ORNL/TM-2019/1109

Evaluation of Machine Learning Approaches to Estimate Aerosol Mixing State Metrics in Atmospheric Models

Zhonghua Zheng
Nicole Riemer
Matthew West
Valentine G. Anantharaj

May 2019

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM--2019/1109

National Center for Computational Sciences

EVALUATION OF MACHINE LEARNING APPROACHES TO ESTIMATE AEROSOL MIXING STATE METRICS IN ATMOSPHERIC MODELS

Zhonghua Zheng
Nicole Riemer
Matthew West
Valentine G. Anantharaj

Date Published: May 2019

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

This research was supported in part by an appointment to the Oak Ridge National Laboratory ASTRO Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

ABSTRACT

Aerosol mixing state describes how aerosol compositions are distributed among atmospheric aerosol particles in a population. Oversimplified assumptions of aerosol mixing state in atmospheric modeling can introduce errors in estimations of weather and climate-relevant aerosol microphysical properties. A more comprehensive representation of the aerosol mixing state can be achieved in principle with a Particle-resolved Monte Carlo (PartMC) model but at added computational cost that may be prohibitive for direct invocation in operational numerical weather prediction or multi-year climate simulations.

The aim of our research is to explore the machine learning (ML) methodologies for estimating aerosol mixing state metrics, which we define here in three different ways: with respect to (a) hygroscopicity; (b) optical properties; and (c) chemical species abundance. We adopted a data-driven approach, leveraging deep learning and statistical learning techniques, to take advantage of massive PartMC model simulations. First, we performed particle-resolved simulations by PartMC to create a series of scenarios considering a range of global environmental conditions. Each scenario consists of aerosol populations and corresponding mixing state metrics. The gas concentration, aerosol mass concentration, environmental variables and mixing state metrics of each population constitute the datasets for machine learning implementations.

We have adopted and evaluated various configurations of machine learning methodologies in this investigation, embracing deep learning, Extreme Gradient Boosting (XGBoost) algorithm, and the ensemble approaches. After a rigorous model selection process, we identified an appropriate model to derive estimates of aerosol mixing state metrics. We used the computational and data resources of the Oak Ridge Leadership Computing Facility (OLCF) and the ORNL Compute and Data Environment for Science (CADES). The NVIDIA DGX-1 hardware was used for the prototyping of the ML models.

Our approach has allowed us to gain a new understanding of how machine learning methodologies can be applied to improve the representation of aerosol mixing state in atmospheric models and benefit the atmospheric research community. Next, we plan to extend our research and methodology to quantify some of the aerosol-related uncertainties in the E3SM Atmospheric Model (EAM) Version 1.

1. INTRODUCTION

1.1 ATMOSPHERIC AEROSOLS

The atmospheric aerosol is a "dispersed condensed phase suspended in a gas," mainly originating from the condensation of gases and from the action of the wind on the Earth's surface (Jacob, 1999). It impacts the lives of human beings given that every day billions of aerosol particles are inhaled with the ambient air by every human being (Heyder, 2004). Aerosol particles can be natural (e.g., fog, dust, forest exudates, and geyser steam, etc.) or anthropogenic (e.g., haze, particulate air pollutants, and smoke, etc.).

Aerosols are known to pose strong adverse impacts on human health (Pope III et al., 2002), and also impact weather (Chin et al., 2007) as well as climate (Ghan et al., 2012). Studies over the past decades implicated that aerosol particle is associated with health disorders including cardiovascular, neurological, and respiratory diseases (Wang et al., 2018), especially aerosol particulate matter below 2.5μm ($PM_{2.5}$) due to its ability to penetrate deeper and deposit in the lower respiratory tract (Martins et al., 2015). Aerosols play vital roles in cloud formation via cloud condensation nuclei (CCN) properties that act as the initial sites for condensation of water vapor into cloud droplets or cloud ice particles. The concentration of CCN in an air parcel influences cloud microphysical and radiative properties, thus leading to the direct and indirect effects on climate systems. For example, a higher concentration of CCN at a given supersaturation results in more droplets with a smaller mean droplet diameter, thus a more reflective cloud, known as Twomey effect (Twomey, 1977). Another indirect effect comes from the inhibition of precipitation in clouds with small mean droplet diameters which alters the extent and lifetime of clouds (Albrecht, 1989).

1.2 AEROSOL MIXING STATE

Atmospheric aerosols vary in their chemical compositions. Individual aerosol particles in the atmosphere are often complex mixtures of a wide variety of chemical species (Bein et al., 2005; Noble & Prather, 2000). We use the term mixing state in this context to describe how the aerosol chemical species are distributed among the aerosol particles in a population (Riemer & West, 2013). Figure 1 shows an example of an aerosol population composed of two chemical species, with four different mixing states. A completely externally mixed population contains only one species per particle, while a completely internally mixed population contains identical particles which are a mixture of two chemical species. An infinite number of intermediate mixing states lie between those two extremes. To quantify the degree of mixing state Riemer and West (2013) introduced the concept of a mixing state index (metric) χ where a

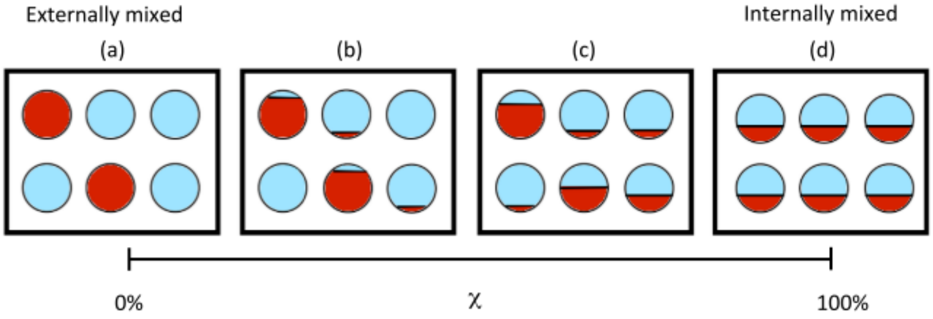

Figure 1. Schematic of aerosol mixing states for four different aerosol populations that have the same bulk composition (Hughes et al., 2018). The blue and red color represent aerosol species: (a) fully external mixture; (b, c) intermediate mixing states; and (d) internal mixture. The mixing state metric χ measures the degree of internal mixing, ranging from 0% to 100%.

scalar 0% stands for externally mixed while 100% for internally mixed. Multiple aerosol mixing state metrics can be defined with respect to hygroscopicity, optical properties, and chemical species abundance.

