AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

Estimating Submicron Aerosol Mixing State at the Global Scale With Machine Learning and Earth System Modeling

Zhonghua Zheng[1] , Jeffrey H. Curtis[2,3] , Yu Yao[2] , Jessica T. Gasparik[2] , Valentine G. Anantharaj[4] , Lei Zhao[1,5] , Matthew West[3] , and Nicole Riemer[2]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA, [2]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA, [3]Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA, [4]National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA, [5]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Abstract This study integrates machine learning and particle-resolved aerosol simulations to develop emulators that predict submicron aerosol mixing state indices from the Earth system model (ESM) simulations. The emulators predict aerosol mixing state using only quantities that are predicted by the ESM, including bulk aerosol species concentrations, which do not by themselves carry mixing state information. We used PartMC-MOSAIC as the particle-resolved model and NCAR's CESM as the ESM. We trained emulators for three different mixing state indices for submicron aerosol in terms of chemical species abundance (χ_a), the mixing of optically absorbing and nonabsorbing species (χ_o), and the mixing of hygroscopic and nonhygroscopic species (χ_h). Our global mixing state maps show considerable spatial and seasonal variability unique to each mixing state index. Seasonal averages varied spatially between 13% and 94% for χ_a, between 38% and 94% for χ_o, and between 20% and 87% for χ_h with global annual averages of 67%, 68%, and 56%, respectively. High values in one index can be consistent with low values in another index depending on the grouping of species and their relative abundance, meaning that each mixing state index captures different aspects of the population mixing state. Although a direct validation with observational data has not been possible yet, our results are consistent with mixing state index values derived from ambient observations. This work is a prototypical example of using machine learning emulators to add information to ESM simulations.

Plain Language Summary Earth system models (ESMs) simulations are computationally expensive, requiring highly simplified representations of aerosol mixing state, a property that describes how different aerosol chemical species are distributed among and within the aerosol particles. The assumption of whether aerosols are internally (multiple species within a particle), externally (one species per single particle), or intermediately mixed greatly influences the properties of aerosol particles and thereby the prediction of the impacts of air pollution on human health and climate change. We built simplified models using machine learning and highly detailed particle-resolved simulations to infer submicron aerosol mixing state from meteorological parameters and pollution levels. These emulators enable us to estimate the degree of aerosol mixing state at a global scale using information that ESMs track. This study provides an example of the integration of detailed aerosol process modeling and a large-scale ESM via machine learning.

1. Introduction

Aerosol particles in the atmosphere are evolving mixtures of different chemical species. The term "aerosol mixing state" is used to describe how different aerosol chemical species are distributed among and within the aerosol particles (Riemer et al., 2019). As stated by Winkler (1973), "the same net composition of an aerosol can be caused by an infinite variety of different internal distributions of the various compounds." An "internally mixed" aerosol refers to the state where the composition of all particles within the population is the same (and equal to the bulk composition of the aerosol), while an "externally mixed" aerosol has all particles of a population consisting of only a single species. In reality, aerosol mixing states are between

internal and external mixtures as shown in many observational studies (e.g., Bondy et al., 2018; Healy et al., 2014; Lee et al., 2019; Ye et al., 2018; Yu et al., 2020). Aerosol mixing state greatly influences the particles' hygroscopicity (Fierce et al., 2017; Holmgren et al., 2014), their optical properties (Fierce et al., 2016; Lesins et al., 2002), their cloud condensation nuclei (CCN) activity (Ching et al., 2012; Wang et al., 2010), their ice nucleation potential (Knopf et al., 2018), their deposition in the human respiratory system (Ching & Kajino, 2018), and the aerosols' lifetime in the atmosphere (Koch et al., 2009).

To quantify mixing state, Riemer and West (2013) introduced a metric, the mixing state index χ, based on diversity measures derived from the information-theoretic (Shannon) entropy of the chemical species distribution among particles, with per-particle mass fractions as the fundamental quantities. The mixing state index χ varies between 0% (for completely external mixtures) and 100% (for completely internal mixtures) for any given aerosol. The metric has been applied to field observations in different environments, for example, Paris during the MEGAPOLI campaign (Healy et al., 2014), in Northern California during CARES (O'Brien et al., 2015), in central London (Giorio et al., 2015), and in Pittsburgh, PA (Ye et al., 2018). It has provided useful insights into the processes that govern diurnal changes in mixing state, and mixing state changes related to air mass origin. It can also be used for error quantification as shown in Ching et al. (2017), Ching et al. (2018), and Ching and Kajino (2018). Ching et al. (2017) quantified errors in CCN concentration prediction when assuming a fully internal mixture compared to using the realistic mixing state predicted by a particle-resolved model. In a follow-up study, Ching et al. (2018) applied this concept to errors in the prediction of black carbon (BC) mass fraction that is incorporated into cloud droplets by nucleation-scavenging. Considering the deposition of soot particles in the human respiratory system, Ching and Kajino (2018) demonstrated that the mixing state index is also an important metric for health impact evaluation.

The most direct way of predicting aerosol mixing state and its index χ is to employ a particle-resolved model. This modeling approach allows for tracking the composition of individual particles and therefore χ can be directly calculated. However, particle-resolved modeling is extremely computationally expensive and thus not practical for use in large-scale models. Alternatively, Hughes et al. (2018) developed a method that uses the output of a large ensemble of particle-resolved box model simulations combined with machine learning techniques to train a model of the mixing state metric χ. This lower-order model for χ uses as inputs only variables known to the global climate model of interest (in Hughes et al. [2018] it was GEOS-Chem-TOMAS, which uses a sectional aerosol modeling approach assuming an internal mixture within each size bin). The outcomes of this procedure were global maps of χ based on GEOS-Chem-TOMAS data.

In this work, we revisit the approach by Hughes et al. (2018), but extend the analysis by defining the mixing state index in three different, but complementary, ways (see Section 3): in terms of chemical species abundance, in terms of the mixing of optically absorbing and nonabsorbing species, and in terms of the mixing of hygroscopic and nonhygroscopic species. We discuss their relationship with each other and their interpretation in Section 5.

We used the Community Earth System Model (CESM; Danabasoglu et al., 2020; Hurrell et al., 2013) in conjunction with the Modal Aerosol Module (MAM4; Liu et al., 2016) as a large-scale model that provides the features for the machine learning procedure. We compared and contrasted the spatial distribution and seasonal variation of aerosol mixing state indices at a global scale.

Our workflow is shown in Figure 1 and is composed of three parts: (1) PartMC-MOSAIC (Riemer et al., 2009; Zaveri et al., 2008) simulations to produce training data (see Section 2), (2) machine learning-enabled aerosol mixing state emulation (Section 4), and (3) global aerosol mixing state indices estimation (Section 5).

2. Ensemble of Particle-Resolved Model Scenarios

To generate the training data, we ran an ensemble of particle-resolved model scenarios using the Part-MC-MOSAIC. This section gives an overview of the model and the design of the scenario ensemble.
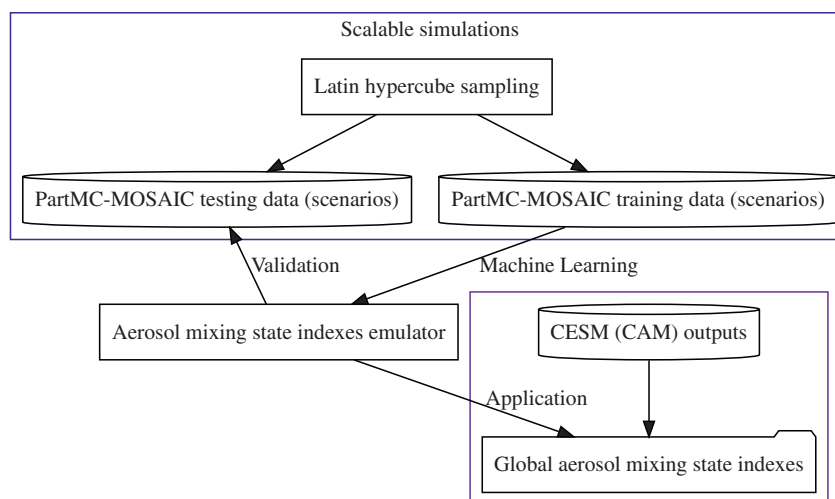
Figure 1. Schematic of the workflow for estimating the mixing state index χ on a global scale.

2.1. PartMC-MOSAIC Model Description

PartMC is a particle-resolved tool to model the aerosol mixing state and its impacts under various meteorological and environmental conditions (Riemer et al., 2009). In brief, this box model simulates individual aerosol particles within a representative volume of air, including stochastic coagulation, gas- and particle-phase chemistries, particle-phase thermodynamics, and dynamic gas-particle mass transfer. The PartMC algorithm has storage cost proportional to the number of particles, computational cost for evaporation/condensation proportional to the number of particles, and computational cost for coagulation proportional to the number of coagulation events. PartMC has been coupled with the state-of-the-art aerosol chemistry model MOSAIC (Model for Simulating Aerosol Interactions and Chemistry; Zaveri et al., 2008) to provide the chemistry modules. MOSAIC consists of four computationally efficient modules: (1) the gas-phase photochemical mechanism CBM-Z (Zaveri & Peters, 1999); (2) the multicomponent Taylor expansion method for estimating activity coefficients of electrolytes and ions in aqueous solutions (Zaveri, Easter, & Wexler, 2005); (3) the multicomponent equilibrium solver for aerosols (MESA) for intraparticle solid-liquid partitioning (Zaveri, Easter, & Peters, 2005); and (4) the adaptive step time-split Euler method for dynamic gas-particle partitioning over size- and composition-resolved aerosol (Zaveri et al., 2008). Secondary organic aerosol (SOA) from anthropogenic and biogenic precursors is included using the Secondary Organic Aerosol Model (Schell et al., 2001). Overall, MOSAIC treats all locally and globally important gas and aerosol species, including sulfate, nitrate, chloride, carbonate, ammonium, sodium, calcium, primary organic aerosol, SOA, BC, and inert inorganic mass (a surrogate species for mineral dust).

The coupled model system, PartMC-MOSAIC, predicts number, mass, and full composition distributions without a priori assumptions about their evolution, which is suitable for use as a numerical benchmark of mixing state for more approximate models (Fierce et al., 2016, 2017; Kaiser et al., 2014). Here, we used PartMC-MOSAIC simulations to provide data for training and testing emulators that predict the mixing state indices on the global scale.

