We thank the referees for their thoughtful and constructive comments. Please find the detailed response to each comment, including text changes to the manuscript. Our response is in blue font while changes to the manuscript are underlined. The line numbers used here refer to those in the change-tracked manuscript file.

Reviewer #1

This paper applies a machine learning approach to try to interpret the POA and SOA data that is observed at the Mace Head atmospheric observatory. These have been reported in previous publications, but this uses a different approach to try to analyse and interpret a long-term dataset.

There are advantages to this type of approach over other methods like PMF, in particular computational cost, interpreting dependencies on independent variables and the ability to gap-fill where necessary. However by using a supervised approach, this can be said to be less objective. This is a particular concern because quite a lot of a priori information is brought into the analysis, in particular how POA is defined, where a number of assumptions are made based on previous experience analysing data. So that being the case, it is perhaps not surprising that the model performs so well. Although it should be noted a certain amount of objectivity is brought in through the use of clustering to generate the representative POA. Furthermore, given that this specific site has proved fairly unique in producing observations such as these (owing to its location), it would remain to be seen whether this technique was applicable to other sites.

All that said however, I still found this an interesting and informative paper. The relationships between the POA and SOA and the other variables in particular, but it was also interesting comparing these data products to the outputs of the HTDMA data, which could have implications for marine CCN populations. I do have some concerns, perhaps the biggest being that I didn't find enough technical information accompanied how the models had been set up (see below) but besides that my comments are pretty minor and I recommend publication after these are addressed.

Response: We thank the reviewer for the insightful and positive comments. Indeed, we acknowledge the model itself might be suitable for this specific site, but we believe the similar methodology is can to be applied to other locations.

Major comments:

The article, as it stands, is severely lacking in the technical details of how the FCM and SVR methods were applied to the data. While it is not necessary to post the code that was used, certain technical details were missing from the article and the supplement. Specific things relating to the FCM I found missing were the distance metric (while Euclidian distance is the default it cannot be inferred), whether there

was any pre-treatment to weight or linearise variables, and how the optimum number of factors were determined. Likewise with the SVR, more explanation should be given regarding the specifics of the input data and parameters. More details such as these should be included, in the supplement if necessary.

Response: We thank the reviewer for these important technical points. We have now included the requested details in the revised manuscript.

We have provided a detailed description of the settings for the FCM model in the revised section 2.3.

Line 154: Fuzzy C-Means (FCM) is a clustering algorithm that enables the grouping of data points into multiple clusters with varying degrees of membership. In this study, the input variables for the FCM model included chemical components (SeaSalt, Org, NO_3, SO_4, NH_4, MSA, and BC), as well as meteorological parameters — temperature (temp), relative humidity (rhum), wind speed (wdsp), and wind direction (wddir). We randomly selected 10,000 samples and retained only those with positive concentrations for all chemical components. The data were then log10-transformed and Z-score standardized. The FCM model was configured with 5 clusters, a fuzziness exponent of 1.2, and Euclidean distance as the distance metric.

The SVR model settings are detailed in the revised Section 2.2.

Line 118: SVR was chosen for its generalizability in handling small datasets and its resistance to overfitting (Ghimire et al., 2022; Juang and Hsieh, 2009). Unlike tree-based models like random forest (Breiman, 2001), SVR model can predict continuous values (Ma et al., 2003; Tang et al., 2024). The hyperparameters, including the penalty coefficient (C) and gamma (γ) of the Radial Basis Functions (RBF) kernel were tuned via grid search. The model targeted OA concentration, using nss-SO_4, MSA, NH_4, and meteorological parameters (temperature, relative humidity, boundary layer height, wind direction, and pressure) as predictors. We also included hours of the day to capture diurnal variations. Predictors not directly linked to secondary production, e.g., sea salt, NO_3, and BC, are excluded to avoid over-fitting and ensure generalizability, even though including these might have enhanced the model performance in the training dataset. Wind speed was used to select the SOA production period, therefore, it was not suitable as a predictor. A summary of the variables employed as predictors is shown in Table 1.