The magnitude of the aerosol impact on climate significantly relies on the mixing state of aerosols (China et al., 2013; Schill et al., 2015). An improper assumption of the aerosol mixing state representation in atmospheric modeling may introduce errors in estimations of weather and climate-relevant aerosol properties. Jacobson (2000) suggested that mixing state of black carbon in atmospheric aerosols impacts radiative heating. Moreover, the properties of fresh emitted atmospheric aerosols can evolve along with time in an ambient environment, resulting in significantly different effects on radiative and microphysical properties. One of the examples can be shown by the shift from an externally mixed state characteristic of fresh emissions to the internally mixed state leading to significant modifications to the optical properties (Doran et al., 2007). Ching et al. (2017) quantified the error in CCN activity due to simplifying assumptions about mixing state, i.e. assuming that the particle population is internally mixed. Their analysis (Figure 2) reveals that for more externally mixed populations (χ below 20%) the relationship between χ and the error in CCN predictions is not similar but ranges from around -40% to 150%, depending on the underlying aerosol population and the environmental supersaturation.

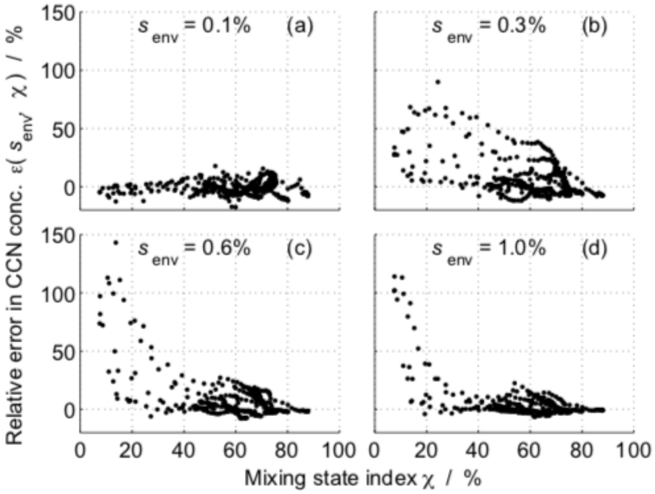

Figure 2. Relative error in CNN concentration for different environmental supersaturations (Ching et al., 2017).

1.3 REPRESENTATION OF AEROSOLS IN ATMOSPHERIC MODELS

A realistic representation of aerosol properties in atmospheric models is required for a proper understanding and interpretation of aerosol effects on climate via clouds, radiation, and precipitation. A comprehensive representation of the aerosol mixing state can be achieved in principle with a Particle-resolved Monte Carlo (PartMC) model (Riemer et al., 2009) which would be free of any assumptions about the mixing state of aerosols. However, the extremely expensive computational cost of this approach is prohibitive for numerical weather and climate modeling. In contrast, aerosol mixing state representation in global climate models has been highly simplified. The first-generation climate models assumed that the aerosols were externally mixed such that each particle is composed of only one type of species (Ghan et al., 2012). The new generation of climate models, on the other hand, assume internal mixtures for the particle of the same size, commonly known as modal and sectional models (Seigneur et al., 1986; Wexler et al, 1994; Zhang et al., 1999). For instance, modal models represent all aerosols within a given mode as internally mixed.

If the real mixing state is closer to an external mixture, it will result in errors in aerosol properties, as discussed above. It is desirable to derive global maps of uncertainties in aerosol mixing state where regions of low χ (externally mixed) can be expected. These would be the areas where we could anticipate large errors in CCN prediction due to a simplified aerosol model (assuming internally mixed). On the contrary, the regions of the globe with high χ (internally mixed) increase the trust in current CCN estimations. Nowadays, the advance of computational resources facilitates the large-scale climate simulation; but we are still many decades away from running a particle-resolved aerosol model directly on a global scale to create a global map of χ.

1.4 ESTIMATING AEROSOL MIXING STATE VIA MACHINE LEARNING

In the era of big data, the recent availability of large volume of datasets and the developments in computer science enable machine learning to resolve the representation of aerosol mixing state metrics in atmospheric models at a global scale. The recent investigation conducted by Hughes et al. (2018) produces the first global distribution of a single category of mixing state metric in terms of hygroscopicity. Their approach is a combination of particle-resolved modeling and global chemical transport model outputs empowered by gradient-boosted regression trees, which provides the capability to estimate the mixing state metric. However, the omission of certain critical feature (predictor) variables (e.g., temperature) of their methodology may inhibit a reliable prediction. On the other hand, features that are either redundant or irrelevant might prohibit the development of a better model because of the redundancy and increased computing resources introduced by those features. In our approach, we are also interested in estimating mixing states with respect to optical properties and chemical species abundance, providing new perspectives and means to study the aerosol properties.

1.5 OBJECTIVES

This project extends our former study and seeks to apply both Deep Learning and *Extreme Gradient Boosting* algorithm (XGBoost) (Chen & Guestrin, 2016) in order to develop a data-driven avenue to predict the multiple mixing state metrics. Deep learning enables computational models to discover intricate structure in large data sets (LeCun et al., 2015). XGBoost is chosen due to its merit in employing a computationally efficient variant of gradient tree boosting methods and fabulous recognition in ML competitions, other studies, and domains (Torlay et al., 2017). During this current phase of the project we have investigated the following:

1) Examine XGBoost, deep learning, and ensemble approaches to representing a single aerosol mixing state metric based on hygroscopicity.
2) Evaluate and develop approaches to contemporaneously represent the multiple aerosol mixing state metrics with respect to hygroscopicity, optical properties, and chemical species abundance.

2. APPROACH

The overarching goals of this project center on two interconnected phases (Figure 3): 1) leverage machine learning to develop a predictive model for mixing state estimation; and 2) facilitate the scientific discovery based on global aerosol mixing state distribution using the machine learning model to characterize the uncertainties in the E3SM Atmosphere Model (EAM). This report describes the outcomes from the first phases of this study (Figure 4).

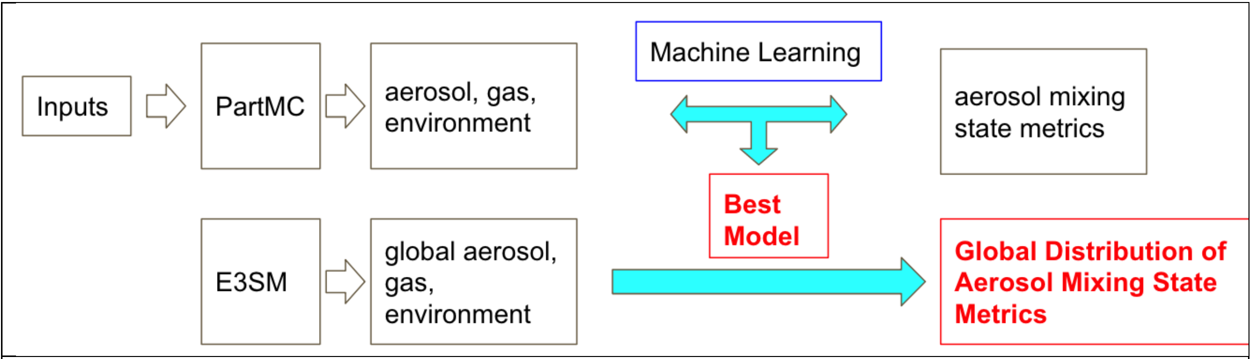
Inputs
PartMC
aerosol, gas, environment
Machine Learning
Best Model
aerosol mixing state metrics
E3SM
global aerosol, gas, environment
Global Distribution of Aerosol Mixing State Metrics

Figure 3. Overarching framework of the entire project.