2.2. Design of Training and Testing Scenarios

The PartMC-MOSAIC training and testing scenarios provide a large number of particle populations with different mixing states aiming at covering the range of conditions found in different environments around the globe. Our strategy to generate this data was to vary the input parameters (primary emissions of different aerosol types, and primary emissions of gas phase species, which served as precursors for secondary aerosol, as well as meteorological parameters; 45 parameters in total), as listed in Table 1. To sample this parameter space efficiently, we employed a Latin Hypercube sampling approach to assemble input parameter

Table 1
List of Input Parameters and Their Sampling Ranges to Construct the Training and Testing Scenarios

Parameters	Range
Environmental variables	
Relative humidity (RH)	[0.1, 1) or [0.4, 1)
Latitude	(70°S, 70°N) or (90°S, 90°N)
Day of year	[1, 365]
Temperature	Based on latitude and day of year
Gas phase emissions scaling factor relative to base case	
SO_2, NO_2, NO, NH_3, CO, CH_3OH,	[0, 200%] or [0, 100%]
ALD2 (Acetaldehyde), ANOL (Ethanol),	
AONE (Acetone), DMS (Dimethyl sulfide),	
ETH (Ethene), HCHO (Formaldehyde),	
ISOP (Isoprene), OLEI (Internal olefin carbons),	
OLET (Terminal olefin carbons),	
PAR (Paraffin carbon), TOL (Toluene), XYL (Xylene)	
Sulf-Aitken: Sulfate emissions (one mode)	
D_g	[15, 53 nm]
σ_g (constant)	1.6
E_a	[0, 2.0×10^9 m^{-2} s^{-1}]
carb: Carbonaceous aerosol emissions (one mode)	
D_g	[25, 250 nm]
σ_g	[1.4, 2.5]
BC/OC mass ratio	[0, 100%]
E_a	[0, 1.6×10^7 m^{-2} s^{-1}]
SS-Aitken: Sea salt emissions (one mode)	
D_g	[15, 53 nm]
σ_g (constant)	1.6
E_a	[0, 2.0×10^8 m^{-2} s^{-1}]
SS: Sea salt emissions (two modes)	
$D_{g,1}$	[180, 720 nm]
$\sigma_{g,1}$	[1.4, 2.5]
$E_{a,1}$	[0, 1.69×10^5 m^{-2} s^{-1}]
$D_{g,2}$	[1 μm, 6 μm]
$\sigma_{g,2}$	[1.4, 2.5]
$E_{a,2}$	[0, 2380 m^{-2} s^{-1}]
OC fraction	[0, 20%]
dust: Dust emissions (two modes)	
$D_{g,1}$	[80, 320 nm]
$\sigma_{g,1}$	[1.4, 2.5]
$E_{a,1}$	[0, 5.86×10^5 m^{-2} s^{-1}]
$D_{g,2}$	[1 μm, 6 μm]
$\sigma_{g,2}$	[1.4, 2.5]
$E_{a,2}$	[0, 2380 m^{-2} s^{-1}]
Restart timestamp	
Timestamp	[0, 25 h]

Table 1
Continued

Parameters	Range
Note. The variables D_g, σ_g, and E_a refer to geometric mean diameter, geometric standard deviation, and particle emission flux, respectively.	

combinations for 2000 different scenarios for the training data set and 180 scenarios as the testing data set, similar to the procedure used in Hughes et al. (2018). All scenarios were run with 10,000 computational particles. They started at 6:00 a.m. local time and used a simulation time of 24 h, saving the hourly output.

We obtained the daily near-surface temperature from the CESM2 (0.9° by 1.25° resolution) fully coupled historical simulations (Danabasoglu, 2019). For each latitude and day of year (DOY) over the years from 1970 to 2014, we retrieved the maximum and minimum temperatures. For a given pair of latitude and DOY in PartMC-MOSAIC, the temperature was then sampled between the maximum and minimum temperatures at the nearest latitude in CESM2 with the same DOY. Gas initial conditions and emission rates were based on Riemer et al. (2009). We sampled the gas phase emissions using a scaling factor (0%–200%) for the gas phase species listed in Table 1.

We set up five blocks of scenarios (blocks A–E) that differed in the way the aerosol types were combined. Table 2 specifies these simulations blocks using abbreviations as defined in Table 1. Block A consisted of 800 scenarios that included carbonaceous aerosol, dust, and sea salt. We uniformly sampled the relative humidity for these scenarios from a range of 40%–100%, and latitude from a range of 90°S to 90°N. The latitude governs the solar zenith angle and the length of the day, which are important for the photochemical production of secondary aerosol. We included initial concentration (Omori et al., 2017), and emissions (Xu et al., 2016) of the biogenic trace gas dimethyl sulfide (DMS) for the scenarios with sea salt emission, as it is a precursor of sulfate aerosol in marine environments (Lana et al., 2011).

Block B had 400 scenarios and the same setup as Block A, but did not contain dust emissions. Block C had 400 scenarios that included only carbonaceous aerosol and dust (no sea salt). For the Block C scenarios, we uniformly sampled the relative humidity from 10% to 100%, and latitude from 70°S to 70°N. Block D comprised 200 scenarios and only contained carbonaceous aerosol. Otherwise, the setup was the same as for Block C. Block E contained carbonaceous aerosol and sea salt similar to Block B, but had additional sea salt and sulfate emissions in the Aitken mode size range. The purpose of Block E was to capture conditions where particles in the Aitken mode size range were present, including those resulting from new particle formation. The latitude range for the Block E scenarios was limited to high latitudes from 70°N–90°N and 70°S–90°S. This was motivated by the fact that, in these areas, CESM simulations showed the Aitken mode particles accounted for an appreciable fraction of submicron aerosol mass. Aitken mode sulfate particles were introduced into the PartMC-MOSAIC simulation by emission rather than by simulating the process of new particle formation and growth explicitly. While PartMC-MOSAIC includes the process of new particle formation (Tian et al., 2014), considerable uncertainty exists regarding the subsequent growth of the freshly nucleated particles (Kulmala et al., 2014), which poses a challenge for a highly detailed aerosol model such as PartMC-MOSAIC.

To create aerosol initial conditions with realistic mixing states we adopted the following approach: We performed a first set of runs, starting with the aerosol initial concentrations set to zero for all the simulations (the "initial runs"). We then repeated the same set of runs, but replaced the aerosol initial condition with a randomly sampled population from the initial runs (the "restart runs"). For the training and testing datasets, we only used the results from the restart runs. The size distribution parameters for the aerosol emissions were prescribed as log-normal, with geometric mean diameter (D_g) and geometric standard deviation (σ_g) depending on the particle type.

Table 2
Specification of Scenario Blocks

Block	Number of scenarios	Latitude range	Relative humidity range (%)	Gas phase emissions scaling factor (%)	Modes (see Table 1)
A	800	90°S-90°N	40–100	0–200	carb, dust, SS
B	400	90°S-90°N	40–100	0–200	carb, SS
C	400	70°S-70°N	10–100	0–200	carb, dust
D	200	70°S-70°N	10–100	0–200	carb
E	200	70°N-90°N, 70°S-90°S	40–100	0–100	carb, SS-Aitken, Sulf-Aitken, SS

3. Aerosol Mixing State Indices Definition

The mixing state index χ (Riemer & West, 2013) quantifies where an aerosol population lies on the continuum of external to internal mixing, that is, how "spread out" the chemical species are over an aerosol population, see Figure 2 for an illustration. We will focus here on the aerosol mixing state of submicron particles (PM$_{1.0}$) due to their impact on light scattering and absorption (Wang et al., 2015), and contribution to CCN formation (Asmi et al., 2011; Pierce et al., 2015; Yu & Luo, 2009).

The mixing state index χ is given by the affine ratio of the average particle species diversity (D_α) and bulk population species diversity (D_γ) as

$$\chi = \frac{D_\alpha - 1}{D_\gamma - 1}. \tag{1}$$

The diversities D_α and D_γ are calculated as follows. First, the per-particle mixing entropies H_i are determined for each particle using the per-particle species mass fractions:

$$H_i = \sum_{a=1}^{A} - p_i^a \ln p_i^a. \tag{2}$$

Here, A is the number of distinct aerosol species and p_i^a is the mass fraction of species a in particle i. These values are then averaged (mass-weighted) over the entire population to find the average particle entropy, H_α, and the average particle species diversity, D_α, by
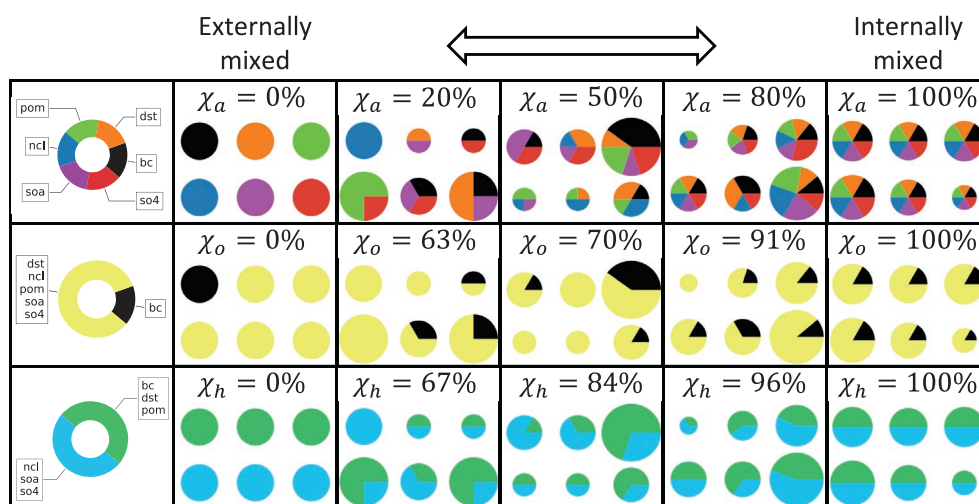

Figure 2. The concept of aerosol mixing state indices χ_a, χ_o, and χ_h for a population consisting of six chemical species, where the bulk mass fractions of species are the same for all six species.

