atmosphere

MDPI

Article

Machine Learning to Predict the Global Distribution of Aerosol Mixing State Metrics

Michael Hughes [1], John K. Kodros [2], Jeffrey R. Pierce [2], Matthew West [3] and Nicole Riemer [1,*]

[1] Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; hughes18@illinois.edu

[2] Department of Atmospheric Sciences, Colorado State University, Fort Collins, CO 80523, USA; jkodros@atmos.colostate.edu (J.K.K.); jeffrey.pierce@colostate.edu (J.R.P.)

[3] Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; mwest@illinois.edu

* Correspondence: nriemer@illinois.edu

Received: 20 November 2017; Accepted: 5 January 2018; Published: 9 January 2018

Abstract: Atmospheric aerosols are evolving mixtures of chemical species. In global climate models (GCMs), this "aerosol mixing state" is represented in a highly simplified manner. This can introduce errors in the estimates of climate-relevant aerosol properties, such as the concentration of cloud condensation nuclei. The goal for this study is to determine a global spatial distribution of aerosol mixing state with respect to hygroscopicity, as quantified by the mixing state metric χ. In this way, areas can be identified where the external or internal mixture assumption is more appropriate. We used the output of a large ensemble of particle-resolved box model simulations in conjunction 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 GCMs, enabling us to create a global map of χ based on GCM data. We found that χ varied between 20% and nearly 100%, and we quantified how this depended on particle diameter, location, and time of the year. This framework demonstrates how machine learning can be applied to bridge the gap between detailed process modeling and a large-scale climate model.

Keywords: aerosol modeling; mixing state; machine learning

1. Introduction

Field measurements show that individual aerosol particles are a complex mixture of a wide variety of species, such as soluble inorganic salts and acids, insoluble crustal materials, trace metals, and carbonaceous materials [1,2]. To characterize this mixture, the term "aerosol mixing state" is frequently used. This, in general, comprises both the distribution of chemical compounds across the aerosol population ("population mixing state") and the distribution of chemical compounds within and on the surface of each particle ("morphological mixing state").

Both the population mixing state and the morphological mixing state are of importance for aerosol impacts, including chemical reactivity, cloud condensation nuclei (CCN) activity, and aerosol optical properties [3]. However, the morphological mixing state is beyond the scope of this study. We will focus here exclusively on the population mixing state, and refer to it for brevity as "mixing state". In this context, the terms "internal" and "external" mixture are frequently used. An external mixture consists of particles that each contain only one species, which may be different for different particles. In contrast, an internal mixture describes a particle population where different species are present within one particle. If all particles consist of the same species mixture, and the relative abundances are identical, the term "fully internal mixture" is commonly used. Considering that aerosol populations contain particles of many different sizes, we can define these terms for the entire populations (comprising all

particle sizes) or for individual size ranges. An aerosol population might be approximately internally mixed for a certain size range, but the internal mixture assumption might not be fulfilled if a large size range is considered. While mixing state can impact both CCN properties and optical properties, here we target CCN properties, and interprete aerosol "species" in terms of hygroscopicity.

An example of an external mixture is shown in Figure 1a, which represents a particle population consisting of six particles, with the blue and the red color symbolizing two different aerosol species with different hygroscopicities. A fully internal mixture is shown in Figure 1d. In reality, aerosol populations assume mixing states that are neither fully externally nor internally mixed, as depicted by Figure 1b,c. Note that each of the four populations in Figure 1 contains the same total amounts of the two species, but the species distribution amongst the particles differs.