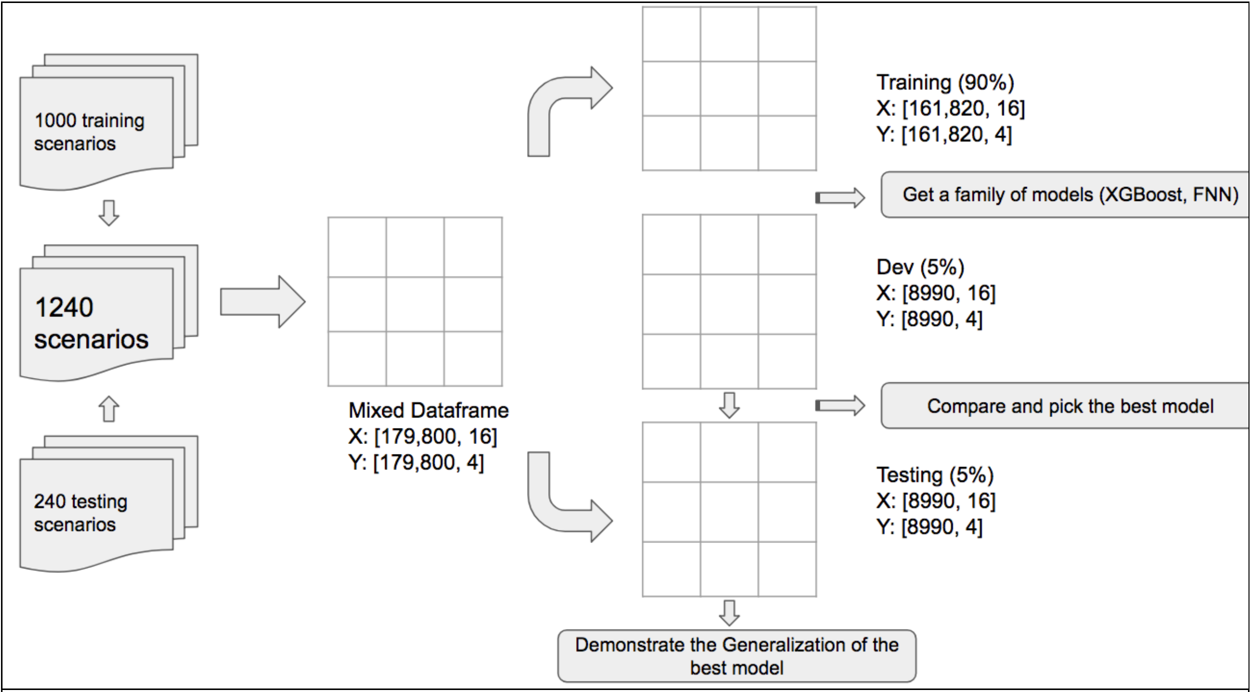

Figure 4. First phase overview.

2.1 CHARACTERIZATION OF MULTIPLE AEROSOL MIXING STATE METRICS

We have quantified aerosol mixing state with the framework designed by Riemer and West (2013), specifically using the mixing state metric χ, which was motivated by diversity metrics used in other disciplines such as ecology (Whittaker, 1972), economics (Drucker, 2013) and genetics (Falush et al., 2007). This metric is an affine ratio of average per-particle species diversity D_α and the bulk population species diversity D_γ, where both are based on information-theoretic entropy measures. The definition of "species" is not only limited to individual chemical species but also can refer to species groups. For example, Dickau et al. (2016) quantified mixing state with respect to volatile and non-volatile components; Hughes et al. (2018) assessed mixing state according to hygroscopicity, defining two species groups.

Here we define multiple aerosol mixing state metrics in three distinct ways: with respect to chemical species abundance (*chi*), hygroscopicity (*chi_hyg*), and optical properties (*chi_opt1* and *chi_opt_2*). In order to calculate the mixing state metric based on hygroscopicity, we have combined black carbon, primary organic matter, soil dust, freshly emitted marine organic matter into one surrogate species, since their hygroscopicity is very low. All other model species are combined into a second surrogate species. Two scenarios were assumed to derive the mixing state metric based on optical properties. The first scenario considers black carbon as absorbing while the rest species are non-absorbing. For the other scenario, both black carbon and soil dust are grouped as one surrogate absorbing species. A detailed description of the group is shown in Table 1.

Table 1. Definition of multiple mixing state metrics

Mixing state metrics	**Justification**	**Group**
chi	Mixing state	/
chi_hyg	Hygroscopicity	black carbon, primary organic matter, soil dust, and marine organic matter
chi_opt1	Optical property	black carbon
chi_opt2	Optical property	black carbon, and soil dust

2.2 DATA

2.2.1 Creation and assembly of data

Aerosol mixing state of an aerosol population evolves by the so-called aerosol aging process including condensation of atmospheric gaseous components and coagulation with other aerosols. We performed particle-resolved simulations by PartMC to create a series of scenarios representing a range of global environmental conditions. A scenario describes the change in aerosol mixing state along with time upon the certain environmental scenario. The gaseous, aerosol properties, environmental variables, and corresponding aerosol mixing state metrics are tracked and recorded by PartMC at each timestamp. Given the mixing state metrics only depend on variables within the current box (grid) due to the nature of PartMC, we make the assumption that every point in time on the same grid can be considered a separate sample for training and testing purposes, leading to a collection of aerosol populations (and corresponding mixing state metrics) within every single scenario. The number of the population samples within each scenario equals to the length of the timestamp of each simulation.

The same set of scenarios created by Hughes et al. (2018) as were adopted in our research. The data consists of 144,000 samples for training and 34,560 samples for testing purposes. A subtle difference between the datasets was in the initial state of the particle populations. We included the initial (timestamp) state samples that were excluded in Hughes et al. (2018) to carry more information. Thus this yields additional 1,000 particle populations for original training, and 240 particle populations for original testing scenarios. Another distinction between the scenarios is that original training and testing scenarios are mixed and shuffled in our methodology. The entire samples (179,800 particle populations) were re-distributed into training, development (dev), and testing sets according to the proportion 90% (training) /

5% (dev) / 5% (testing). Although the percentage of testing samples account for 5% of the overall data sets, the 8,990 samples provide sufficient cases, corresponding to 8,990 particle populations. We applied the various learning algorithms on training set, and conducted the hyperparameter tuning on dev set. The generalization performance of the algorithm was evaluated on the test set.