Line 128: The SVR were trained using 'tidymodel 1.3.0' framework using R programming software (version 4.4.3)

Reference

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/a:1010933404324, 2001.

Ghimire, S., Bhandari, B., Casillas-Pérez, D., Deo, R. C., and Salcedo-Sanz, S.: Hybrid deep CNN-SVR algorithm for solar radiation prediction problems in Queensland, Australia, Engineering Applications of Artificial Intelligence, 112, 104860, https://doi.org/10.1016/j.engappai.2022.104860, 2022.

Juang, C.-F. and Hsieh, C.-D.: TS-fuzzy system-based support vector regression, Fuzzy Sets and Systems, 160, 2486–2504, https://doi.org/10.1016/j.fss.2008.11.022, 2009.

Ma, J., Theiler, J., and Perkins, S.: Accurate On-line Support Vector Regression, Neural Computation, 15, 2683–2703, https://doi.org/10.1162/089976603322385117, 2003.

Tang, D., Zhan, Y., and Yang, F.: A review of machine learning for modeling air quality: Overlooked but important issues, Atmospheric Research, 300, 107261, https://doi.org/10.1016/j.atmosres.2024.107261, 2024.

Minor comments:

L185: While the Monte Carlo bootstrapping is a powerful method of assessing random uncertainties in the data, it does not address the issue of systematic uncertainties, which given that these AMS OA types have been reported from few locations, is potentially substantial. This should be noted.

Response: we thank the reviewer for pointing this out, we have now included the following statement in the revised manuscript to highlight the potential uncertainties associated with Monte Carlo bootstrapping.

Line 225: “While it should be noted that the Monte Carlo bootstrapping is used to assess the random uncertainties, potential uncertainties associated with possible systematic uncertainties require further investigation.”

L224: A note of caution should be added regarding treating nss-SO4 as a secondary marker because while it is indeed formed through secondary timescales, the precursors and formation timescales are different to SOA and the relationship between the two is inconsistent when looking at terrestrial environments. There is a case to be made for them being correlated in the marine boundary layer if they are both assumed to originate from biological activity in the sea surface, but explicit correlation should not necessarily be expected.

Response: We thank the reviewer for the important comment. Indeed, the nss-SO4 formation timescale and dynamics are expected to be different to marine SOA. However, nss-SO4 were selected as a secondary marker as it is almost ubiquitous in the marine boundary layer and represents the most well-known secondary formation pathway of marine secondary aerosols. Actually we have not constrained the marine SOA to be correlated with nss-SO4, but the decoupled correlation between marine POA and SOA with nss-SO4 have led further confidence on the OA apportionment.

MSA, another typical marine secondary marker, seems to be equally important as nss-SO4. Although MSA and nss-SO4 have different formation pathways when DMS underwent atmospheric reactions, their high correlation suggests the marine SOA is also expected to be highly relevant in terms of correlation. We have now included additional statement in the revised manuscript to highlight possible bias by using the selected set of variables, including nss-SO4.

Line 194: "It should be noted that the nss-SO4, which albeit being marine secondary species, exhibit different formation dynamics and timescales with marine SOA, which might induce some extent of uncertainty. However, both marine SOA and nss-SO4 might originate from marine biological activities. Furthermore, marine air masses arriving in MHD are expected to advected over Northeast Atlantic for several days. The use of nss-SO4 can also be supported by the high correlation between nss-SO4 and MSA, which exhibit different atmospheric formation dynamics."

Figure S3: I actually found this one of the more interesting aspects of this work, and I would consider moving it to the main article.

Response: We thank the reviewer for the suggestions, we have now removed the Figure S3 to the main text.

Figure S7: The figure caption doesn't currently make sense. Reword.

Response: We have now reworded the caption of Figure S7:

Figure S7. Observed OA versus predicted (SOA +POA_{bg}) for (a) training, (b) validation and (c) test datasets. Data density is illustrated using a color gradient with darker colour indicating lower data density. Black lines denote the 1:1 correspondence lines, blue lines represent regression lines. The ML model does not use MSA as predictors.