Table 3
Aerosol Mixing State Indices Definition

Aerosol mixing state index (abbr)	Grouped species
Abundance (χ_a)	(bc), (dst), (ncl), (pom), (soa), (so4)
Optical property (χ_o)	(bc), (dst, ncl, pom, soa, so4)
Hygroscopicity (χ_h)	(bc, dst, pom), (ncl, soa, so4)

Note. Six aerosol species (see Table 4) are used in calculating the aerosol mixing state indices based on different definitions. We calculate χ_a based on all six aerosol model species. The mixing state indices χ_o and χ_h are based on two surrogate species.

$$H_\alpha = \sum_{i=1}^{N_p} p_i H_i, \tag{3}$$

$$D_\alpha = e^{H_\alpha}, \tag{4}$$

where N_p is the total number of particles in the population, and p_i is the mass fraction of particle i in the population. Finally, the bulk entropy, H_γ, and bulk diversity, D_γ, is calculated as:

$$H_\gamma = \sum_{a=1}^{A} -p^a \ln p^a, \tag{5}$$

$$D_\gamma = e^{H_\gamma}. \tag{6}$$

Importantly, the definition of the "species" to calculate D_α and D_γ depends on the application or may be dictated by the instrumentation used to estimate per-particle mass fractions. For example, in some previous studies, elemental species (e.g., N, O, and C) have been used (Bondy et al., 2018; Fraund et al., 2017; O'Brien et al., 2015), while others used molecular species (e.g., SO_4, NH_4, and NO_3) (Healy et al., 2014; Lee et al., 2019; Ye et al., 2018). Yet other studies employed surrogate species where several species with similar properties were grouped together, such as hygroscopic/nonhygroscopic species (Ching et al., 2017; Hughes et al., 2018), or volatile/nonvolatile species (Dickau et al., 2016). The choice of which species to group as one surrogate species is then motivated by the science question under investigation.

Here, we compare and contrast the aerosol mixing state indices defined in three different ways, namely based on model chemical species abundance (χ_a), based on the mixing of optically absorbing and nonabsorbing species (χ_o), and based on the mixing of hygroscopic and nonhygroscopic species (χ_h). Table 3 shows the specific definitions of these aerosol mixing state indices. The index χ_a was defined based on all six model aerosol species. For χ_o, we considered two surrogate species, BC (absorbing) and all other aerosol species grouped together (assumed to be nonabsorbing). Thus, a lower value in χ_o refers to the case where the absorbing species BC and the sum of all other (nonabsorbing) species are more externally mixed. Similarly, χ_h was also calculated from two surrogate species. We combined BC, dust, and primary organic matter as one surrogate species, given their comparatively low hygroscopicities, and NaCl, SOA, and sulfate as the other surrogate species. Here, a lower value in χ_h represents the case where hygroscopic and nonhygroscopic species tend to be present in separate particles.

Figure 2 illustrates the concept of different aerosol mixing state indices corresponding to a particle population which consists of six particles and six chemical species that are present in equal amounts in the bulk. Moving from left to right, the bulk composition for each population remains the same, but the population becomes more internally mixed. That is, the particles become more similar to each other. This is reflected in an increasing mixing state index, and this applies to all three mixing state indices individually. Another lesson is that the mixing state index depends on how we define the surrogate species. In our example, the population with a mixing state index of $\chi_a = 20\%$ has $\chi_o = 63\%$ and $\chi_h = 67\%$. For this example, the population is quite externally mixed with respect to all chemical model species, but appears more internally mixed with respect to hygroscopicity or optical properties. A consequence of the definition of the surrogate species as supersets of the chemical model species is that, for the extreme cases of $\chi_a = 100\%$ and $\chi_a = 0\%$, χ_o and χ_h are also 100% or 0%, respectively. The reverse, however, is not true. That is, $\chi_o = 100\%$ does not imply $\chi_a = 100\%$. In general, the exact mapping between χ_a, χ_h, and χ_o is not straightforward and depends on the relative abundance of the different (surrogate) species.

Table 4
Feature (Variable) Definitions in CESM and PartMC-MOSAIC

Feature name	CESM	PartMC	Type	Unit
Black carbon	(bc_a1_SRF, bc_a4_SRF)	BC	A	kg/kg
Mineral dust	(dst_a1_SRF, dst_a2_SRF)	OIN	A	kg/kg
Sea salt	(ncl_a1_SRF, ncl_a2_SRF)	(Na, Cl)	A	kg/kg
Primary organic matter	(pom_a1_SRF, pom_a4_SRF)	OC	A	kg/kg
Secondary organic aerosol	(soa_a1_SRF, soa_a2_SRF)	(ARO1, ARO2, ALK1, OLE1, API1, API2, LIM1, LIM2)	A	kg/kg
Sulfate	(so4_a1_SRF, so4_a2_SRF)	SO4	A	kg/kg
Dimethyl sulfide	DMS	DMS	G	mol/mol
Hydrogen peroxide	H2O2	H2O2	G	mol/mol
Sulfuric acid	H2SO4	H2SO4	G	mol/mol
Zone	O3	O3	G	mol/mol
Semi-volatile organic gas	SOAG	(ARO1, ARO2, ALK1, OLE1, API1, API2, LIM1, LIM2)	G	mol/mol
Sulfur dioxide	SO2	SO2	G	mol/mol
Air temperature	T	temperature	E	K
Relative humidity	RELHUM	relative_humidity	E	1
Solar zenith angle	SZA	solar_zenith_angle	E	radian

Note. Type: A (aerosol), G (gas), E (environmental). The words "SRF," "a1," "a2," and "a4" refer to surface, Accumulation mode, Aitken mode, and Primary carbon mode.

4. PartMC-MOSAIC Emulator Development and Application

4.1. Feature Definitions

The features for the emulator were selected according to the following two criteria. First, the features should be available within both the PartMC-MOSAIC and CESM simulations, so that we have consistent features for training, validation, and application processes. In other words, quantities that are only available in Part-MC-MOSAIC but not in CESM simulations cannot serve as features for the emulators. Second, the features should be physically or chemically related to the aerosol properties. A subset of the aerosol and gas phase species and environmental variables were selected as features in this study (Table 4). The specific aerosol species were defined by the modal aerosol module (MAM4; Liu et al., 2012, 2016) and PartMC-MOSAIC.

4.2. Emulator Development

We used eXtreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016) with a L_2 regularization term on weights as the machine learning algorithm for our emulator development. Built upon the Gradient Boosting
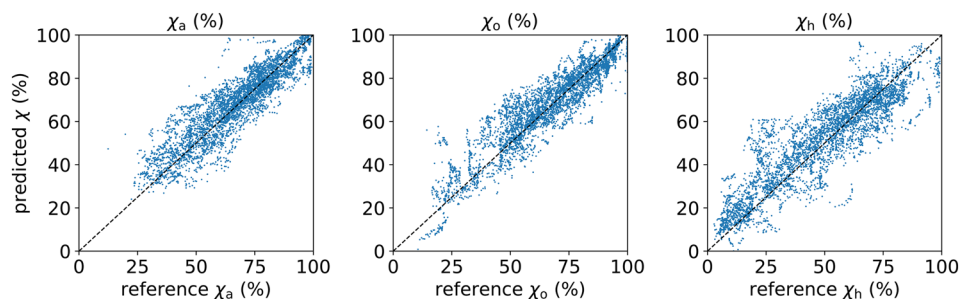
Figure 3. Emulator performance comparing the predicted mixing state indices with the reference indices, as measured on the testing data set.

(Friedman, 2001) framework, XGBoost is highly efficient, flexible, and portable. It can not only handle complex nonlinear interactions and collinearity among predictors, but also provides a parallel implementation that solves data science problems (particularly with structured data) efficiently. It has been successfully used in various fields such as physics (Mott et al., 2017), medical research (Abelson et al., 2018), and ornithology (Rosenberg et al., 2019). Here, we developed three emulators, corresponding to the three aerosol mixing state indices outlined in Section 3.

For each aerosol mixing state index, the emulator was trained from the entire 1800 training scenarios with hyperparameters (number of gradient boosted trees, maximum tree depth for base learners, and boosting learning rate) previously determined by using grid search with 10-fold cross-validation. Figure 3 shows the performance of emulators using the testing scenarios, which have never been seen by the training process. Table 5 shows the predictive accuracy of the XGBoost-based emulators compared to the benchmark using ordinary least squares. These metrics are important to keep in mind when interpreting the results in Section 5. The error for any predicted mixing state index can be roughly thought of as ±10%, and somewhat higher for χ_h.

Figure 4 shows the overall feature importance based on the "gain" metric, which quantifies the improvement in accuracy brought by a feature to the branches it is on. A higher value of this metric when compared to another feature implies it is more important for generating a prediction. In other words, the predictions are more sensitive to the features with higher gain values. This shows that the aerosol species are most important, although the ranking is different for the different mixing state indices. Environmental conditions (relative humidity and temperature) are of medium importance, and the gas phase concentrations are of lower importance.

Table 5
Evaluation of Mixing State Emulators Using the Testing Data Set

χ	RMSE	MAE	MAD	d	PCC	r^2
			XGBoost Emulators			
χ_a	0.077	0.058	0.045	0.945	0.899	0.802
χ_o	0.079	0.060	0.046	0.957	0.922	0.845
χ_h	0.108	0.082	0.065	0.938	0.892	0.789
	Linear regression with ordinary least squares (OLS)					
χ_a	0.137	0.108	0.088	0.726	0.622	0.379
χ_o	0.152	0.116	0.093	0.783	0.680	0.434
χ_h	0.195	0.156	0.134	0.721	0.599	0.315

Note. The metrics are root-mean-square error (RMSE), mean absolute error (MAE), median absolute deviation (MAD), index of agreement (*d*; Willmott, 1981), Pearson correlation coefficient (PCC), and coefficient of determination (*r²*). The emulators from linear regression with ordinary least squares (OLS) are examined as a benchmark. The p value (<0.001) applies for both methods using the F-test.