A subset of the atmospheric variables (also referred to as features), common to both EAM and PartMC, were selected as inputs in model construction. In this way, some features that are either redundant or irrelevant were removed while the entire datasets keep sufficient information (Bermingham et al., 2015). The common input features consist of gas concentrations, aerosol mass concentration, and environmental variables. More information of the input features is available in Support Information Table 1.

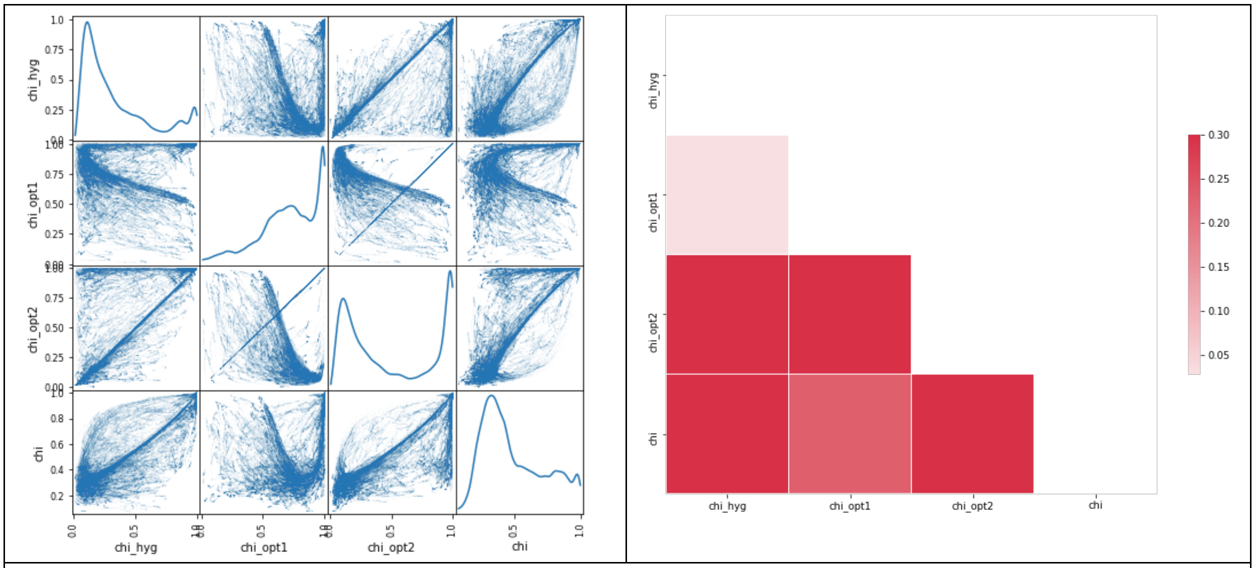

Figure 5. (a) Scatter plots and probability density functions (a); and (right) correlation coefficient of multi-aerosol mixing state metrics (b) in training set.

2.2.2 Characteristics of data

The characteristics of multi-aerosol mixing state metrics in the training set is explored in Figure 5(a) and Figure 5(b). Since the dev and testing sets are from the same distribution, their characteristics are in line with the training. All the mixing state metrics are strictly confined within values of 0 to 1 due to the nature of their definitions. Certain linear correlations are identified between a pair of different mixing state metrics. Unexpectedly, there is a 1:1 line between chi_opt1 (optical property only considering black carbon) and chi_opt2 (optical property considering black carbon and soil dust), which might be caused by the lack of the soil dust in the population, resulting in the black carbon dominating the optical property. On the other hand, chi_opt1 is not correlated to chi and chi_hyg. The probability density functions for the 4 mixing state metrics are clearly distinguished from one another. For example, two modes are identified in chi_op2 which may owe to the different distribution of the black carbon and soil dust compared to the chi_opt1.

2.3 METHODOLOGY

2.3.1 Implementation

2.3.1.1 Deep learning

For deep learning, the activation function of nodes is critical in the neural network which offers the nonlinearity relationship and bounds the results in a range. In this investigation, both sigmoid (Cybenko, 1989) and Rectified Linear Units, ReLUs (Krizhevsky et al., 2012) functions were adopted as the activation functions. The sigmoid function is a monotonic, bounded, and differentiable function that has the non-negative output (ranging from 0 to 1) and derivative at each point. Krizhevsky et al. (2012) suggest that deep convolutional neural networks with ReLUs outperform the other activation function (tanh) by several times faster training speed. However, the ReLUs function cannot ensure the results are within the boundary (given the mixing state metric ranging from 0 to 1). Additional steps need to be taken to limit the predictions within the continuity interval [0, 1]. This is quite often accomplished by simple truncation. We consider two hidden layers in this study, and each layer encompasses 32 neurons (Figure 6).

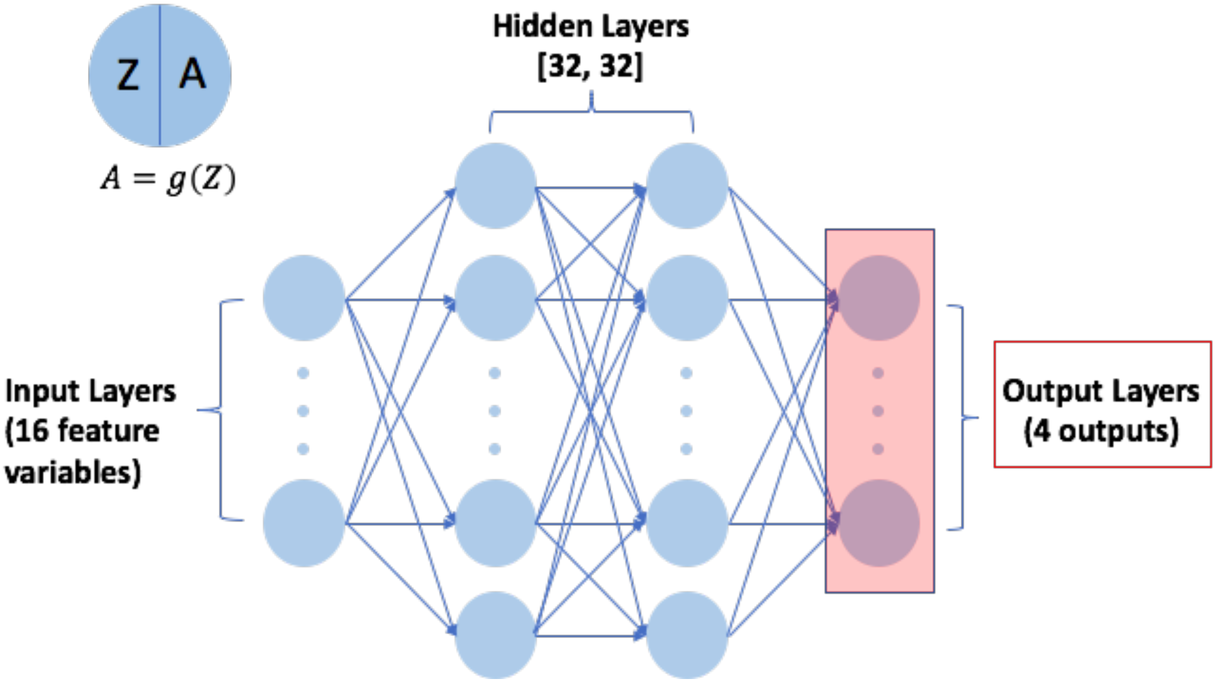

Figure 6. The neural network architecture of this project.

