Review of Karnezi et al, 'Simulation of Organic Aerosol using its Volatility-Oxygen Content Distribtuion during the PEGASOS 2012 campaign', ACPD 2018

This paper describes the simulation of ground-based and vertical profile measurements of organic aerosol mass loading and O:C ratio during the PEGASOS campaign. The multiple simulations use a variety of different parameterizations of functionalization and fragmentation, different for anthropogenic and biogenic SOA, and also vary assumed vaporization enthalpy. The offsetting nature of functionalization, fragmentation, and vaporization enthalpy result in no unique simulation scheme having clearly superior prediction skill metrics.

General comments:

In the introduction, you describe the Murphy et al modeling studies and point out that they found the simplest approach to parameterization (for anthro compound aging) did better than the more complex formulation (lines 72-75). It seems like this is your conclusion, too, based on your figures focusing on your simplest 1-bin parameterization. But, I didn't see that stated clearly as the conclusion. I suggest to explain somewhere why you focused on that 1-bin model for the figures, and if you conclude it was the best, make that clearer. In the end I came away feeling that no parameterization was necessarily superior but then was confused by your focus on the one in the figures. (Also in that discussion of Murphy 2011 and 2012: I would have found it easier to make the comparison between parameterizations if you included some metrics from each model on average volatility reduction and O addition at each step.)

The discussion of temperature dependence / enthalphy of vaporization confused me. In the introduction (lines 97-103), can you comment a bit on the stark difference between the Sheehan & Bowman and Murphy results you report? Because you also find a surprising result, it would be useful to understand what is known about why this has not been found to have consistent results in past studies. – Related, around lines 431-434: I don't understand your result. Why would aging mechanism change in response to deltaHvap? Or is this just that your optimization will adjust fragmentation correspondingly? This discussion could be clearer. – And, lines 472-484: these conclusions could use more discussion. Why is the low sensitivity surprising? Is this result mainly a function of not having explored a wide enough temperature range because of the season? If so, could say this at the end: e.g., "wintertime predictions would explore a larger range of temperature and thus better constrain delHvap."

I also had some confusion about the exact use of the back-trajectories. On line 115, are you simulating along the clustered average back-trajectory, or each one? Perhaps a bit more additional explanation of how the CTM incorporates the back-trajectories would be helpful. At line 142, I suggest to begin this paragraph with stating that these are 72-hour back trajectories (you mention this much later), and clarify that the ground site is the traj receptor – the list of times at the beginning was confusing. At line 161: "receptor site around Po Valley" –does this refer to the Zeppelin flights? Were 72-hr backtrajectories run for a series of points along the flight trajectory too? Or did you only use zeppelin measurements above the ground site and thus use the same back-trajectories? At the end of this paragraph, maybe segue to the next section with a brief recap of how they are fed into the chemistry schemes. (If you discuss this more above it may not be necessary to do much here)

Around lines 178-182: What is the reasoning for using different aging rate constants for some classes of OA and not others? Explain choices.

Around line 208-209: Why would aging of bSOA result in negligible change in volatility and an increase in O:C? And around line 215: why would bSOA be different than other OA aging?

Fragmentation parameterization, around lines 223-224 and following section: Were fragmentation probabilities simulated separatedly per time step or at all all times, or at endpoint, or at "average" timestep? Why would b not be dependent on O:C? At line 233: what is Figure C.4?

Lines 243-253: Aren't these skill metrics pretty standard definitions? Is it necessary to include the equations in the body text, or could they be in the supplemental info, or referenced?

On the binning of model results (lines 269-270): why 4 hours? Did you try single hours and how did it look? If 4 hours is all that works from the model perspective, why not bin your data onto the same time base as the model output that you are comparing too? It looks awkward to have the difference. The wording also made me wonder if you binned over the same hour of the day on multiple days – clarify how you did the averaging, how many days averaged, etc.

A big question I have about the observations is why the error bar suddenly changes above 700 m (the top 3 points in the vertical profiles), and there is a striking discontinuity in the data at those points as well. This needs to be addressed in the text (around lines 279-287, caption for Fig. 3) – what could explain this unusual behavior? You mention it is a single flight – how many flights do the other points represent? How were the locations different? Are those points truly better known (as suggested visually by the smaller error bars) or were the others just spatially or temporally more diverse? If the latter, maybe rethink how you determine your error bars, or how you portray the different points to show this. One thought: Are the top 3 points at lower temperature? Although, this would suggest a stronger delHvap sensitivity …

Around line 314: It doesn't make sense to me that aggressive functionalization would underpredict O:C – would it oxidize faster?

Around lines 347-354: The range in b's you find optimized for the various parameterizations is enormous – it seems to just be a correction factor you can use to tune your model. Can you contextualize what would be a reasonable value for b, e.g. by citing some experimental results that might help one choose? Same questions generically at the end of section 3.5 (line 412) about bSOA parameterizations – these also span a huge range. How could we decide which assumptions are more reasonable? If no data exists to constrain this, might you suggest some experiments that should be done?

On line 456, you refer to "more than a hundred" tested aging schemes, but your tables have a much smaller number. Do you mean the multiple tested b values in each scheme?

Specific comments / technical suggestions:

Line 26: define SOA (first instance)

Line 28: "contribute around 5%"

Line 29: define HOA, define (& explain) PMF-AMS

Line 32: "intricate interplay" is unclear, replace with something more descriptive?

Line 37: perhaps you meant to cite the IPCC 2013 WG1 report?

Line 68: “(PMCAMx-Trj) as the” (this is one of several places where I think a comma should be removed or added)

Line 75: “chemical aging assumed”

Line 82: “(2012), formation”

Line 98: “Bowman (2001) concluded”

Line 99: are you refering to SOA mass yields increase? Or mass concentrations?

Line 106: “The Po Valley”

Line 108: “evaluated by comparing the resulting 2D-VBS”

Line 115: “2011; 2012) simulating the”

Line 118-119:”aerosols, vertical turbulent dispersion, and area and”

Line 132: “from terrestrial ecosystems”

line 133: “emissions, and wildfire emissions are included following Sofiev et al. (2008a, b).”

line 267: “ground level is”

line 270: “concentrations of 4 hours instead of”? I think this is what you mean? But not clear why you chose 4 hours. See general comment above.

line 278: “measurement in Figure 3a … predictions agree with”

line 299: “error for altitudes”

line 311: “scheme underpredicted the”

line 313: “; 2012) about the”

line 461-462: “with similar average”

line 464: “evaporation of primary OA and subsequent”

lines 479-480: “small temperature differences for altitudes up to 600 m … this low sensitivity to … evidence that the higher values (150 kJ mol-1) are”

Fig 1: In caption, mention that these are 20 72-hour trajectories. On maps: can you make the map boundaries the same in both panels?

Fig. 2: As mentioned above, it feels strange to me that the binning is different for model and measurements – consider unifying? Maybe state in caption how many days of coverage this is to help contextualize the variability / error bars. And, could interrupt / truncate the y axis to miss all the white space up 0 – 0.4 and see the variation better.

Fig. 3: As mentioned, say something about the weird discontinuity at the top. Mention in caption how many measurement are averaged at each height (or find a way to show this with symbology in the figure)

Fig 2-4: somewhere, explain why you picked the 1-bin case for all comparisons.

Fig 6 caption: “Diurnally averaged (a)”