4.3. Emulator Application

Applying the emulator is fast, compared to embedding PartMC-MOSAIC into regional models or ESMs for each grid cell. Here, we used CESM (version 2.1.0; Bogenschutz et al., 2018) component set, FHIST (Community Atmospheric Model standalone runs), as a testbed for our emulator application. This component set represents a typical historical simulation in the Community Atmospheric Model (CAM6), which uses an active atmosphere and land (satellite phenology) with prescribed sea-surface temperatures and sea-ice extent, and a 1-degree finite volume dycore with the forcing data available from 1979 to 2015. MAM4 is the default aerosol module of this component set, which predicts the aerosol size distribution with four lognormal modes (Aitken, accumulation, coarse, and primary carbon modes). The model treats the microphysical aging of the primary carbonaceous aerosols in the atmosphere (Liu et al., 2016). It contains six aerosol species, namely BC, dust, sea salt, primary organic matter, SOA, and sulfate across four modes. CESM/CAM6 is suitable as
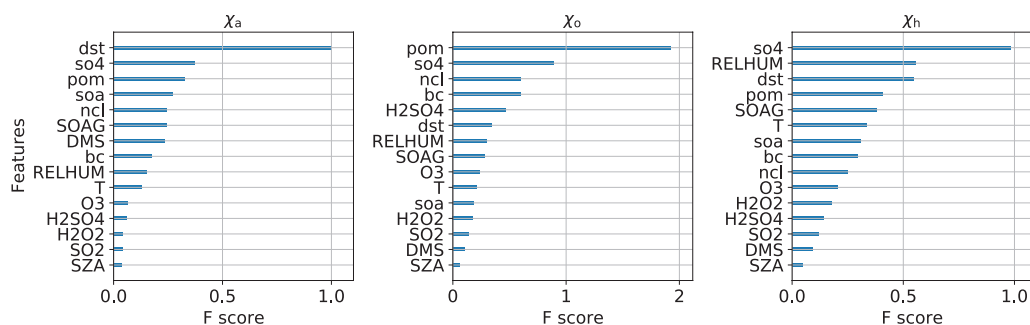
Figure 4. Overall feature importance based on the "gain" metric for the three mixing state indices χ_a, χ_o, and χ_h.

a testbed because CESM is a state-of-the-art and open-source ESM with many substantial science and infrastructure improvements (e.g., improved treatment of dust [Albani et al., 2014; Bogenschutz et al., 2018] and primary organic matter [Liu et al., 2016]), and improved historical simulations compared to available observations (Danabasoglu et al., 2020). The aerosol configuration in CESM/CAM has been calibrated by many observations (Liu et al., 2016; Tilmes et al., 2015). We ran the model for the year 2011 with 6 years (2005–2010) spin-up. See Gantt et al. (2014) and He and Zhang (2014) for details of the spin-up procedure. The simulation was conducted at a resolution of 0.9° latitude × 1.25° longitude with the emission inventories from CMIP6 emissions (Emmons et al., 2020). We stored the instantaneous outputs every 3 h in the simulation, which yielded 2920 timestamps for each surface-layer grid cell for the entire year of simulation time.

For reference, Figures S1 and S2 illustrate the seasonal variation and spatial distribution of the submicron aerosol species at the global scale. As we discussed in Text S1 (supporting information), the distributions were in line with previous studies (Bond et al., 2004, 2013; Hommel et al., 2011; Lack et al., 2004; Mahowald et al., 2005; Ramanathan & Carmichael, 2008; Sofiev et al., 2011; Tanaka & Chiba, 2006). It is worth mentioning that our simulation does not focus on optimizing the parameterizations of CESM. We used the default setting to conduct the simulations as a basis for estimating the different mixing state indices with the corresponding emulator in each grid cell. However, it should be noted that limitations and systematic biases in CESM (Hodzic et al., 2020; Schwantes et al., 2020) may be translated into the mixing state estimates.

4.4. Emulator Validation With Observational Data

To validate our χ emulators requires an observational data set that has both measured χ values as well as measured values for the emulator features (Table 4). Measuring χ requires quantitative per-particle mass fractions, which are difficult to estimate from observations (see Section 5.2). The data set that comes closest to having the required information was produced during the MEGAPOLI winter campaign (Healy et al., 2014) for a site in Paris, France from January 15 to February 11, 2010. We used hourly measurements from this observational data.

Healy et al. (2014) used single particle mass fraction estimates for BC, organic aerosol, ammonium, nitrate and sulfate to calculate single-particle diversities and the mixing state index χ_{obs}. The mass fraction estimates were derived from a combination of single particle mass spectrometer, aerosol mass spectrometer, and multiangle absorption photometer measurements.

Using this data set for validating our emulator poses some challenges. The species list used to calculate χ_{obs} was somewhat different from that used for χ_a in this paper (Table 3). Further, not all emulator features (Table 4) were measured during the campaign. The missing features (dust, sea salt, DMS, H2O2, H2SO4, SO2, and SOAG) were set as "NaN" when applying the emulator, which is how XGBoost indicates missing data. XGBoost has explicit support for missing data and it learns default tree branch directions for missing values during training. To partition the measured organic aerosol and create an estimate of the features "pom" and "soa," we assumed a ratio of 1:1, based on CESM results for the same grid cell.
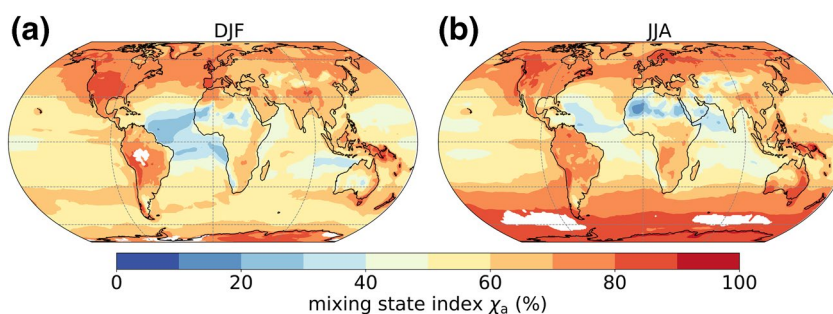
Figure 5. Average mixing state index χ_a based on chemical abundance in seasons (a) DJF and (b) JJA. The areas with the mass fraction of any one species higher than 97.5% are masked.

We should not expect a perfect agreement between the emulator values, χ_a, and observational values, χ_{obs}, due to the difference in species lists used to compute χ and the fact that 7 out of 15 features were missing, including the most important feature "dst" (left panel of Figure 4). Nevertheless, over all MEGOPOLI campaign data the emulator had a mean value of $\chi_a = 63\%$, which was close to the measured mean value of $\chi_{obs} = 59\%$. In terms of the distribution, the 25th percentile, median, and 75th percentile values for the emulator were $\chi_a = 56\%$, 63%, and 70%, which were also similar to the measured values of $\chi_{obs} = 54\%$, 61%, and 65%, respectively. The minimum and maximum values during the campaign were $\chi_a = 42\%$ and 88%, compared to $\chi_{obs} = 37\%$ and 72%. Although the emulator produced a good match of the distribution compared to the observations, the predictions did not capture the diurnal cycle, which may be due to the differences in species lists and the missing features. Despite the difficulties in validating the emulator with observational data, this distributional agreement provides some support for the accuracy of the emulator.

5. Global Aerosol Mixing State Indices

In this section, we will first present the seasonal variation and spatial distribution of the three aerosol mixing state indices as obtained from applying the emulators. We will then put these findings in context with available observations, and finish this section with quantifying the relationship between the three indices as they were predicted for the entire model domain.

5.1. Seasonal Variation and Spatial Distribution of Aerosol Mixing State Indices

We averaged the 3-h instantaneous values of aerosol mixing state index for the two seasons (December-January-February, DJF, and June-July-August, JJA) to determine their seasonal variations and spatial distributions. We masked the areas where the mass fraction of any one species is higher than 99% (for χ_o) or 97.5% (for χ_a and χ_h). The rationale for applying these masks is that only when at least two species are present in a given location does it make sense to quantify mixing state.

Figure 5 shows the global maps of the aerosol mixing state index χ_a based on chemical species abundance. In DJF, χ_a ranged from 19% to 93% over the globe, with a mean of 64%. Of note is the region across the Atlantic from West and Central Africa to the northeast coast of South America, where χ_a was smaller than 40%. In this region, dust and primary organic matter were prevalent, with contributions of sea salt (Figure S2), and so this result can be interpreted to mean that these species are comparatively, although not completely, externally mixed.

In contrast, aerosol species were rather internally mixed ($\chi_a > 70\%$) over North America, the North Atlantic Ocean, the Arctic Ocean, East Asia, and the North Pacific Ocean. The aerosol composition varied from region to region. Briefly, in DJF, aerosols over the continents of those regions were dominated by primary organic matter, SOA, and sulfate, while sea salt and sulfate governed the aerosol composition over the oceans. The high values of χ_a suggest that the species in these locations are rather internally mixed.
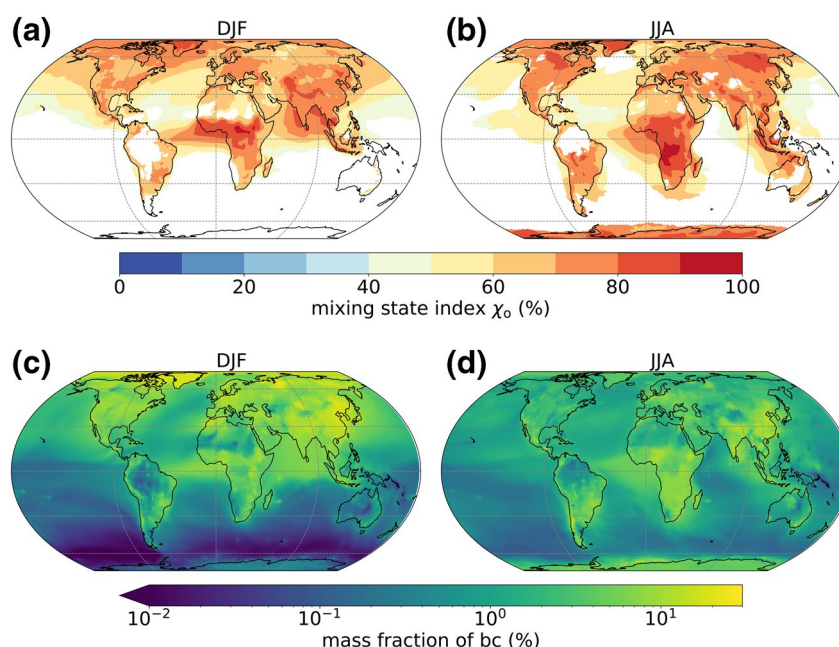
Figure 6. Average mixing state index χ_o (based on the mixing of optically absorbing and nonabsorbing species) and the mass fraction of black carbon in submicron aerosol. Mixing state index χ_o in seasons (a) DJF and (b) JJA. Mass fraction of black carbon in seasons (c) DJF and (d) JJA. The mask in (a) and (b) is based on the mass fraction in (c) and (d), respectively, where the areas with mass fraction of black carbon lower than 1% or higher than 99% are masked.

