Reviewer report Uruci et al., 2023, AMT

Uruci et al. extended an existing algorithm to derive particle volatility information from the combination of thermal evaporation in a thermodenuder (TD), isothermal evaporation in a dilution chamber (DC), and yield experiments (YE). They used artificial data to evaluate the performance of their algorithm.

The topic is suitable for publication in AMT and highly relevant for a broad audience in atmospheric science. The overall presentation is good, and the descriptions are generally easy to follow. But there are two major issues that must be address prior to publication.

Major comments

1) From how I understand the generation of the artificial data set, the authors use circular reasoning when evaluating their algorithm.
   In their algorithm, they compare the "measurement" data (yield curve, thermogram, and areogram) with a lookup table of yield curves, thermograms, and areograms generated for a large number of combinations of VBS, enthalpy of evaporation ($\Delta H_{evap,}$), and mass accommodation coefficient (α_M) values. The calculated curve within 5% of the "measured" values are picked and the underlying VBS, ΔH_{evap}, and α_M combinations are presented.
   To generate "measurement" data for a known set of VBS, ΔH_{evap}, and α_M values they start with values derived from yield experiments and then generate a thermogram and an areogram using the same thermodenuder and evaporation model as in their algorithm.
   When they now compare these generated curves with the look up table, they will, of course, find good matches if the input parameters for the measured data (true VBS) were covered in the lookup table generation (see also specific comments 2 and 3). The good agreement and narrow range of the <5% solutions only shows that there is low ambiguity in the method. I.e., there are not many combinations of values far away from the true ones that produce matching yield, thermogram, and aerogram curves. In other words: if you use the values a, b, c to calculate thermograms and areograms, the algorithm will tell you that you used the values a, b, c if you included a, b, c when calculating your lookup table. To be blunt, the authors just showed that their equations work both ways. But they have not shown how well their method works with data that was not calculated with the model used in the algorithm.

2) The manuscript closes with the recommendation to use the combination of YE, TD, and DC data for future parametrisation. The manuscript did not convince me that the addition of YE data truly improves the results. Yes, the results using all three data sets look good. But nowhere do the authors show that their results are better than those from the method of Karnezi et al. (2014) with just TD and DC data. How much differ the results (combination of values for VBS distribution, ΔH_{evap}, and α_M) when only TD and DC data is combined vs using all three (TD, DC, and YE)? Unfortunately, the used data sets all derive the thermogram and areogram data from the yield data (see first major comment). Thus, I am not sure if this test will really be convincing or again simply show that the algorithm in itself is sound.
   However, the authors need to present stronger arguments why the inclusion of YE data is beneficial, especially considering the much higher experimental effort needed to obtain YE data.

Specific comments

1) Line 189: Why was sum(α_i)<1 chosen as a criterium? There is no reason for that as α_i are stochiometric coefficients and not mass or mole fractions. The true VBS of case B shows a violation of that rule (10^3 + 10^4 bin signal is already >1).

2) Following up on the previous comment: Because of this rule, the lookup table combinations do not cover the true VBS values in case B and more discrepancies are seen, especially in the areogram as that is most affected by the higher C* bins. It seems that the α_M value may be compensating the absence of the highest volatility bin somehow. This behaviour should be investigated as it has implications for the role of the α_M parameter in the algorithm which may not be desired.

3) Why was Case B only tested with 4 VBS bins? 60% of the signal is assigned to the 10^4 bin which is not part of the lookup range. Comparing the estimated VBS distributions from case A and B one could argue that the estimations for Case B (blue, see Fig R1) are more similar to the true VBS distribution in case A (red) than to the true distribution of case B (yellow). Since the "true" VBS distributions in case A and B are derived from the same SOA data, one could come to the conclusion that the algorithm wants to find a solution close to the true case A VBS distribution. What are the authors thoughts on such reasoning? Would using more VBS bins (and including the 10^4 bin) change this behaviour? I.e., would the algorithm suggest an estimate more similar to the "true" case B VBS values?

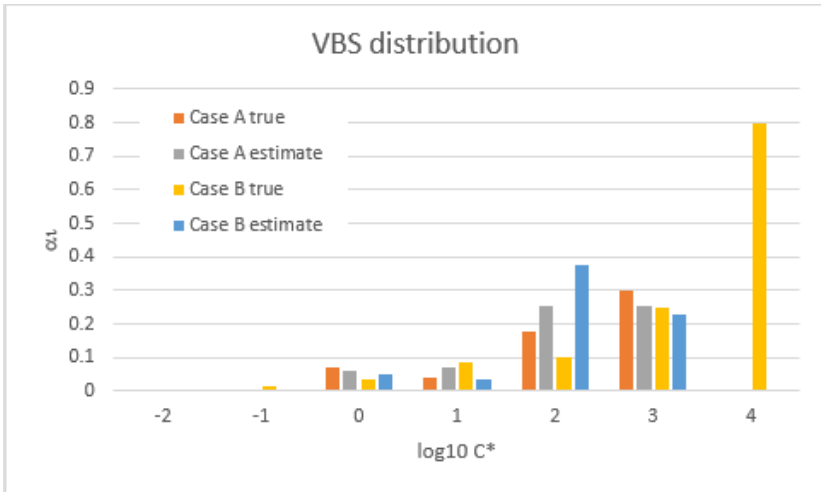

Figure R1: True and estimated VBS distributions for cases A1 and B1 visualised using values presented in Table 1 of the main manuscript.