2.3.1.2 XGBoost

Similarly, the outputs of XGBoost overflow the boundary occasionally, same steps were taken to secure the results within the boundary. Various hyperparameters were evaluated in this investigation, here we only present the best combination and the corresponding results for each machine learning algorithm. Detailed descriptions of algorithms were presented in Table 2 and Table 3. What should be noted is that the design of Table 3 is slightly different from the Table 2. For examples, the predictions by $\mathrm{FNN}_{\mathrm{RELU}}$ in Table 2 are within the boundary, while the others mixing state metrics predicted by $\mathrm{FNN}_{\mathrm{RELU}}$ in Table 3 overflow the boundary. For this reason, FNN$_{\text{RELU\_f}}$ was considered in Table 3. The following metrics were examined: RMSE (Root Mean Square Error), R^2 (Coefficient of Determination), and IOA (Index of Agreement). A comprehensive comparison of different implementations will be interpreted in Section 3.

Table 2. Description of Algorithms for Single Aerosol Mixing State Metric with respect to Hygroscopicity

Approaches	**Ensembling or not**	**Forced the boundary of XGBoost (as inputs) or not**	**Forced the boundary of ensemble results or not**	**Overflow the boundary or not**
XGBoost (XGB, benchmark)	N	N	N/A	Y (from former studies)
Fully-Connected Neural Network with ReLUs function (FNN$_{\text{RELU}}$)	N	N	N/A	Unknown
XGB with confined boundary (XGB$_{\text{f}}$)	N	Y	N/A	N
Ensemble Approach (EA, a linear combination of XGB and FNN$_{\text{RELU}}$)	Y	N	N	Unknown
EA with confined boundary (EA$_{\text{f}}$)	Y	N	Y	N
Confined EA$_{\text{f}}$ (EA$_{\text{ff}}$, a linear combination of XGB$_{\text{f}}$ and FNN$_{\text{RELU}}$ with confined boundary)	Y	Y	Y	N

Table 3. Description of Algorithms for Multi-Aerosol Mixing State Metrics

Approaches	**Ensembling or not**	**Forced the boundary of XGBoost (as inputs) or not**	**Forced the boundary of ensemble results or not**	**Overflow the boundary or not**
XGBoost (XGB, benchmark)	N	N	N/A	Y (from former studies)
XGB$_{\text{f}}$	N	Y	N/A	N
Fully-Connected Neural Network with sigmoid function (FNN$_{\text{SIG}}$)	N	N	N/A	N
FNN$_{\text{RELU}}$	N	N	N/A	Unknown
FNN$_{\text{RELU}}$ with confined boundary (FNN$_{\text{RELU\_f}}$)	N	Y	N/A	N
Ensemble Approach	Y	Y	Y	N

(EA*, a linear combination of XGB_f and FNN_{RELU_f} with confined boundary)				

2.3.2 Hyperparameter tuning, cross-validation and model selection for XGBoost

Model selection is the indispensable element of the entire workflow. Hyperparameters stand for the parameters whose value is set before the learning process begins rather than during the course of the machine learning process. In this project, the hyperparameter tuning follows the idea of grid-searching technique that scans the performance of each combination to configure the optimal combination of hyperparameters for a given model. Although cross-validation is widely adopted for machine learning, it is not worth the trouble for the deep learning (Bengio, 2016). Instead, we use the train/dev/test split to tune the hyperparameters for deep learning. Same strategies were taken for XGB to keep consistent with deep learning. With respect to deep learning, the learning rate, batch size, and the number of training epochs were considered as the hyperparameters. Whilst we explored the maximum depth of a tree (max_depth), step size shrinkage used in the update to prevents overfitting (learning_rate), and the number of boosted trees to fit (n_estimators) for XGB. Both training set and dev set were utilized to select the models. We derived models with a various combination of hyperparameters from the training set and applied them into the dev set. The model with the lowest RMSE or highest R^2 was determined as the optimal model for both deep learning and XGB.

3. RESULTS

3.1 ESTIMATION OF SINGLE AEROSOL MIXING STATE METRIC

To assess the ability of original algorithms without confining the results to the boundary we start by comparing the XGB, FNN_{RELU}, and EA. The predictions by XGB [-0.0007, 1.01578]) (in dev set and EA [-0.00285, 1.01883] (in dev set) overflow the boundary whereas the estimations by FNN_{RELU} [0, 0.938501] (in dev set) surprisingly within the boundary. The results suggest that neither XGB and EA could not be the best model for single aerosol mixing state prediction, although the RMSE of XGB (0.01716) and EA (0.01712) are low. The RMSE, R^2, and IOA of FNN_{RLUE} are 0.06622, 0.95032, and 0.98710, respectively, indicating that FNN_{RELU} is promising to predict the single aerosol mixing state metric with respect to hygroscopicity.

The XGB_f, EA_f, EA_{ff} are evaluated against the dev sets as well. The XGB_f forces the predictions of XGB within the boundary, which only marginally improves the RMSE from 0.01716 to 0.01715 and slightly boosts the R^2 from 0.996663 to 0.996667. EA_f leverages the results from XGB and FNN_{RELU} to group as a linear combination, while EA_{ff} linearly combines the results from XGB_f and FNN_{RELU}. All the predictions are forced to within the boundary. The evaluation based on RMSE, R^2 and IOA offer the conclusions, in this order (from most accurate): EA_{ff} > EA_f > EA (outflow) > XGB_f > XGB (outflow) > FNN_{RELU} (Table 4). The EA_{ff} offers the best performance with the lowest RMSE (0.01709 in dev set, and 0.01726 in testing dev), R^2 (0.99669 in dev set, and 0.99653 in testing set), and IOA (0.99917 in dev set, and 0.99913 in testing dev). In general, the ensemble approaches outperform the non-ensemble approach, and the XGB show an advantage over the neural network.