In JJA, χ_a varied within a range of 13%–94% and a mean of 69%, which is similar to DJF, but exhibited a different spatial pattern compared to DJF. The extent of regions with predominantly externally mixed aerosols ($\chi_a < 40\%$) decreased, while those with internally mixed aerosols ($\chi_a > 80\%$) increased. The lowest values of χ_a shifted from the North Atlantic Ocean to Western and Northern Africa. Significant differences between seasons existed in the Gulf of Guinea and the Arabian Sea. Compared to DJF, the mixing state index increased greatly over the Gulf of Guinea (12°S-12°N, 12°W-12°E) for JJA, from $\chi_a = 40\%$ to $\chi_a = 54\%$. The bulk composition of the aerosol also changed in this region between seasons. Combining this information, in DJF the submicron aerosol was mainly primary organic matter and dust, and it was comparatively externally mixed, while in JJA it was mainly primary organic matter, secondary organic matter, and sulfate, and was comparatively internally mixed. In contrast, χ_a over the Arabian Sea (0°N-25°N, 50°E-75°E) experienced a drastic decrease of χ_a from DJF to JJA, where a more internal mixture of sulfate and primary organic matter was replaced by a more external mixture dominated by sea salt, sulfate, and dust. Over the Southern Ocean (30°S-60°S), χ_a values increased during JJA to values of nearly 80%. The bulk composition of the aerosol, however, was mainly sea salt (over 80% mass fraction), with only small contributions of the other aerosol species.

Figures 6a and 6b show the global distribution of the aerosol mixing state index χ_o (based on the mixing of optically absorbing and nonabsorbing species) for DJF and JJA, and, for reference, Figures 6c and 6d depict the corresponding bulk mass fractions of BC. We masked out the areas where the bulk mass fraction of BC is lower than 1% to only evaluate χ_o where BC is present in nonnegligible amounts.

In DJF, χ_o ranged from 40% to 93%. This range is smaller than the range for χ_a although the global mean (67%) was close to that of χ_a. The high values of χ_o suggest that in most of the regions around the globe, BC tended to be internally mixed, and even more so in the continental regions, although a complete internal mixture was not reached anywhere. This means that BC is not evenly distributed over all particles, but it is also not exclusively confined to some particles. We can refer to the middle row in Figure 2 for an illustration of what the particle population might look like going from $\chi_o = 63\%$ to 91%. Note that BC-free particles may coexist with BC-containing particles under these conditions.
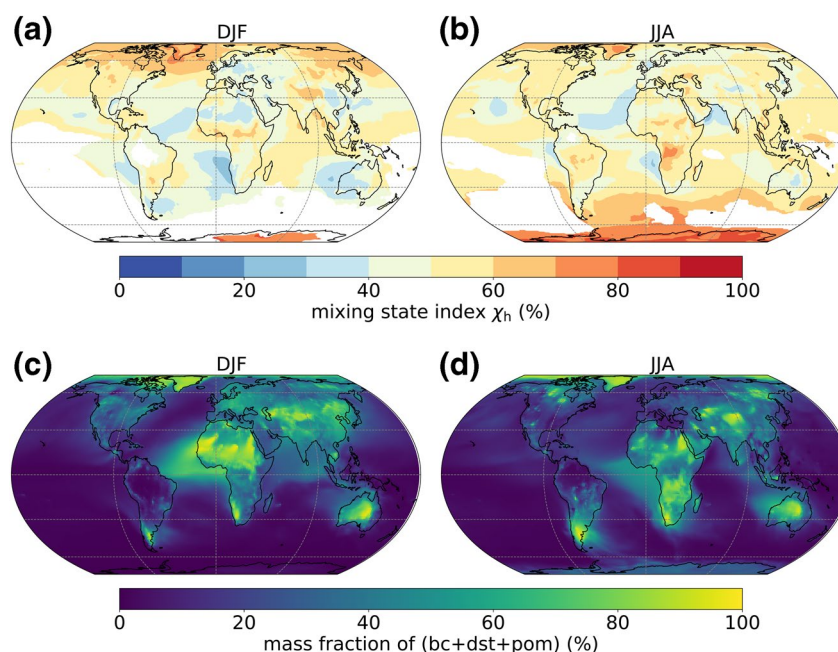
Figure 7. Average mixing state index χ_h (based on the mixing of hygroscopic and nonhygroscopic species) and the mass fraction of nonhygroscopic species (black carbon, dust, and primary organic matter) in submicron aerosol. Mixing state index χ_h in seasons (a) DJF and (b) JJA. Mass fraction of nonhygroscopic species in seasons (c) DJF and (d) JJA. The mask in (a) and (b) is based on the mass fraction in (c) and (d), respectively, where the areas with mass fraction of black carbon, dust, and primary organic matter lower than 2.5% or higher than 97.5% are masked.

A high degree of internal mixture (χ_o larger than 70%, meaning BC is rather spread out over the aerosol population) was predicted over the Arabian Sea, Bay of Bengal, India, and south-west of China. These are areas where the mass mixing ratios of BC were relatively high, and χ_a was also relatively high. It suggests that BC was internally mixed with other species, with the most abundant species being primary organic matter and sulfate. However, high values of χ_o coincide with low values of χ_a over the Gulf of Guinea. This can be attributed to the external mixture of the nonblack-carbon species (dust and SOA) at this location.

In JJA, χ_o had a range of 38%–94% and a global mean of 69%, overall similar to the DJF season. BC was more internally mixed in Southern Africa in JJA. In this area, it was mainly primary organic matter that formed an internal mixture with BC. Another pronounced change compared to DJF occurred in Eastern China, where the BC was internally mixed with an elevated level of primary organic matter, sulfate, and SOA. In general, areas with high mass fractions of BC were associated with higher values of χ_o, meaning that in polluted urban regions, BC existed in an internal mixture with other aerosol species. In contrast, areas with low BC mass fractions, for example over the oceans, were associated with lower values of χ_o (between an internal and external mixture).

Figure 7 shows the distributions of the aerosol mixing state index χ_h (based on the mixing of hygroscopic and nonhygroscopic species), and the corresponding mass fraction of nonhygroscopic species (BC, dust, and primary organic matter). Similar to Figure 5, we focus on areas where both surrogate species are present with a mass fraction threshold of 2.5%. The mixing state index χ_h in DJF ranged from 20% to 82%, with a global mean of 54%. The Bay of Bengal, south-west of China, and Mongolia were areas where hygroscopic and nonhygroscopic species existed in a more internal mixture (χ_h larger than 55%). Sulfate and primary organic matter were the most abundant species over these regions (Figure S2). In contrast, hygroscopic and nonhygroscopic species were more externally mixed over the North Atlantic Ocean (near the equator), Southern Africa, Australia, Eastern Europe, and East China. Here, the dominant aerosol species were mineral dust and sea salt for the ocean, and sulfate, primary organic carbon, as well SOA for the land. In JJA, χ_h varied between 26% and 87%, with a mean of 58%. Over Central Africa and Southern Africa, χ_h increased to higher values (greater than 55%) than the DJF season. Here, primary organic carbon was the dominant

nonhygroscopic species, while SOA and sulfate were both present as hygroscopic species. In contrast, χ_h over the Arabian Sea decreased from DJF to JJA, similar to χ_a. In this region, a predominantly sea salt/dust mixture that was more externally mixed replaced the more internally mixed primary organic carbon/sulfate mixture present in DJF.

The χ_h distribution maps can be related to the framework of CCN error quantification by Ching et al. (2017). Their Figure 6 established a relationship between χ_h and the error in CCN concentration prediction when a fully internal mixture was assumed ($\chi_h = 100\%$). In our case, three-quarters of the grid cells have seasonally averaged χ_h values between 20% and 63%, which corresponds to an error in CCN concentration between 50% and 100%. For 11% of the grid cells, χ_h is over 75% and the error is expected to be less than 10%. We do not have any grid cells with seasonally averaged χ_h values less than 20%, which would be associated with errors larger than 100%. However, instantaneous χ_h values smaller than 20% do occur and therefore the seasonally averaged error based on instantaneous χ_h values would be higher since the error in CCN concentration is a convex function of χ_h.

The mixing state indices varied considerably over the oceans. The outflow of biomass burning or dust emission regions in Africa were areas of comparatively external mixtures in terms of species abundance (quantified by χ_a), owing to the coexistence of mineral dust or biomass burning aerosol and sea salt aerosol. Judging from the maps for χ_o, plumes of BC originating from major biomass burning regions in Africa were initially comparatively internally mixed and then became more externally mixed as the plume was transported over the ocean. This makes sense considering that an increasingly external mixture should form as the carbonaceous plumes mix with sea salt aerosol as they are transported away from the coasts. Another finding is that BC is fairly (albeit not completely) internally mixed in highly polluted areas, which is consistent with prior work from modeling (Fierce et al., 2016) and measurements (Wang et al., 2010). These studies found that short aging time scales apply to BC aerosol in polluted regions. However, this may also be due to the crude resolution of horizontal grids in the global model. Very close to the source an external mixture may apply (Ching et al., 2019), but capturing this would require a high spatial resolution down to the kilometer scale.

5.2. Comparison to Observational Data

The comparison to observational data is still challenging at this point, as per-particle mass fractions are needed to determine the mixing state indices and they are difficult to estimate from observational data. While many data sets exist where single-particle data was collected in different environments (Ault et al., 2009; Gunsch et al., 2018; Murphy et al., 1998; Murphy, Thomson, et al., 1998; Qin et al., 2012), the data has rarely been analyzed in a way that yields per-particle mass fractions based on chemical species that are compatible with the species used in chemical transport models. Notable exceptions are the studies by Healy et al. (2014), Ye et al. (2018), and Ching et al. (2019) where single-particle mass spectrometers were used that produce estimates of per-particle mass fractions of sulfate, ammonium, nitrate, organics BC, which are species that chemical transport models track.

