Review for: “Differentiation of primary and secondary marine organic aerosol with machine learning” by Chen et al., 2025 - ACP

Overview and recommendations

Chen et al., 2025 provides a technique for estimating marine POA by using a machine learning technique to estimate marine SOA. I found the layout and logic in this study confusing and hard to follow at times. My interpretation of the work is that ML was used to predict marine SOA, and then POA = OA - SOA. But the authors focused on filtering the data down to periods in which marine SOA were most likely and little marine POA was predicted during these times. It was not clear to me whether this was used to *train* the model and that then the model was applied to the whole time period, or if the model was only used in these highly filtered times. If the latter case (seems unlikely!), then the ML work seems unnecessary. I’ve tried to point out where I found the methods and application confusing in my comments below. I do think that if the authors can make their intent and applicability of methods more clear that this study could be published in EUGsphere.

Major comments

Why did the authors choose to predict SOA with the ML model, rather than POA? Neither seems inherently correct or incorrect but the logic should be explained.

Line 85: Downscaling hourly meteorological measurements to 10 minutes could introduce significant bias to the study, depending on how sensitive the ML predictions were to the various met parameters. Were any sensitivity studies performed on this assumption? Can the authors comment on any potential limitations from this assumption?

Sect 2.2 The filtering processes don’t make sense how they’re presented - for example the paragraph that begins “We then applied additional filtering processes to reduce the impact of POA production…” If this study is intended to use ML to predict SOA and then obtain POA via POA = OA - SOA….then what is done for the periods that are filtered out to exclude POA production times? Are those times assumed to be POA only or just removed from the dataset? Basically overall, this section reads as “We find the periods of time in which SOA is most likely and then predict POA from that .” Which then begs the question - wouldn’t POA be very low and why is a ML model needed? Presumably this is not the authors’ intent so clarity is needed throughout this section. Perhaps once SOA is quantified in the “SOA-only” times, then the authors deploy their methods to the whole data set?

Assuming that the authors trained the model on SOA production periods only and then applied the model to the whole dataset - have similar aerosol/ML studies been performed? Does it seem valid that the model can easily sort out anthropogenic influences to correctly predict marine SOA (mSOA) and therefore mPOA? Another way of asking this question is: what happened to the time periods with anthropogenic influence? How is the model supposed to know that everything that is not mSOA could be a mixture of mPOA and aSOA/aPOA? Again, what portions of the dataset were used needs to be clarified throughout.

Line 155 - the SVR model is used to predict OA. I thought that OA was available via AMS data, and that the model is being used to predict SOA to find POA.

Section 3.3 - I'm confused where the k_HTDMA and spread values come from. Is this from the observations? The ML model? Other? This section seems somewhat out of place at the moment.

Minor comments

The introduction refers to numerous marine studies but does not provide any geographical context to any of them. It would be helpful to understand to what regions each studies' conclusions are most likely appropriate to.

Lines 78-79: Please note what was used to provide meteorological conditions.

Line 101-110: Please provide citations for the assertions about SVR's skills and merits.

Line 113: please explain why 2015 was used; are there any indicators that POA and SOA could or could not have shifted in any way over the entire time period (2009-2018)? For example, given that 27.8% of SOA periods happened in the 4-year challenge period (2015-2018) but this period represents ~44% of the total time, that seems like it could indicate that a change in aerosol balance between POA and SOA may have occurred over the time period. Did the authors attempt using data from each year in the training and challenge datasets as a sensitivity study?

Figure 2 - is the y-axis mislabeled? Isn't it predicted SOA + POA_bg?

Line 172 - 183 and Figure S5: Figure S5 would be more meaningful with an indication of % of data points the y axis represents. Having 25,000 negative values as the optimal POA_bg assumption still seems quite large. Why are there so many negative values - is this to be expected with the ML methods deployed? How were the negative values handled?

Technical comments

Line 16-17: " …introduce a machine learning approach to differentiate and quantify the contribution of marine POA from marine secondary organic aerosol (SOA)." - the phrasing here makes it sound like the team is quantifying the contribution of POA that is formed from SOA. Suggest rephrasing (something like quantify the contribution of marine POA to total OA).

Line 89: MHD - I assume the Mace Head site? Please be sure to define.

Line 120 - please provide reference for FCM

Line 139 - "periods" not "period"

Throughout the manuscript (including figures) the authors refer to "train" or "training" data - I believe training is more typically used. Please use one or the other throughout.

Figures and Tables

Please provide in the SI a key for the abbreviations used on Figures S1, S2 and elsewhere (e.g. bih, rhum…)