupiainen-Määttä, O., Loukonen, V., Kurtén, T., and Vehkamäki, H. (2014). Electrical charging changes the composition of sulfuric acid–ammonia/dimethylamine clusters. *Atmos. Chem. Phys.*, 14(15):7995 – 8007.

Su, T. and Chesnavich, W. J. (1982). Parametrization of the ion-polar molecule collision rate constant by trajectory calculations. *J. Chem. Phys.*, 76(10):5183 – 5185.

ter Braak, C. J. F. and Vrugt, J. A. (2008). Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435 – 446.