Table 4. Evaluation of Single Aerosol Mixing State Metric with respect to Hygroscopicity (dev set)

Approaches	**RMSE**	**R^2**	**IOA**	**Overflow the boundary or not**
XGBoost (XGB, benchmark)	0.017163	0.996663	0.999161	Y
Fully-Connected Neural Network with ReLUs function (FNN$_{RELU}$)	0.066223	0.950320	0.987102	N (but not applicable to other mixing state metrics)
XGB with confined boundary (XGB$_f$)	0.017152	0.996667	0.999162	N
Ensemble Approach (EA, a linear combination of XGB and FNN$_{RELU}$)	0.017118	0.996681	0.999169	Y
EA with confined boundary (EA$_f$)	0.0170921	0.9966906	0.9991707	N
Confined EA$_f$ (EA$_{ff}$, a linear combination of XGB$_f$ and FNN$_{RELU}$ with confined boundary)	0.0170917	0.9966907	0.9991709	N

3.2 MULTI-AEROSOL MIXING STATE METRICS PREDICTIONS

The model selections in single aerosol mixing state metric prediction contribute a guideline, which facilitates the model selections toward multi-aerosol mixing state metrics predictions. Multi-Target Regression (MTR) is considered when it comes to multiple dependent variables. Theoretically, the same number of XGB models can be trained with respect to the same number of mixing state metrics. For instance, four XGB models are easy to attain from training the same features/inputs but different outputs in our case. However, when we define more than four aerosol mixing state metrics, models using single outputs take longer and are more computationally expensive. In addition, models using single outputs with the same features omit the potential relationship between target outputs, since certain target outputs may share the similar features during learning. The deep learning demonstrates its superiority over multi-target predictions since the shallow neural layers are shared by all the outputs, which is simpler than a stack of single output models. Here we investigate the performance of 1) multiple XGB models, 2) different options for a single neural network with multi-outputs, and 3) the ensemble approach leveraging the above models.

3.2.1 Non-ensemble approaches

As we discussed previously, The relationship among different mixing state metrics shows the potential of applying MTR. The non-ensemble approaches, namely, XGB, XGB$_f$, FNN$_{SIG}$, FNN$_{RELU}$, and FNN$_{RELU\_f}$ were evaluated against the dev sets. The following order is arranged in rank order according to the performance (from best): XGB$_f$ > XGB > FNN$_{RELU\_f}$ > FNN$_{RELU}$ > FNN$_{SIG}$. The XGB$_f$ stands out to be the best model for each mixing state metrics in highest R^2 for chi (0.99710), chi_hyg (0.99667), chi_opt1

(0.99825), chi_opt2 (0.99894), respectively. With respect to the models related to deep learning in our case, FNN_{RELU_f} performs the best with highest R^2 for chi (0.94859), chi_hyg (0.93221), chi_opt1 (0.95452), chi_opt2 (0.94324). Here we emphasize again that FNN_{RELU_f} is a single neural network which predicts the multi outputs simultaneously, while the XGB_f are multiple models for multiple mixing state metrics correspondingly. For example, four XGB_f models are trained then leveraged to predict the mixing state metrics. Amongst them, the XGB and FNN_{RELU} could not be chosen the final models since both of them overflow the boundary, albeit XGB provides high predictive performance. Here we adopted the XGB_f (hereinafter XGB), and FNN_{RELU_f} (hereinafter ANN) as the feasible model for predicting the multi-aerosol mixing state metrics given there is a tradeoff between runtime and accuracy. Figure 7 displays the predictions of the dev set against the actual values using ANN and XGB.

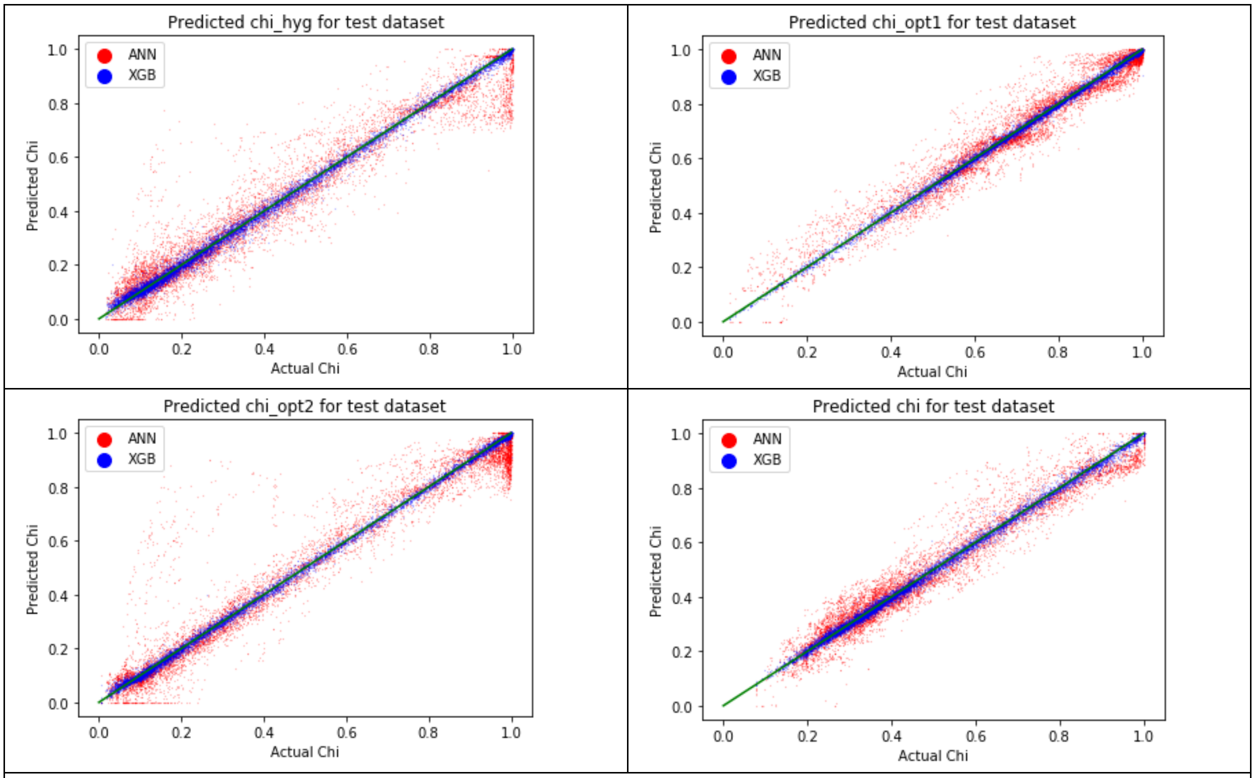

Figure 7. Predictions of multi-aerosol mixing state metrics in the testing set using XGB (XGB_f in the context,) and ANN (FNN_{RELU_f} in the context).