Here, we attempted to compare at least qualitatively to the work by Healy et al. (2014), Ye et al. (2018), and Ching et al. (2019) since they used mixing state indices comparable to our χ_a. Compared to the values measured during the MEGAPOLI campaign (with a mean of 59%), the global mixing state estimates obtained a higher mean value ($\chi_a = 79\%$) for DJF and the grid cell where Paris is located. Ye et al. (2018) found a very similar range of mixing state indices (30%–70%) for the Pittsburgh, PA, region, with values closer to 70% outside of the downtown area. They used measurements from August 2016 and January/February 2017 for this analysis. Our global estimates for DJF and JJA for this region are both 78%. Ching et al. (2019) calculated the mixing state index in Tokyo ranging from 2.6% to 81.9% and indicated that samples were more internally mixed in August 2012. Our global estimates for the nearest grid cell also showed a relative high value ($\chi_a = 67\%$) for the JJA season.

A direct comparison of our results and observations like these is generally challenging, since point measurements in an urban environment are not directly comparable to a global modeling grid cell. However, it is encouraging that our results are relatively close to the observations. Clearly, it would be beneficial to have more observations available for model-observation comparisons, especially from regions that are not
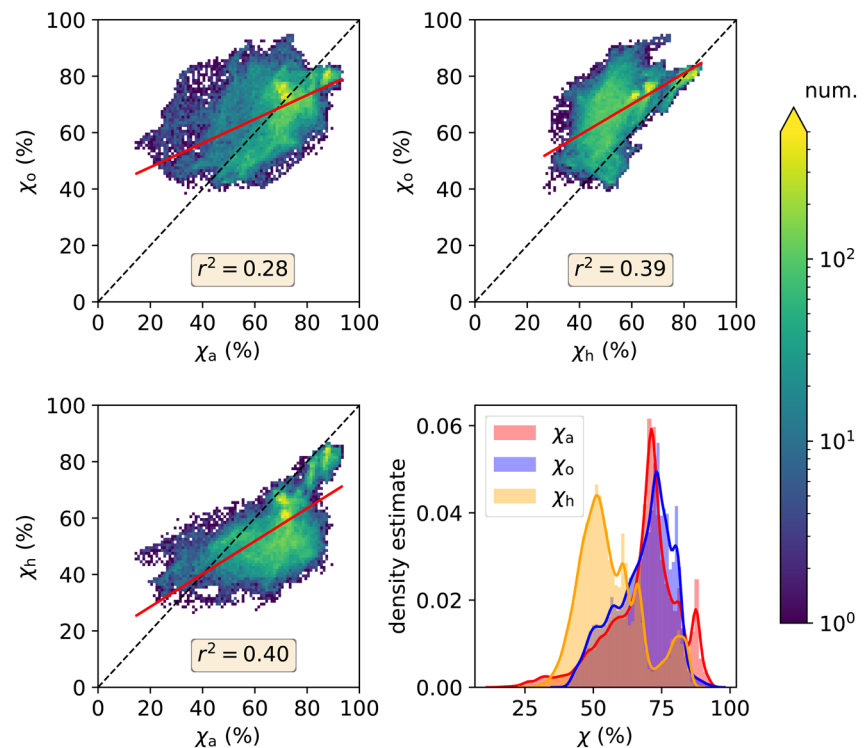
Figure 8. Relationship among aerosol mixing state indices χ_a, χ_o, and χ_h. Datapoints included here are seasonally averaged values from each surface-layer CESM gridpoint for DJF and JJA seasons.

dominated by local sources. In addition, difficulties in a comparison may arise when the species list used in the observations is not easily mapped onto model species, as is for example the case with methods using electron microscopy or X-ray spectroscopy where χ was calculated but based on elemental composition (Bondy et al., 2018; Fraund et al., 2017).

Our χ_o index can be related to observational studies that use SP2 measurements. Raatikainen et al. (2015) conducted measurements in the Finish Arctic during winter 2011–2012 and found that the number fraction of BC-containing particles was on average 24%, and that those BC-containing particles were thickly coated. While this study did not provide quantitative mixing state index calculations, it is an important finding that BC-containing particles (with various amounts of coatings) coexist with BC-free particles, and our result of $\chi_o \approx 72\%$ for this area is consistent with this. As we see from Figure 2, a χ_o value of 72% could correspond to a population where only some particles contain BC, but where BC-containing particles have a substantial nonBC fraction. Representing this mixing state is a challenge for many aerosol models used on the global scale that assume internal mixtures, but important for estimating aerosol optical properties accurately (Oshima et al., 2009).

5.3. Relationship Amongst the Three Mixing State Indices

Figure 8 shows the correlation (with p value < 0.001) among the aerosol mixing state indices χ_a, χ_o, and χ_h. The datapoints included in this figure are seasonally averaged values from each surface-layer CESM gridpoint for both DJF and JJA seasons. There is no strong relationship between χ_a and χ_o, which shows that each mixing state index captures different aspects of the population mixing state. As mentioned above, a population can appear as completely internally mixed when we only consider "black carbon" and "other species" as the two surrogate species. At the same time, the distribution of model species can vary within the "other species" class, resulting in a comparatively externally mixed population ($<50\%$) in terms of χ_a. In contrast, a high value for χ_a can correspond to a lower χ_o in cases where the average particle species diversity
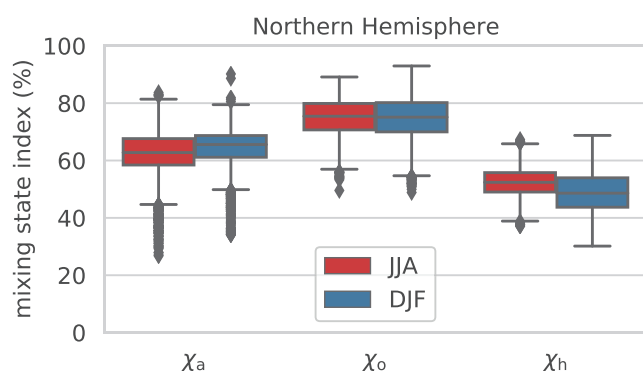
Figure 9. Aggregated statistics of aerosol mixing indices χ_a, χ_o, and χ_h in the areas of high anthropogenic pollution in the Northern in DJF and JJA.

D_α (in terms of optical properties) is low. This situation typically occurs when the particles are dominated by the "other species." A small fraction of "black carbon" exerts more influence on χ_o than on χ_a. Similar principles apply to the relationship between χ_h and χ_a, and χ_h and χ_o.

Overall, these results illustrate that, when determining mixing state from observed data using a certain set of species (determined by the instrumentation), it is difficult to infer the mixing state based on a different set of species. Mixing state indices for different purposes capture different information about the aerosol. Therefore, care has to be taken regarding the comparison of mixing states in different environments when the measurement techniques to estimate particle mass fractions are different.

Figure 9 summarizes the seasonally averaged aerosol mixing indices for grid points with high anthropogenic pollution in the Northern hemisphere. We defined these grid points as those where the mass mixing ration of the sum of BC, POM, SOA, and SO4 is higher than the 90% percentile for both DJF and JJA. Regardless of which mixing state index we consider, extremely external mixtures (χ lower than 20%) were not predicted for the seasonal averages. However, individual 3-h values reached as low as 3%, 11%, and 9% for χ_a, χ_o, and χ_h, respectively. Areas of high anthropogenic pollution in the northern hemisphere, such as central Europe and south-east Asia had larger χ_h-values during JJA which is consistent with faster chemical aging during northern hemisphere summer.

6. Conclusions

In this paper, we presented a framework for estimating submicron aerosol mixing state indices at a global scale. A machine learning model (an emulator) was trained on high-detail simulations, using the particle-resolved PartMC-MOSAIC, and then applied to the output from a coarser model (MAM4 within CESM) to enhance its information content.

We developed three emulators based on the XGBoost algorithm to determine the aerosol mixing state indices for submicron aerosol in terms of chemical species abundance (χ_a), the mixing of optically absorbing and nonabsorbing species (χ_o), and the mixing of hygroscopic and nonhygroscopic species (χ_h) using variables available in CESM. Based on error metrics for the testing data set, the predictions of mixing state metrics have an error of $\sim\pm10\%$ associated with them, and somewhat higher for χ_h. The emulators were applied to CESM simulations to produce global maps of aerosol mixing state indices for every 3 h timestamp. For this work, we focused on seasonal variation and spatial distribution of these aerosol mixing state indices.

The seasonally averaged indices varied spatially between 13% and 94% for χ_a, between 38% and 94% for χ_o, and 20% and 87% for χ_h, with global annual averages of 67%, 68%, and 56%, respectively. Hygroscopic and nonhygroscopic species were more externally mixed over the North Atlantic Ocean near the equator, off the coasts of Southern Africa, and Australia. An internal mixing state assumption for those areas could result in errors in CCN concentration of 50%–100%.

High values in one mixing state index can be consistent with low values in another index, depending on the grouping of species and their relative abundance. This indicates that the different indices capture different aspects about the mixing state of an aerosol. When comparing mixing state indices for different environments from observations, care needs to be taken that they use the same choice of species as the basis.

Comparing our results with available ambient observations of mixing state indices in the literature is still challenging, as estimates of per-particle mass fractions needed to determine the mixing state indices are difficult to obtain and long-term records are not yet available. Qualitative comparisons with studies by Healy et al. (2014), Ye et al. (2018), and Ching et al. (2019) show that our results are similar in magnitude to the observations.

Data Availability Statement

Scripts and instruction to create the PartMC-MOSIAC scenarios are available at https://doi.org/10.5281/zenodo.4033985 or https://github.com/zzheng93/code_scenario_generator. Notebooks and data to reproduce the emulator development, emulator application, and global mixing state indices analysis are available at https://doi.org/10.5281/zenodo.4033938 or https://github.com/zzheng93/code_global_ms. The temperature data from CESM CMIP6 output is available at https://esgf-node.llnl.gov/search/cmip6/.

