Authors' Response to Reviews of

Deep learning representation of the aerosol size distribution

Donifan Barahona, Katherine Breen, Karoline Block, Anton Darmenov
Geoscientific Model Development, 2025

RC: ***Reviewers' Comment***, AR: Authors' Response, □ Manuscript Text

We appreciate the constructive comments. We have clarified these points in the revised manuscript, as detailed below.

1. General Comments

RC: ***How does the trained model perform during a different time period? Aerosols in the 90s were much higher than they are today, is the model that is practically only driven by temperature (and air density, which does not change much with climate change) able to capture that time period? More generally, what is the validity range of the model, given its training dataset?***

AR: We appreciate the reviewer's comment. MAMnet was tested using data from a different year not used during training, showing that it is able to generalize beyond its training data. MAMnet is not intended to fully emulate the modal aerosol model (MAM7). Rather, its purpose is to map the simulated aerosol mass across species into a 7-modal aerosol size distribution (ASD). This is arguably a simpler task than emulating the full range of aerosol processes represented in MAM7, since the aerosol mass fields used as input already encapsulate the effects of meteorology, clouds, as well as trends in aerosol emissions. The mass-number relationship on the other hand is not expected to depend strongly on such factors, since in many cases it can be approximated to some degree using prescribed formulations for the ASD. Therefore it is likely that it can be learned only from the relative abundances of the aerosol species and their vertical variation (encapsulated in the density and temperature). As MAMnet performs well using data from a different year of simulation, not used during training, we don't anticipate a major effect of the time period.

RC: ***How much computational time is saved? There is no MERRA-2+MAM model, but the comparison between GEOS, GEOS+MAM, MERRA-2, and MERRA-2+MAMnet should be able to provide the necessary information.***

AR: Thanks for the comment. We haven't completed the implementation of MAMnet in an online model (i.e., GEOS+MAMnet), and focused instead on its offline performance. Whether such an implementation leads to a speed up depends strongly on technical details such as code base (i.e., Python vs. Fortran), input and output speeds, and parallelization performance. Expanding on these details would deviate substantially from the focus of this work, and it is left for future research. It must be mentioned that MAMnet not only offers potential speedup, but also enables a better estimation of the ASD where limited information is available, like in satellite retrievals and data assimilation systems.

RC: ***I guess it is MAM7 used in this work; shouldn't you be using this name to separate it from other MAM versions?***

AR: Yes, we agree and it has been corrected in the revised work.

RC: ***I am really surprised that only temperature and air density have been used for the meteorological state. I***

would expect that 3-dimensional wind fields (long-range transport), clouds and precipitation (wet removal, CCN, activation), and surface type (dry deposition) would be of key importance. Clouds can be also important for sulfate formation in the aqueous phase, and then cloud evaporation should affect sulfate size distribution. How can a model be accurate without these processes included?

AR: We appreciate the reviewer's comment.

As mentioned above MAMnet maps the simulated aerosol mass across species to the ASD, but it is not intended to simulate the full range of aerosol processes. Instead it leverages such physics from the bulk model. We agree that aerosol number concentrations and mixing states are influenced by meteorological and cloud processes in ways that differ from aerosol mass. However, our results suggest that the neural network can effectively learn the nonlinear relationship between aerosol species mass and number, likely by leveraging the relative abundances of different species within each mode. This may reduce the model's sensitivity to the explicit meteorological state.

This has been clarified in the revised work.

RC: ***The lifetime of a single species in MAM (e.g. SU) would depend by the removal rates in each mode, which differs in terms of mode solubility (a function of mode composition) and sedimentation velocity (a function of mode size). The NN training is implicitly using this information, but the NN application in a bulk model like GOCART does not have that distinction when calculating SU mass, so inherently SU is different across models by design. The NN will likely try to compensate that, but can you make a comment on this?***

AR: This is a valid point. The evolution of aerosol species in MAM7 depends on processes such as solubility and sedimentation, not explicitly represented in bulk models. As a result, applying MAMnet to bulk aerosol mass fields introduces uncertainty due to differences in model assumptions. However, MAMnet can be fine-tuned for such applications if needed. Furthermore, our comparison against observations also shows that when driven by assimilated aerosol fields MAMnet still produces realistic ASDs. This suggests that the network is able to learn a robust, nonlinear mapping from species mass to the ASD, even under varying model conditions.

This discussion has been added to the conclusion section.