4) Why was case A only tested for shifting to lower volatility bins? Especially, since the alternative parametrisation from the original paper (= case B) has very strong contributions from the 10^4 bin. Also, what results would be obtained if the full range of VBS bins (10^{-2} to 10^4) was used?

5) Lines 482ff: The authors need to define more clearly what they mean by "robustness of their algorithm". Are they primarily interested in predicting yields? Then the algorithm indeed seems robust and reliable. But the tests with the shifted volatility range show that for case C a completely different VBS distribution recreates the yield curve and areogram as well as the true VBS distribution. This behaviour could be called being subjective to the choice of input parameters by the user – so very much not "robust". How would the user know if the yields are "right for the wrong reasons"? I.e., which volatility range would they choose without prior knowledge?
In the case C2, it seems that the lower ΔH_{evap} compensates the shift in volatility (again this should tell as something about the mechanism of the algorithm/model). It is hard to predict how these values will behave when they are used in a different context (e.g., new particle formation in regional model). They do distort the shape of the thermogram somewhat which could be used as an indicator for a "right for wrong reasons" case. But what would be objective criteria for "too much deviation" to identify a not that good solution?

6) After (hopefully) showing that the YE data really improves the predictions, I wonder if the combination of TD and YE data works as well as the combination of all three. I.e., do the DC and YE data sets essentially cover the same aspects of the underlying true values? The aim of the question is: What should be measured to obtain the most reliable VBS distribution, ΔH_{evap}, and α_M combination? Especially when considering the time and effort needed for the measurements.

7) What about when the method is applied to real measurement data and that there was a process not covered in the model (e.g., the occurrence of thermal decomposition in the TD which shifts the thermogram towards apparent higher volatility). What would the algorithm make of that? Would the user be able to see that something is not right? Or would everything look fine, and the user will base their evaluation on incorrect VBS, ΔH_{evap}, and α_M values?

8) Line 63ff and 325ff: The method by Stainer et al. and the assumptions for predicting the yield curves at different T ignores that with changing temperature the chemical formation pathways may change. Especially, HOM and/or dimer formation can be strongly affected and thus have a unexpected effect on the observed VBS and yield (e.g. Quelever et al., 2019; Gao et al., 2022). They authors should at least mention this aspect somewhere when discussing yield curves at different temperature.

9) Line 180ff: Should not the dilution ratio also play a role for the isothermal evaporation in a dilution chamber?

10) Line 190ff: The authors provide the number of combinations that need to be calculated. How does that translate to computational time/effort? Can this be run on an office PC at reasonable time? Can the lookup table be created once and then used for the comparison with multiple "measurements "?

11) Why was NMSE used to get the overall error to find the "closest" solutions, but in section 4.2 and later the solutions are compared using MNE?

12) Line 373ff: The wording here makes it sound as if the Experiment B data was from a completely different SOA system. But it is based on the same measured data as Experiment A. Only the number of VBS bins is changed. The authors should make this fact clearer.

13) Fig1 – 3 and S1 - S3: The information about the content of each panel in these figures is there. But labelling the panels with a, b, c etc in each figure will make it easier for the reader. Currently, only the Figure is referenced in the text and the reader has to figure out which of the panels is meant in the text.

14) When the authors compare the different cases, e.g., when adding the additional high c_{OA} data point, it is difficult to judge how much the reconstructed VBS distribution, yield curves, etc. really change from the base case. E.g., Fig 8a needs to be compared with the 25 °C panel in Fig 1 which has different x and y axis scaling. It would be very helpful to add the base case lines/bars to Fig 8 and 9. Especially the extrapolation of the base case to 200 ug m^{-3} should provide an even stronger argument why the extra data point is useful.

15) I need more information about the weighted averaging of the selected <5% solutions. When the averages are calculated, is each data point treated individually? I.e., data point 1 in solution 1 is close to the true data and gets a high weight. But data point 3 of solution 1 is far away (i.e., the slope of the solution is wrong). Does data point 3 then get a different weight than data point 1? Or do all data points of a solution get the same weight factor?

16) How many solutions usually are in the <5% group? Did this number differ between the investigated cases? E.g., were there less acceptable solutions when the "wrong" VBS range was chosen? If data points were treated individually, how much did the number of <5% solutions vary for the data points?

17) Following up on the previous comment assuming that each data point is treated individually: The authors could consider improving the visualisation of the range of estimates. Instead of a uniform grey area, they could indicate the density of solution curves with a colour scale. That would preserve the range of slopes of the solutions. Two examples of such figures are shown below.