References

Abelson, S., Collord, G., Ng, S. W. K., Weissbrod, O., Mendelson Cohen, N., Niemeyer, E., & Shlush, L. I. (2018). Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*, *559*(7714), 400–404. https://doi.org/10.1038/s41586-018-0317-6

Albani, S., Mahowald, N. M., Perry, A. T., Scanza, R. A., Zender, C. S., Heavens, N. G., & Otto-Bliesner, B. L. (2014). Improved dust representation in the community atmosphere model. *Journal of Advances in Modeling Earth Systems*, *6*(3), 541–570. https://doi.org/10.1002/2013MS000279

Asmi, A., Wiedensohler, A., Laj, P., Fjaeraa, A.-M., Sellegri, K., Birmili, W., & Kulmala, M. (2011). Number size distributions and seasonality of submicron particles in Europe 2008–2009. *Atmospheric Chemistry and Physics*, *11*(11), 5505–5538. https://doi.org/10.5194/acp-11-5505-2011

Ault, A. P., Moore, M. J., Furutani, H., & Prather, K. A. (2009). Impact of emissions from the Los Angeles port region on San Diego air quality during regional transport events. *Environmental Science & Technology*, *43*(10), 3500–3506. https://doi.org/10.1021/es8018918

Bogenschutz, P. A., Gettelman, A., Hannay, C., Larson, V. E., Neale, R. B., Craig, C., & Chen, C.-C. (2018). The path to CAM6: Coupled simulations with CAM5.4 and CAM5.5. *Geoscientific Model Development*, *11*(1), 235–255. https://doi.org/10.5194/gmd-11-235-2018

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., & Zender, C. S. (2013). Bounding the role of black carbon in the climate system: A scientific assessment. *Journal of Geophysical Research: Atmospheres*, *118*(11), 5380–5552. https://doi.org/10.1002/jgrd.50171

Bond, T. C., Streets, D. G., Yarber, K. F., Nelson, S. M., Woo, J.-H., & Klimont, Z. (2004). A technology-based global inventory of black and organic carbon emissions from combustion. *Journal of Geophysical Research*, *109*, D14. https://doi.org/10.1029/2003JD003697

Bondy, A. L., Bonanno, D., Moffet, R. C., Wang, B., Laskin, A., & Ault, A. P. (2018). The diverse chemical mixing state of aerosol particles in the southeastern United States. *Atmospheric Chemistry and Physics*, *18*(16), 12595–12612. https://doi.org/10.5194/acp-18-12595-2018

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). San Francisco, CA: ACM Press. https://doi.org/10.1145/2939672.2939785

Ching, J., Adachi, K., Zaizen, Y., Igarashi, Y., & Kajino, M. (2019). Aerosol mixing state revealed by transmission electron microscopy pertaining to cloud formation and human airway deposition. *npj Climate and Atmospheric Science*, *2*(1), 22. https://doi.org/10.1038/s41612-019-0081-9

Ching, J., Fast, J., West, M., & Riemer, N. (2017). Metrics to quantify the importance of mixing state for CCN activity. *Atmospheric Chemistry and Physics*, *17*(12), 7445–7458. https://doi.org/10.5194/acp-17-7445-2017

Ching, J., & Kajino, M. (2018). Aerosol mixing state matters for particles deposition in human respiratory system. *Scientific Reports*, *8*(1), 8864. https://doi.org/10.1038/s41598-018-27156-z

Ching, J., Riemer, N., & West, M. (2012). Impacts of black carbon mixing state on black carbon nucleation scavenging: Insights from a particle-resolved model. *Journal of Geophysical Research*, *117*, D23209. https://doi.org/10.1029/2012JD018269

Ching, J., West, M., & Riemer, N. (2018). Quantifying impacts of aerosol mixing state on nucleation-scavenging of black carbon aerosol particles. *Atmosphere*, *9*(1), 17. https://doi.org/10.3390/atmos9010017

Danabasoglu, G. (2019). *NCAR CESM2 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.7627

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., & Strand, W. G. (2020). The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. https://doi.org/10.1029/2019MS001916

Dickau, M., Olfert, J., Stettler, M. E. J., Boies, A., Momenimovahed, A., Thomson, K., & Johnson, M. (2016). Methodology for quantifying the volatile mixing state of an aerosol. *Aerosol Science & Technology*, *50*(8), 759–772. https://doi.org/10.1080/02786826.2016.1185509

Emmons, L. K., Schwantes, R. H., Orlando, J. J., Tyndall, G., Kinnison, D., Lamarque, J.-F., et al. (2020). The chemistry mechanism in the Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(4), e2019MS001882. https://doi.org/10.1029/2019MS001882

Fierce, L., Bond, T. C., Bauer, S. E., Mena, F., & Riemer, N. (2016). Black carbon absorption at the global scale is affected by particle-scale diversity in composition. *Nature Communications*, *7*(1), 12361. https://doi.org/10.1038/ncomms12361

Fierce, L., Riemer, N., & Bond, T. C. (2017). Toward reduced representation of mixing state for simulating aerosol effects on climate. *Bulletin of the American Meteorological Society*, *98*(5), 971–980. https://doi.org/10.1175/BAMS-D-16-0028.1

Fraund, M., Pham, D., Bonanno, D., Harder, T., Wang, B., Brito, J., & Moffet, R. (2017). Elemental mixing state of aerosol particles collected in Central Amazonia during GoAmazon2014/15. *Atmosphere*, *8*(12), 173. https://doi.org/10.3390/atmos8090173

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gantt, B., He, J., Zhang, X., Zhang, Y., & Nenes, A. (2014). Incorporation of advanced aerosol activation treatments into CESM/CAM5: Model evaluation and impacts on aerosol indirect effects. *Atmospheric Chemistry and Physics*, *14*(14), 7485–7497. https://doi.org/10.5194/acp-14-7485-2014

Giorio, C., Tapparo, A., Dall'Osto, M., Beddows, D. C. S., Esser-Gietl, J. K., Healy, R. M., & Harrison, R. M. (2015). Local and regional components of aerosol in a heavily trafficked street canyon in central London derived from PMF and cluster analysis of single-particle ATOFMS spectra. *Environmental Science & Technology*, *49*(6), 3330–3340. https://doi.org/10.1021/es506249z

Gunsch, M. J., May, N. W., Wen, M., Bottenus, C. L. H., Gardner, D. J., VanReken, T. M., & Pratt, K. A. (2018). Ubiquitous influence of wildfire emissions and secondary organic aerosol on summertime atmospheric aerosol in the forested Great Lakes region. *Atmospheric Chemistry and Physics*, *18*(5), 3701–3715. https://doi.org/10.5194/acp-18-3701-2018

He, J., & Zhang, Y. (2014). Improvement and further development in CESM/CAM5: Gas-phase chemistry and inorganic aerosol treatments. *Atmospheric Chemistry and Physics*, *14*(17), 9171–9200. https://doi.org/10.5194/acp-14-9171-2014

Healy, R. M., Riemer, N., Wenger, J. C., Murphy, M., West, M., Poulain, L., & Evans, G. J. (2014). Single particle diversity and mixing state measurements. *Atmospheric Chemistry and Physics*, *14*(12), 6289–6299. https://doi.org/10.5194/acp-14-6289-2014

Hodzic, A., Campuzano-Jost, P., Bian, H., Chin, M., Colarco, P. R., Day, D. A., & Jimenez, J. L. (2020). Characterization of organic aerosol across the global remote troposphere: A comparison of ATom measurements and global chemistry models. *Atmospheric Chemistry and Physics*, *20*(8), 4607–4635. https://doi.org/10.5194/acp-20-4607-2020

Holmgren, H., Sellegri, K., Hervo, M., Rose, C., Freney, E., Villani, P., & Laj, P. (2014). Hygroscopic properties and mixing state of aerosol measured at the high-altitude site Puy de Dôme (1465 m a.s.l.), France. *Atmospheric Chemistry and Physics*, *14*(18), 9537–9554. https://doi.org/10.5194/acp-14-9537-2014

Hommel, R., Timmreck, C., & Graf, H. F. (2011). The global middle-atmosphere aerosol model MAECHAM5-SAM2: Comparison with satellite and in-situ observations. *Geoscientific Model Development*, *4*(3), 809–834. https://doi.org/10.5194/gmd-4-809-2011

Hughes, M., Kodros, J., Pierce, J., West, M., & Riemer, N. (2018). Machine learning to predict the global distribution of aerosol mixing state metrics. *Atmosphere*, *9*(1), 15. https://doi.org/10.3390/atmos9010015

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., & Marshall, S. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*(9), 1339–1360. https://doi.org/10.1175/BAMS-D-12-00121.1

Kaiser, J. C., Hendricks, J., Righi, M., Riemer, N., Zaveri, R. A., Metzger, S., & Aquila, V. (2014). The MESSy aerosol submodel MADE3 (v2.0b): Description and a box model test. *Geoscientific Model Development*, *7*(3), 1137–1157. https://doi.org/10.5194/gmd-7-1137-2014

Knopf, D. A., Alpert, P. A., & Wang, B. (2018). The role of organic aerosol in atmospheric ice nucleation: A review. *ACS Earth Space Chemistry*, *2*(3), 168–202. https://doi.org/10.1021/acsearthspacechem.7b00120

Koch, D., Schulz, M., Kinne, S., McNaughton, C., Spackman, J. R., Balkanski, Y., & Zhao, Y. (2009). Evaluation of black carbon estimations in global aerosol models. *Atmospheric Chemistry and Physics*, *9*(22), 9001–9026. https://doi.org/10.5194/acp-9-9001-2009

Kulmala, M., Petäjä, T., Ehn, M., Thornton, J., Sipilä, M., Worsnop, D., & Kerminen, V.-M. (2014). Chemistry of atmospheric nucleation: On the recent advances on precursor characterization and atmospheric cluster composition in connection with atmospheric new particle formation. *Annual Review of Physical Chemistry*, *65*(1), 21–37. https://doi.org/10.1146/annurev-physchem-040412-110014

Lack, D. A., Tie, X. X., Bofinger, N. D., Wiegand, A. N., & Madronich, S. (2004). Seasonal variability of secondary organic aerosol: A global modeling study. *Journal of Geophysical Research*, *109*, D3. https://doi.org/10.1029/2003JD003418

Lana, A., Bell, T. G., Simó, R., Vallina, S. M., Ballabrera-Poy, J., Kettle, A. J., & Liss, P. S. (2011). An updated climatology of surface dimethylsulfide concentrations and emission fluxes in the global ocean. *Global Biogeochemical Cycles*, *25*(1), GB1004. https://doi.org/10.1029/2010GB003850

Lee, A. K., Rivellini, L.-H., Chen, C.-L., Liu, J., Price, D. J., Betha, R., & Cappa, C. D. (2019). Influences of primary emission and secondary coating formation on the particle diversity and mixing state of black carbon particles. *Environmental Science & Technology*, *53*(16), 9429–9438. https://doi.org/10.1021/acs.est.9b03064