2. Specific Comments

RC: ***Line 9: Replace "physical representation" with "aerosol microphysics representation". A machine-learned approach is not physics***

AR: The statement now reads: "Our model paves the way to improve the representation of aerosols in atmospheric models while maintaining the versatility and efficiency required in large scale applications"

RC: ***Line 24: "of the same size" should be "in the same bin". Bulk approaches allow particles in different bins to have the same size but different composition, e.g. sulfate vs. nitrate. Line 25: "they fail to distinguish" is too harsh, please replace with "they are not designed to resolve". They would fail if they would try to resolve ASD, but they don't.***

AR: We agree the definition is ambiguous. The statement has been rewritten as:

"The bulk mass approach predicts the transport and evolution of aerosols by tracking the mass concentration of individual chemical species [Jones et al., 1994, Langner and Rodhe, 1991, Ginoux et al., 2001, Chin et al., 2000]. It inherently treats aerosols as externally mixed, since each particle is assumed to consist of a single

chemical component or their surrogate [Riemer et al., 2019]. Because each species is typically represented by a single prognostic variable, the bulk approach is not designed to resolve the ASD or the mixing state, which are often prescribed from climatological data."

RC: ***Lines 38-39: "These models offer the most physically consistent representation of the ASD" is not necessarily correct, since modal models assume a shape of the size distribution per mode, typically a lognormal, which is an approximation of reality. One could argue that sectional models, which are even more expensive than modal ones, are better, since they can freely calculate the ASD shape without the need of a lognormal, but they also suffer from assumptions needed when moving mass and number from one section to another. Particle-resolved models might be the most realistic ones, but these are practically impossible to use in large-scale models. The point is that mentioning that modal schemes are the most physically consistent is incorrect.***

AR: This was referring to more sophisticated models. There is also typo where we meant moments "instead" of "modes". We agree that the way it is written is ambiguous. The statement has been corrected and now reads:

"More sophisticated aerosol schemes either compute additional moments of the ASD [Zhang et al., 2020], explicitly resolve it using a binned approach [e.g., Adams and Seinfeld, 2002], or represent it on a particle-by-particle basis [Riemer et al., 2019]. While these models provide the most physically consistent representation of the ASD, they are often too computationally expensive for operational forecasting and long-term climate simulations."

RC: ***Line 96: Which years were simulated, and 72 vertical levels up to what altitude?***

AR: The model top pressure is 0.01 hPa. The simulated years were 2001-2006. This is now clarified in the paper.

RC: ***Line 97: Please elaborate on the choice of 9 AM/PM UTC time for the output and especially the 12-hour frequency. Understandably this is a lot of output already, but I would argue that sampling any individual location just twice a day has a high probability to miss the diurnal variability of ASD. I would expect that 4 times a day would be the minimum reasonable sampling frequency, as a first guess.***

AR: Sampling at 12-hour intervals allowed us to use more data for training while still capturing differences between day and night. We agree that higher-frequency sampling could better resolve the diurnal cycle, but this comes at the cost of fewer training time steps due to memory limitations. However, this is not expected to be critical, as the relationship between mass and number likely exhibits weaker diurnal variability than the aerosol mass itself, which is already represented by the bulk model.

RC: ***Section 2.2.1: I do not follow the files counting and usage. 25 were "randomly selected without replacement for training" (what does that mean?), 10 were used "for the testing of the trained model", 100 were "not used during training" (how were they used?). What are these files? Each instantaneous output produces one file, so 2 per day, times 365 times 5 years files? If yes, what happens with the remaining thousands of files? And how many have been used for training?***

AR: Thanks for bringing this up. Yes, out of the thousands of files of the run we selected 25, at random for training, with no duplicates (without replacement). Each file represents global instantaneous output from the GEOS+MAM7 model. During training the loss is calculated on data not used to update the parameters of the network, termed validation loss, taken as 10 additional files. Since the validation loss still guides optimization choices it is considered part of the training step. The testing data is completely independent, and we have used 5 additional files taken from the year 2006, which was not used at all during training.

The reason that we select just a few files is that each one represents an already very large number of samples.

For our final training we used $N_s = 72 * 180 * 360 * 25 \approx 1.12 \times 10^8$ samples. Given this large number, this is assumed as representative of all possible mass-number combinations produced by MAM7, which is what the model needs to train. In our sampling strategy we maximize the number of samples the model could handle. Again we want to emphasize that a MAMnet does not attempt to learn the spatial distribution of the ASD, just its relation to the aerosol mass.

This has been clarified in the work.

RC: ***I see later (lines 139-140) stated "5 output files for training, 2 for validation" which makes even less sense. Please explain.***

AR: As there are several free parameters in the design of the neural network, the optimal architecture is found by doing a guided search, training on a small set of data, running for a fixed number of epochs (50 in our case). This was done using 5 files training and 2 for validation.