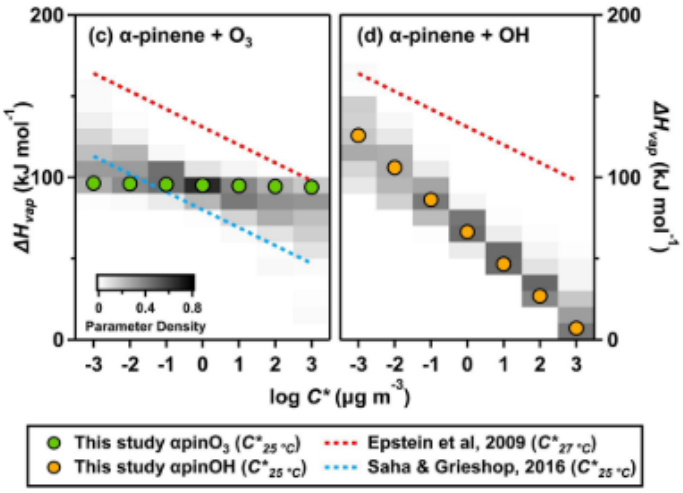

Figure R1: Lower half of Fig 2 in Li et al. (2019). Enthalpy of evaporation (ΔHvap) values from the best-fit simulations are presented with circles and the parameter density with shades of grey. The parameter density is derived by dividing ΔHvap and log C* space into grid cells, counting the frequency of simulated parameters inside each cell, and normalizing to the sum of frequency in each log C* column.

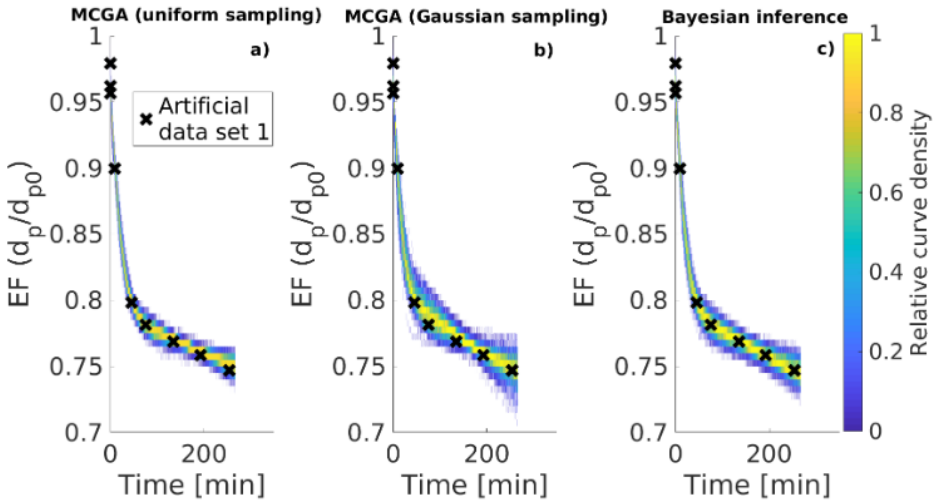

Figure S1: Relative evaporation curve density of the three different optimization schemes applied to artificial data set 1. The relative evaporation curve density is calculated by dividing the EF and time space to grids and counting how many of the simulated evaporation curve goes through a particular grid box. The counts are then normalized by the highest count in every time column. White color indicates that no simulation goes through that area. a) MCGA method with uniform sampling of the parameter space. b) MCGA with sampling similar to the Bayesian inversion method (see main text, Sect. 3) c) Bayesian inference method.

Figure R2: Figure S1 from Tikkanen et al., (2019)

Typos and language comments

1) Line 29: “The predicted yield uncertainty…” I find this sentence hard to understand.
2) Line 35: “IPCC, 2013” should be updated to IPCC, 2021 unless the authors are referring to something very specific which is only contained in the 5th assessment report.
3) Line 103: comma before “respectively”
4) Line 184: “SOA partitioning model” does that refer to the Eq. 1-3 in section 2.1?
5) Eq. 9: One sum goes up to “N_0” the other to “n”. Are these indeed different numbers or is it a typo?
6) Line 219: make it clear that these are the same models that are used by the analysis algorithm.
7) line 365: “wider range” is not a precise description here. Better say “at higher SOA concentrations”. Wider could also mean extending the range to lower concentrations.
8) Line 304 “(20 to 200 °C with a step of 5 °C but including TD MFR values greater than zero)” I do not understand the “but…” part. If “but only including” was meant, how could there even be MFR values <0?
9) Line 306 “[…] we used a higher resolution for the first 0.5h (step of 2 min), in which the evaporation is usually faster, and lower then (step of 10 min) up to 3 hours.” What is meant with the “lower then …“ part?

Line 445 “[…] instead of the 4 bins used in Test A1 and covering the 10^3 µg m^{-3} material.“ I am not sure about the meaning of the underlined part in this sentence.