3.2.2 Ensemble approaches

As we discussed earlier, the ensemble approach considers the linear combination of outputs predicted by different models, leading to better estimations for single mixing state metric. Here we explored the same technique for multi-aerosol mixing state metrics. The output from XGB_f and FNN_{RELU_f} are treated as features to formalize as a linear regression problem. Assume we have n samples, the ensemble approach can be expressed as

$$y_{ensemble} = \beta_1 \, y_{XGB} + \beta_2 \, y_{ANN} + \beta_0$$

where y_{XGB} and y_{ANN} are calculated by the XGB_f and FNN_{RELU_f} models. The ensemble approach allows for a nudging for the prediction by individual model, ensuring stable predictions. Figure 5 reveals the predictions by combining the XGB_f and FNN_{RELU_f}, with better RMSE, R^2, and IOA in all the mixing state metrics than the best non-ensemble approach above (XGB_f).

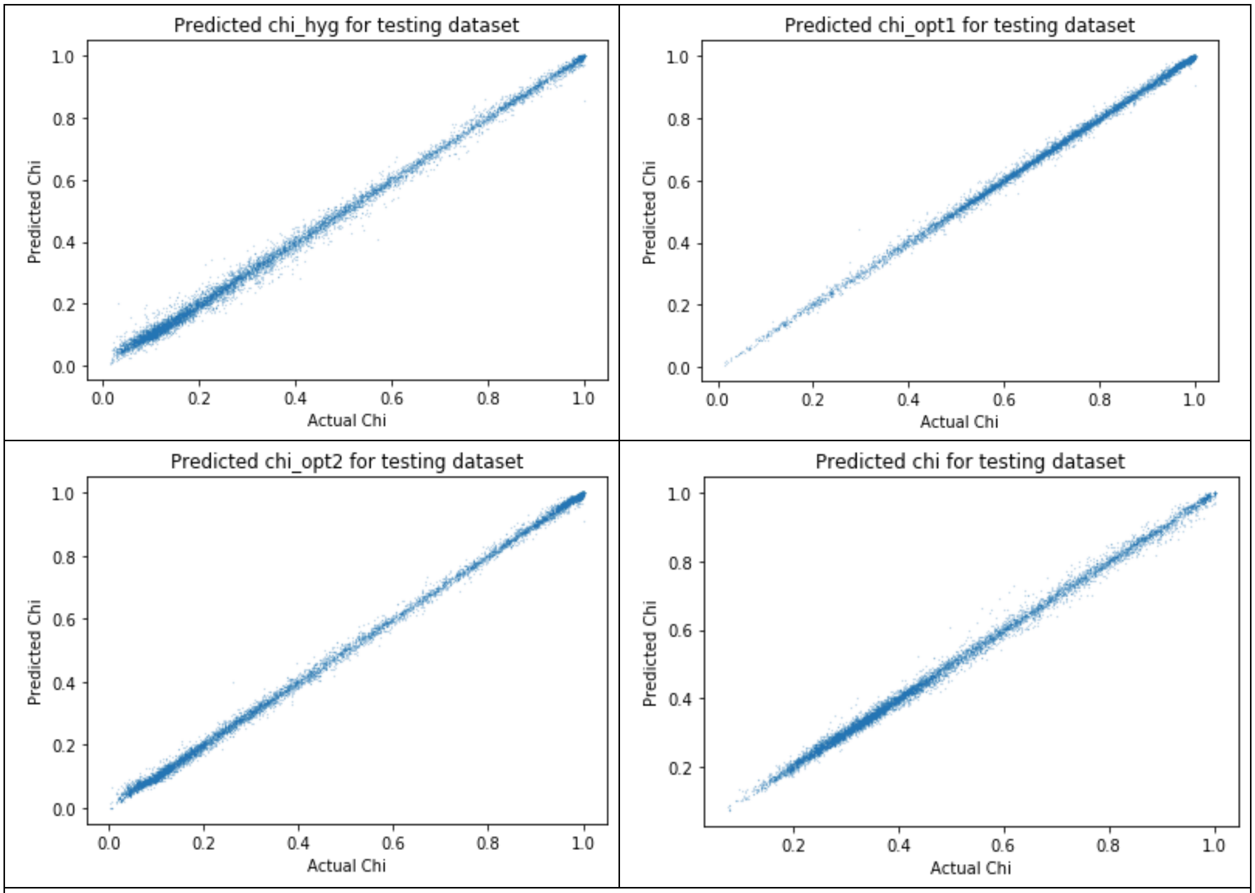

Figure 8. Predictions of multi-aerosol mixing state metrics in the testing set.

4. SUMMARY

4.1 CONCLUSIONS

This study advances the frontier of atmospheric modeling by creating the first-generation machine learning models to predict the multi-aerosol mixing state metrics, which enable us to gain new fundamental understanding about 1) how machine learning can be applied to improve the representation of aerosol mixing state and; and 2) where the inappropriate assumptions of aerosol mixing state may lead to large errors at a global scale. The major conclusion from this investigation are summarized as follows:

(1) Ensemble approach outperforms the non-ensemble XGB and deep learning approaches. Truncating the results within the boundary for the non-ensemble approach not only ensures the prediction but also improves the accuracy of model predictions.

(2) An ensemble approach was established to couple the XGB and deep learning methods. This approach can predict the multi-aerosol mixing state metrics with acceptable predictive power ($R^2 = 0.99$, IOA = 0.99).

(3) The tradeoff between runtime and accuracy needs to be considered when choosing the predictive models.

4.2 APPLICATIONS

This study provides feasible modeling options to predict the mixing state metric(s) as required. For instance, the deep learning model might be adopted when considering the long-term or high-resolution simulations given the shorter runtime, while the ensemble approach might be possible when the users aspire sufficiently confident. Since the features in all the models are a subset of features (outputs) of E3SM. Global distributions of multi-aerosol mixing state metrics can be generated by feeding the E3SM simulations to the models. This global distribution will re-envision the understanding of current aerosol representations in a global model.

4.3 FUTURE DIRECTIONS

Further studies are foreseen to investigate the robustness of current approaches. Currently, we utilize 90% of the data set as the training set and create the models. Using a reduced data set as a training set may enable us to develop stable approaches that are loosely coupled with the distribution of data.

The Energy Exascale Earth System Model Version 1 (E3SM-V1), a state-of-the-science Earth system model was released recently. This model is able to provide modeling, simulation, and prediction that optimize the use of DOE laboratory resources to meet the science needs, which offers a new solution to the development of the global mixing state distribution. A significant benefit of the E3SM is its high-resolution global simulation, which allows producing a higher resolution global maps of the mixing state metrics compared to existing map.

At the same time, efforts should also be invested in creating a wide variety of the scenarios. Given our current dataset only contains 1240 scenarios in total, massive scenarios similar to the E3SM simulation will advance better model development.

5. REFERENCES

Albrecht, B. A. (1989). Aerosols, cloud microphysics, and fractional cloudiness. *Science*, *245*(4923), 1227-1230.

Bein, K. J., Zhao, Y., Wexler, A. S., & Johnston, M. V. (2005). Speciation of size-resolved individual ultrafine particles in Pittsburgh, Pennsylvania. *Journal of Geophysical Research: Atmospheres*, *110*(D7).