The optimization and training section has been moved to the Appendix for clarity.

RC: ***Figure 1: Please explain what MAMnet loss is. It is not referenced anywhere else in the manuscript. Also, why GOCART is mentioned? This figure is for the development of the NN, not its application. Isn't GOCART only used for application?***

AR: Thank you for pointing this out. The MAMnet loss refers to the minimum mean squared error between the network predictions and the corresponding MAM7 fields. We have clarified this in the figure caption. Additionally, we have removed the GOCART reference from the figure, as it is not used in this study.

RC: ***Table 3: Too many new concepts there which are not explained. Please help the reader understand what these are, or move this table in an appendix, if you consider it too technical to expand.***

AR: Thanks for the suggestion. We have moved Table 3 and the related discussion to the Appendix.

RC: ***Section 3: I would recommend adding a section 3.1 "evaluation against GEOS+MAM", similar as to what current section 3.1 says "evaluation against observations", instead of having it under the generic section 3.***

AR: Thanks for the suggestion, the headline has been added.

RC: ***Figure 2: Are these global means per layer? Assuming that yes, is this a good metric, especially for number concentration? Wouldn't doing this regionally be much more meaningful? I appreciate the zonal means and maps later, but my question stands. To be more specific, how can you say "systematic errors emerge" in line 199, without knowing whether this error is widespread or just some very large scattered errors that overwhelm the mean?***

AR: Thank you for the comment. We agree that number concentration varies greatly across regions, and that averaging globally can hide local differences. In Figure 2, we used global mean bias by vertical level to give an overview of model performance. To account for the large range in number concentration we calculated the bias on a logarithmic scale. We also included additional plots such as the zonal averages and spatial maps in later figures. Finally, the second panel in Figure 2, which shows spatial correlation, helps identify whether large errors spread across regions.

Regarding the statement on line 199 about "systematic errors," we have clarified in the revised text that this refers to consistent patterns in the bias. It now reads:

"In summary, MAMnet captures overall modal number patterns well, but errors remain in the Aitken mode, primary carbon and coarse dust."

RC: ***Figures 2-3, regarding mass concentrations: what is the model performance in terms of mass conservation? The results per mode do not need to conserve mass, but per species across modes mass conservation is paramount. Thinking even further, how will the mass conservation concept be applied when using MAMnet in production runs? Lines 253-262, and Figure 6: These are an evaluation against MERRA-2, not observations, as the title of section 3.1 denotes. This whole paragraph and figure are a good conclusion in the discussion just before this section, so moving it right after line 247 and before section 3.1 starts should be considered.***

AR: This is an important point, which we aimed to illustrate in Figure 6. There, we show that when MERRA-2 fields are used as input and all species predicted by MAMnet are combined (as in Figure 1), the resulting bias is very small. While we originally presented this as evidence that MAMnet does not inherit the biases of GEOS+MAM7, it also demonstrates that the model conserves mass, as it accurately maintains the total mass of each species. Additional support for this comes from the the modal geometric mean diameter, $D_{pg,i}$, as it remains very close to MAM7. Since $D_{pg,i}$ is not predicted by MAMnet, but instead calculated from its output, it indicates that both mass and number concentrations evolve in a physically consistent manner. In an online implementation MAMnet would be diagnostic to the bulk model, hence would not directly influence mass conservation.

We have moved Figure 6 and the corresponding discussion before Figures 2 and 3 at the start of the new Section 3.1 to clarify this point.

RC: ***Section 3.1: Although I agree with the motivational 1st paragraph of this section (lines 249-252), it sounds more than wishful thinking. MAMnet is trained with model data, not measurements, so at its peak performance it will be able to emulate the modeled data. In terms of measurements, it can only be as good as GEOS+MAM or MERRA-2 models, and any improvement in skill when compared with measurements (if at all evident) will be coincidental, thus irrelevant. What is really missing from both sections 3.1.1 and 3.1.2 is a baseline discussion: how does MERRA-2 alone perform when comparing with measurements? Of course MERRA-2 does not simulate ASD, but biases in the total aerosol mass (per species or not) will impact ASD. Even more, GEOS+MAM does not include assimilation, so other sorts of biases are likely present in the ASD of the training data set. Since this paper is about MAMnet, and since section 3.1 as a whole is to demonstrate its overall skill, not knowing the skill of the training dataset is a major shortcoming. To the very least, GEOS+MAM should be presented in figures 7 and 8, but a mass concentration comparison (or citation of past evaluation efforts) should be presented as well.***

AR: Thank you for this comment.