Lesins, G., Chylek, P., & Lohmann, U. (2002). A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing. *Journal of Geophysical Research*, *107*, D10. https://doi.org/10.1029/2001JD000973

Liu, X., Easter, R. C., Ghan, S. J., Zaveri, R., Rasch, P., Shi, X., & Mitchell, D. (2012). Toward a minimal representation of aerosols in climate models: Description and evaluation in the Community Atmosphere Model CAM5. *Geoscientific Model Development*, *5*(3), 709–739. https://doi.org/10.5194/gmd-5-709-2012

Liu, X., Ma, P.-L., Wang, H., Tilmes, S., Singh, B., Easter, R. C., & Rasch, P. J. (2016). Description and evaluation of a new four-mode version of the Modal Aerosol Module (MAM4) within version 5.3 of the community atmosphere model. *Geoscientific Model Development*, *9*(2), 505–522. https://doi.org/10.5194/gmd-9-505-2016

Mahowald, N. M., Baker, A. R., Bergametti, G., Brooks, N., Duce, R. A., Jickells, T. D., & Tegen, I. (2005). Atmospheric global dust cycle and iron inputs to the ocean. *Global Biogeochemical Cycles*, *19*(4), GB4025. https://doi.org/10.1029/2004GB002402

Mott, A., Job, J., Vlimant, J.-R., Lidar, D., & Spiropulu, M. (2017). Solving a Higgs optimization problem with quantum annealing for machine learning. *Nature*, *550*(7676), 375–379. https://doi.org/10.1038/nature24047

Murphy, D. M., Anderson, J. R., Quinn, P. K., McInnes, L. M., Brechtel, F. J., Kreidenweis, S. M., & Buseck, P. R. (1998). Influence of sea-salt on aerosol radiative properties in the Southern Ocean marine boundary layer. *Nature*, *392*(6671), 62–65. https://doi.org/10.1038/32138

Murphy, D. M., Thomson, D. S., Middlebrook, A. M., & Schein, M. E. (1998). In situ single-particle characterization at Cape Grim. *Journal of Geophysical Research*, *103*(D13), 16485–16491. https://doi.org/10.1029/97JD03281

O'Brien, R. E., Wang, B., Laskin, A., Riemer, N., West, M., Zhang, Q., & Moffet, R. C. (2015). Chemical imaging of ambient aerosol particles: Observational constraints on mixing state parameterization. *Journal of Geophysical Research: Atmospheres*, *120*(18), 9591–9605. https://doi.org/10.1002/2015JD023480

Omori, Y., Tanimoto, H., Inomata, S., Ikeda, K., Iwata, T., Kameyama, S., & Furuya, K. (2017). Sea-to-air flux of dimethyl sulfide in the South and North Pacific Ocean as measured by proton transfer reaction-mass spectrometry coupled with the gradient flux technique. *Journal of Geophysical Research: Atmospheres*, *122*(13), 7216–7231. https://doi.org/10.1002/2017JD026527

Oshima, N., Koike, M., Zhang, Y., & Kondo, Y. (2009). Aging of black carbon in outflow from anthropogenic sources using a mixing state resolved model: 2. Aerosol optical properties and cloud condensation nuclei activities. *Journal of Geophysical Research*, *114*(D18), D18202. https://doi.org/10.1029/2008JD011681

Pierce, J. R., Croft, B., Kodros, J. K., D'Andrea, S. D., & Martin, R. V. (2015). The importance of interstitial particle scavenging by cloud droplets in shaping the remote aerosol size distribution and global aerosol-climate effects. *Atmospheric Chemistry and Physics*, *15*(11), 6147–6158. https://doi.org/10.5194/acp-15-6147-2015

Qin, X., Pratt, K. A., Shields, L. G., Toner, S. M., & Prather, K. A. (2012). Seasonal comparisons of single-particle chemical mixing state in Riverside, CA. *Atmospheric Environment*, *59*, 587–596. https://doi.org/10.1016/j.atmosenv.2012.05.032

Raatikainen, T., Brus, D., Hyvärinen, A.-P., Svensson, J., Asmi, E., & Lihavainen, H. (2015). Black carbon concentrations and mixing state in the Finnish Arctic. *Atmospheric Chemistry and Physics*, *15*(17), 10057–10070. https://doi.org/10.5194/acp-15-10057-2015

Ramanathan, V., & Carmichael, G. (2008). Global and regional climate changes due to black carbon. *Nature Geoscience*, *1*(4), 221–227. https://doi.org/10.1038/ngeo156

Riemer, N., Ault, A. P., West, M., Craig, R. L., & Curtis, J. H. (2019). Aerosol mixing state: Measurements, modeling, and impacts. *Reviews of Geophysics*, *57*(2), 187–249. https://doi.org/10.1029/2018RG000615

Riemer, N., & West, M. (2013). Quantifying aerosol mixing state with entropy and diversity measures. *Atmospheric Chemistry and Physics*, *13*(22), 11423–11439. https://doi.org/10.5194/acp-13-11423-2013

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*, *114*(D9), D09202. https://doi.org/10.1029/2008JD011073

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., & Marra, P. P. (2019). Decline of the North American avifauna. *Science*, *366*(6461), 120–124. https://doi.org/10.1126/science.aaw1313

Schell, B., Ackermann, I. J., Hass, H., Binkowski, F. S., & Ebel, A. (2001). Modeling the formation of secondary organic aerosol within a comprehensive air quality model system. *Journal of Geophysical Research*, *106*(D22), 28275–28293. https://doi.org/10.1029/2001JD000384

Schwantes, R. H., Emmons, L. K., Orlando, J. J., Barth, M. C., Tyndall, G. S., Hall, S. R., & Bui, T. P. V. (2020). Comprehensive isoprene and terpene gas-phase chemistry improves simulated surface ozone in the southeastern US. *Atmospheric Chemistry and Physics*, *20*(6), 3739–3776. https://doi.org/10.5194/acp-20-3739-2020

Sofiev, M., Soares, J., Prank, M., de Leeuw, G., & Kukkonen, J. (2011). A regional-to-global model of emission and transport of sea salt particles in the atmosphere. *Journal of Geophysical Research*, *116*, D21. https://doi.org/10.1029/2010JD014713

Tanaka, T. Y., & Chiba, M. (2006). A numerical study of the contributions of dust source regions to the global dust budget. *Global and Planetary Change*, *52*(1), 88–104. https://doi.org/10.1016/j.gloplacha.2006.02.002

Tian, J., Riemer, N., West, M., Pfaffenberger, L., Schlager, H., & Petzold, A. (2014). Modeling the evolution of aerosol particles in a ship plume using PartMC-MOSAIC. *Atmospheric Chemistry and Physics*, *14*(11), 5327–5347. https://doi.org/10.5194/acp-14-5327-2014

Tilmes, S., Lamarque, J.-F., Emmons, L. K., Kinnison, D. E., Ma, P.-L., Liu, X., & Val Martin, M. (2015). Description and evaluation of tropospheric chemistry and aerosols in the Community Earth System Model (CESM1.2). *Geoscientific Model Development*, *8*(5), 1395–1426. https://doi.org/10.5194/gmd-8-1395-2015

Wang, J., Cubison, M. J., Aiken, A. C., Jimenez, J. L., & Collins, D. R. (2010). The importance of aerosol mixing state and size-resolved composition on CCN concentration and the variation of the importance with atmospheric aging of aerosols. *Atmospheric Chemistry and Physics*, *10*(15), 7267–7283. https://doi.org/10.5194/acp-10-7267-2010

Wang, Y. H., Liu, Z. R., Zhang, J. K., Hu, B., Ji, D. S., Yu, Y. C., & Wang, Y. S. (2015). Aerosol physicochemical properties and implications for visibility during an intense haze episode during winter in Beijing. *Atmospheric Chemistry and Physics*, *15*(6), 3205–3215. https://doi.org/10.5194/acp-15-3205-2015

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Winkler, P. (1973). The growth of atmospheric aerosol particles as a function of the relative humidity—II. An improved concept of mixed nuclei. *Journal of Aerosol Science*, *4*(5), 373–387. https://doi.org/10.1016/0021-8502(73)90027-X

Xu, L., Cameron-Smith, P., Russell, L. M., Ghan, S. J., Liu, Y., Elliott, S., & Manizza, M. (2016). DMS role in ENSO cycle in the tropics. *Journal of Geophysical Research: Atmospheres*, *121*(22), 13537–13558. https://doi.org/10.1002/2016JD025333

Ye, Q., Gu, P., Li, H. Z., Robinson, E. S., Lipsky, E., Kaltsonoudis, C., & Donahue, N. M. (2018). Spatial variability of sources and mixing state of atmospheric particles in a metropolitan area. *Environmental Science & Technology*, *52*(12), 6807–6815. https://doi.org/10.1021/acs.est.8b01011

Yu, C., Liu, D., Broda, K., Joshi, R., Olfert, J., Sun, Y., & Allan, J. D. (2020). Characterising mass-resolved mixing state of black carbon in Beijing using a morphology-independent measurement method. *Atmospheric Chemistry and Physics*, *20*(6), 3645–3661. https://doi.org/10.5194/acp-20-3645-2020

Yu, F., & Luo, G. (2009). Simulation of particle size distribution with a global aerosol model: Contribution of nucleation to aerosol and CCN number concentrations. *Atmospheric Chemistry and Physics*, *9*(20), 7691–7710. https://doi.org/10.5194/acp-9-7691-2009

Zaveri, R. A., Easter, R. C., Fast, J. D., & Peters, L. K. (2008). Model for simulating aerosol interactions and chemistry (MOSAIC). *Journal of Geophysical Research*, *113*, D13. https://doi.org/10.1029/2007JD008782

Zaveri, R. A., Easter, R. C., & Peters, L. K. (2005). A computationally efficient multicomponent equilibrium solver for aerosols (MESA). *Journal of Geophysical Research*, *110*(D24), D24203. https://doi.org/10.1029/2004JD005618

Zaveri, R. A., Easter, R. C., & Wexler, A. S. (2005). A new method for multicomponent activity coefficients of electrolytes in aqueous atmospheric aerosols. *Journal of Geophysical Research*, *110*(D2), D02201. https://doi.org/10.1029/2004JD004681

Zaveri, R. A., & Peters, L. K. (1999). A new lumped structure photochemical mechanism for large-scale applications. *Journal of Geophysical Research*, *104*(D23), 30387–30415. https://doi.org/10.1029/1999JD900876