Bengio Yoshua (2016, January 19). *Bengio Yoshua's Answer to Is cross-validation heavily used in deep learning or is it too expensive to be used?* Retrieved from https://www.quora.com/Is-cross-validation-heavily-used-in-deep-learning-or-is-it-too-expensive-to-be-used

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., ... & Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, *5*, 10312.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

Chin, M., Diehl, T., Ginoux, P., & Malm, W. (2007). Intercontinental transport of pollution and dust aerosols: implications for regional air quality. *Atmospheric Chemistry and Physics*, *7*(21), 5501-5517.

China, S., Mazzoleni, C., Gorkowski, K., Aiken, A. C., & Dubey, M. K. (2013). Morphology and mixing state of individual freshly emitted wildfire carbonaceous particles. *Nature communications*, *4*, 2122.

Ching, J., Fast, J., West, M., and Riemer, N. (2017). Metrics to quantify the importance of mixing state for CCN activity. *Atmospheric Chemistry Physics, 17,* 7445-7458, https://doi.org/10.5194/acp-17-7445-2017.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

Dickau, M., Olfert, J., Stettler, M. E., Boies, A., Momenimovahed, A., Thomson, K., ... & Johnson, M. (2016). Methodology for quantifying the volatile mixing state of an aerosol. *Aerosol Science and Technology*, *50*(8), 759-772.

Doran, J. C., Barnard, J. C., Arnott, W. P., Cary, R., Coulter, R., Fast, J. D., ... & Paredes-Miranda, G. (2007). The T1-T2 study: evolution of aerosol properties downwind of Mexico City. *Atmospheric Chemistry and Physics*, *7*(6), 1585-1598.

Drucker, J. (2013). Industrial structure and the sources of agglomeration economies: evidence from manufacturing plant production. *Growth and Change*, *44*(1), 54-91.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes*, *7*(4), 574-578.

Ghan, S. J., Liu, X., Easter, R. C., Zaveri, R., Rasch, P. J., Yoon, J. H., & Eaton, B. (2012). Toward a minimal representation of aerosols in climate models: Comparative decomposition of aerosol direct, semidirect, and indirect radiative forcing. *Journal of Climate*, *25*(19), 6461-6476.

Heyder, J. (2004). Deposition of inhaled particles in the human respiratory tract and consequences for regional targeting in respiratory drug delivery. *Proceedings of the American Thoracic Society*, *1*(4), 315-320.

Hinds, W. C. (2012). *Aerosol technology: properties, behavior, and measurement of airborne particles*. John Wiley & Sons.

Hughes, M., Kodros, J. K., Pierce, J. R., West, M., & Riemer, N. (2018). Machine Learning to Predict the Global Distribution of Aerosol Mixing State Metrics. *Atmosphere*, *9*(1), 15.

Jacob, Daniel. *Introduction to atmospheric chemistry*. Princeton University Press, 1999.

Jacobson, M. Z. (2001). Strong radiative heating due to the mixing state of black carbon in atmospheric aerosols. *Nature*, *409*(6821), 695.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Martins, V., Minguillón, M. C., Moreno, T., Querol, X., de Miguel, E., Capdevila, M., ... & Lazaridis, M. (2015). Deposition of aerosol particles from a subway microenvironment in the human respiratory tract. *Journal of Aerosol Science*, *90*, 103-113.

Noble, C. A., & Prather, K. A. (2000). Real-time single particle mass spectrometry: A historical review of a quarter century of the chemical analysis of aerosols. *Mass Spectrometry Reviews*, *19*(4), 248-274.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama, 287(9), 1132-1141.

Riemer, N., West, M., Zaveri, R. A., & Easter, R. C. (2009). Simulating the evolution of soot mixing state with a particle-resolved aerosol model. *Journal of Geophysical Research: Atmospheres*, *114*(D9).

Riemer, N., & West, M. (2013). Quantifying aerosol mixing state with entropy and diversity measures. *Atmospheric Chemistry and Physics*, *13*(22), 11423-11439.

Schill, S. R., Collins, D. B., Lee, C., Morris, H. S., Novak, G. A., Prather, K. A., ... & Cappa, C. D. (2015). The impact of aerosol particle mixing state on the hygroscopicity of sea spray aerosol. *ACS central science*, *1*(3), 132-141.

Seigneur, C., Hudischewskyj, A. B., Seinfeld, J. H., Whitby, K. T., Whitby, E. R., Brock, J. R., & Barnes, H. M. (1986). Simulation of aerosol dynamics: A comparative review of mathematical models. *Aerosol Science and Technology*, *5*(2), 205-222.

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain informatics*, *4*(3), 159.

Twomey, S. (1977). The influence of pollution on the shortwave albedo of clouds. *Journal of the atmospheric sciences*, *34*(7), 1149-1152.

Wang, Y., Plewa, M. J., Mukherjee, U. K., & Verma, V. (2018). Assessing the cytotoxicity of ambient particulate matter (PM) using Chinese hamster ovary (CHO) cells and its relationship with the PM chemical composition and oxidative potential. *Atmospheric Environment*, *179*, 132-141.

Wexler, A. S., Lurmann, F. W., & Seinfeld, J. H. (1994). Modelling urban and regional aerosols—I. Model development. *Atmospheric Environment*, *28*(3), 531-546.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, 213-251.

Zhang, Y., Seigneur, C., Seinfeld, J. H., Jacobson, M. Z., & Binkowski, F. S. (1999). Simulation of aerosol dynamics: A comparative review of algorithms used in air quality models. *Aerosol Science & Technology*, *31*(6), 487-514.

APPENDIX A. REPRODUCIBILITY

APPENDIX A. REPRODUCIBILITY

The code and data

The code repository is available at https://code.ornl.gov/vga/aerosol-msm-partmc-v0.git

Brief instruction for re-implementation

a. Systems and platforms

   NVIDIA DGX-1 artificial intelligence supercomputer

   ```
   ssh -L 8880:localhost:8880 <username>@deep.ornl.gov
   ```

b. Which containers to use

   ```
   docker_image=zzjn-image
   external_folder=/home/<username>
   internal_folder=/workspace/<username>

   sudo nvidia-docker run --shm-size=1g --ulimit memlock=-1 -p 8880:8880 --ulimit stack=67108864 -it -v $external_folder:$internal_folder $docker_image
   ```

c. Any necessary configuration changes

   Use chrome to launch the Jupyter Notebook on local machine: http://localhost:8880/

d. Installing additional software

   The packages for the implementations include: math, numpy, pandas, matplotlib, tensorflow, sklearn, xgboost, pickle.

e. Additional instructions

   Please contact zzheng25@illinois.edu if you have any questions and suggestions.

Verification of the results

All the Jupyter notebooks within the “/tutorial/ipynb” are self-explanatory. The results within each cell should be similar with the reference results.

Please contact zzheng25@illinois.edu if you have any questions and suggestions.