To clarify, MERRA-2 is an observation-constrained dataset created by assimilating a wide range of measurements, including satellite-derived aerosol optical depth, into the GEOS model. Because of this assimilation, MERRA-2 aerosol fields are in principle closer to observations than those from GEOS+MAM7. Several studies have evaluated the performance of MERRA-2 aerosol fields against observations [e.g., Buchard et al., 2017, Sun et al., 2019, Ukhov et al., 2020, Gueymard and Yang, 2020, Su et al., 2023], and we now reference these in the revised manuscript.

We agree that biases in the training data (GEOS+MAM7) can influence the learned mapping, and that the skill of the input data (MERRA-2) affects the final output. However, we do not present comparisons between GEOS+MAM7 and observations at specific sites, since such comparisons are not meaningful for a free-running model. Instead, we emphasize that comparisons between MERRA-2+MAMnet and observations reflect the model's performance when driven by realistic aerosol fields.

We use GEOS+MAM7 as the training dataset because it provides internally consistent mass and number

concentrations needed to learn the relationship between them. MAMnet is trained to approximate this relationship, not to replicate the exact output of GEOS+MAM7. When applied to MERRA-2, MAMnet combines the more realistic aerosol mass fields from MERRA-2 with the learned mass–number mapping. This allows us to assess its performance in a more observationally constrained setting.

We have added the following clarification before Section 3.1 to the manuscript to address this point:

"MERRA-2 includes aerosol mass fields that are constrained by satellite observations through data assimilation [Buchard et al., 2017, Sun et al., 2019, Ukhov et al., 2020, Gueymard and Yang, 2020, Su et al., 2023], and thus provides a more realistic input compared to free-running model simulations. Although GEOS+MAM7, which was used to train MAMnet, does not assimilate aerosols and cannot be directly compared to observations at specific sites, it provides physically consistent mass and number concentrations from which the network learns the relationship between these quantities. When applied to MERRA-2, MAMnet combines this learned relationship with more observation-constrained aerosol mass fields, allowing us to evaluate how well it maintains physical consistency in a more realistic setting. This comparison does not validate MAMnet independently of its training data but serves to assess its performance when driven by the best available mass estimates."

RC: ***Section 3.2: please explain what Shapley values are exactly. There is some information in the figure legend, but a short introduction would be useful. Also, since this is a comparison against the model data, I would recommend moving it before the observations sections, so swapping sections 3.1 and 3.2.***

AR: The following explanation has been added to the section, which has been moved before the observations section.

"Shapley values [Winter, 2002], originally developed in cooperative game theory, are now widely used to interpret predictions from neural networks [Kwon et al., 2023, Jeggle et al., 2023, Jia et al., 2023, Ma and Stinis, 2020, Lundberg and Lee, 2017]. A Shapley value quantifies the contribution of a single input feature to a specific model prediction by comparing the prediction for a given sample to the average prediction across all samples. This contribution is averaged over all possible combinations of the remaining input features, referred to as coalitions. Because the number of such combinations grows rapidly with the number of features, we approximate Shapley values using 1,000 randomly selected coalitions per calculation, facilitated by the SHAP python library using the kernel explainer method [Lundberg et al., 2020]. In this study, Shapley values are used to assess the influence of each input feature on the predicted aerosol number concentrations for each mode."

RC: ***Line 334: What do you mean by "possibly by promoting secondary aerosol formation" here? Secondary organics will evaporate more at higher temperatures, while secondary inorganic aerosols will have a more complex relationship depending on relative humidity as well.***

AR: Thanks for pointing it out. We have removed the statement as it is speculative.

3. Technical Corrections

RC:

1. ***Line 44: Change "ML models, we can" to "ML models can".***
2. ***Line 79: Add "of different sizes" after "five mass bins".***

3. ***Line 80: Replace "hydrophilics" with "hydrophilic".***
4. ***Line 86: Table 2 is referenced before Table 1.***
5. ***Line 97: Replace "these" with "that".***
6. ***Figure 1:*** rho_{air} ***is mentioned in the legend, but it is termed AIRD in the figure.***
7. ***Line 109: Replace "Kg" with "kg".***
8. ***Lines 179 and 181: "the original MAM" and "GEOS+MAM" are the same thing, right? Please use one terminology throughout, for clarity.***
9. ***Line 214: "smaller and less massive" is the same, why not just say "smaller"?***
10. ***Line 217: Replace "near-perfect" with "very high".***
11. ***Line 223: Replace "sulfates" with "sulfate".***
12. ***Line 260: Replace "accurate" with "accurately"***
13. ***Line 311: Replace "tends align" with "tends to align".***
14. ***Figure 9: Please add a figure legend that explains the color lines, on top of the verbal description present in the caption.***
15. ***Line 363: Replace "predicted concentrations" with "predicted number concentrations".***

AR: All technical corrections have been incorporated.

References

ADLA Jones, DL Roberts, and A Slingo. A climate model study of indirect radiative forcing by anthropogenic sulphate aerosols. *Nature*, 370(6489):450–453, 1994.

J Langner and H Rodhe. A global three-dimensional model of the tropospheric sulfur cycle. *Journal of Atmospheric Chemistry*, 13:225–263, 1991.

Paul Ginoux, Mian Chin, Ina Tegen, Joseph M Prospero, Brent Holben, Oleg Dubovik, and Shian-Jiann Lin. Sources and distributions of dust aerosols simulated with the gocart model. *Journal of Geophysical Research: Atmospheres*, 106(D17):20255–20273, 2001.

Mian Chin, Richard B Rood, Shian-Jiann Lin, Jean-Francois Müller, and Anne M Thompson. Atmospheric sulfur cycle simulated in the global model gocart: Model description and global properties. *Journal of Geophysical Research: Atmospheres*, 105(D20):24671–24687, 2000.

N Riemer, AP Ault, M West, RL Craig, and JH Curtis. Aerosol mixing state: Measurements, modeling, and impacts. *Reviews of Geophysics*, 57(2):187–249, 2019. .

Huang Zhang, Girish Sharma, Sukrant Dhawan, David Dhanraj, Zhichao Li, and Pratim Biswas. Comparison of discrete, discrete-sectional, modal and moment models for aerosol dynamics simulations. *Aerosol Science and Technology*, 54(7):739–760, 2020. .

Peter J Adams and John H Seinfeld. Predicting global aerosol size distributions in general circulation models. *Journal of Geophysical Research: Atmospheres*, 107(D19):AAC–4, 2002.

V Buchard, CA Randles, AM Da Silva, A Darmenov, PR Colarco, R Govindaraju, R Ferrare, J Hair, AJ Beyersdorf, LD Ziemba, et al. The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies. *Journal of Climate*, 30(17):6851–6872, 2017.

Enwei Sun, Huizheng Che, Xiaofeng Xu, Zhenzhu Wang, Chunsong Lu, Ke Gui, Hujia Zhao, Yu Zheng, Yaqiang Wang, Hong Wang, et al. Variation in merra-2 aerosol optical depth over the yangtze river delta from 1980 to 2016. *Theoretical and Applied Climatology*, 136(1):363–375, 2019. .

Alexander Ukhov, Suleiman Mostamandi, Arlindo Da Silva, Johannes Flemming, Yasser Alshehri, Illia Shevchenko, and Georgiy Stenchikov. Assessment of natural and anthropogenic aerosol air pollution in the middle east using merra-2, cams data assimilation products, and high-resolution wrf-chem model simulations. *Atmospheric Chemistry and Physics Discussions*, 2020:1–42, 2020. .

Christian A Gueymard and Dazhi Yang. Worldwide validation of cams and merra-2 reanalysis aerosol optical depth products using 15 years of aeronet observations. *Atmospheric Environment*, 225:117216, 2020. .

Xin Su, Yuhang Huang, Lunche Wang, Mengdan Cao, and Lan Feng. Validation and diurnal variation evaluation of merra-2 multiple aerosol properties on a global scale. *Atmospheric Environment*, 311:120019, 2023. .

Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

Youngchae Kwon, Seung A An, Hyo-Jong Song, and Kwangjae Sung. Particulate matter prediction and shapley value interpretation in korea through a deep learning model. *SOLA*, 19:225–231, 2023.

Kai Jeggle, David Neubauer, Gustau Camps-Valls, and Ulrike Lohmann. Understanding cirrus clouds using explainable machine learning. *Environmental Data Science*, 2:e19, 2023.

Yichen Jia, Hendrik Andersen, and Jan Cermak. Analysis of cloud fraction adjustment to aerosols and its dependence on meteorological controls using explainable machine learning. *EGUsphere*, 2023:1–25, 2023.

Po Lun Ma and Panagiotis Stinis. Developing a simulator-based satellite dataset for using machine learning techniques to derive aerosol-cloud-precipitation interactions in models and observations in a consistent framework. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020. .