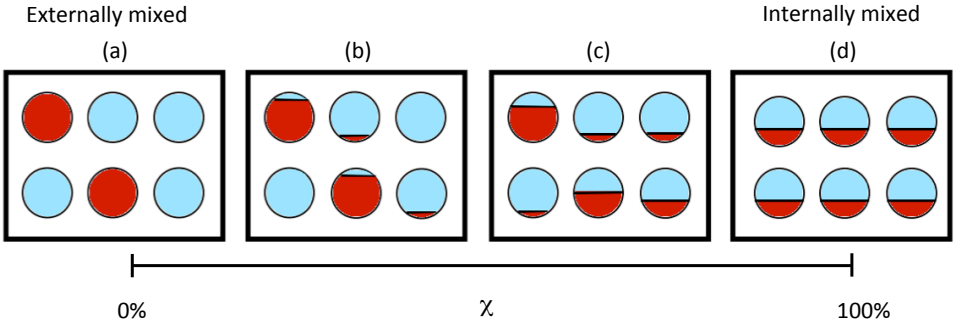

Figure 1. Schematic of aerosol mixing states for four different aerosol populations that have the same bulk composition. The blue and red color represent aerosol species with different hygroscopicity: (**a**) fully external mixture; (**b**,**c**) intermediate mixing states; and (**d**) internal mixture.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

Aerosol mixing state is challenging to represent in atmospheric aerosol models. The most rigorous approach is the particle-resolved approach by Riemer et al. [4], which explicitly resolves population mixing state. However, this method is too computationally demanding for routine use in spatially-resolved regional or global chemical transport models. Instead of resolving the full aerosol mixing state, regional and global models therefore use distribution-based methods, commonly known as modal and sectional models [5–7]. An inherent assumptions of these methods is that within one mode or within one size section, the aerosol particles are assumed to be internally mixed. This assumption can lead to misprediction in climate-relevant aerosol properties such as CCN concentrations and optical properties [8–12].

To illustrate this concept, Figure 2 shows the global distribution of the fraction of hygroscopic species (sulfate, ammonium, sea-salt, and aged organics) as simulated by GEOS-Chem-TOMAS for the month of January 2010 for particles of ~358 nm. For areas where this fraction is close to 100% (oceans) or close to 0% (parts of the Saharan desert), the aerosol consists essentially of only hygroscopic or only non-hygroscopic species, respectively, so mixing state is not an issue in these areas. However, there are many regions such as the continental US or Europe where the fraction is between the two extremes. For these regions, the question is, given the local conditions, what degree of internal/external mixing is most likely? Our approach seeks to answer this question for different particle sizes, different geographic locations, and different seasons.

To quantify the degree of internal/external mixture Riemer and West [13] introduced the mixing state index χ. This is a scalar quantity that varies between 0% for completely external mixtures and 100% for completely internal mixtures, as indicated by Figure 1. It can be calculated from per-particle species mass fractions (see Section 2.1), which requires either simulations with computationally expensive, high-detail aerosol models [13,14] or observations with a sophisticated suite of instruments [15,16].

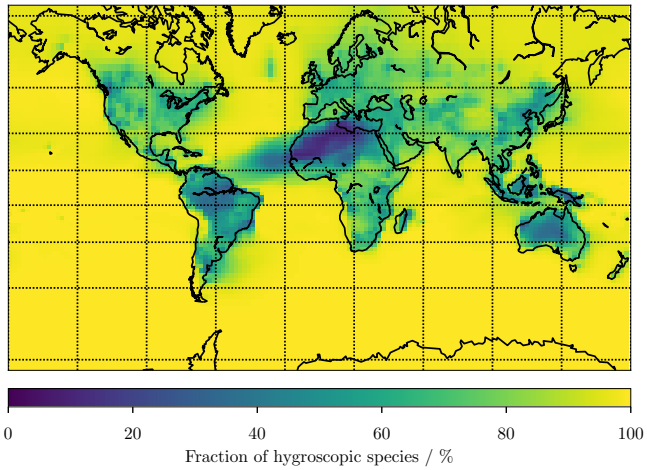

Figure 2. Global distribution of fraction of hygroscopic species as simulated by GEOS-Chem-TOMAS for the month of January for particles of ∼358 nm.

Ching et al. [14] quantified the relationship of mixing state index χ and the error in CCN concentrations when neglecting mixing state information by assuming a fully internal mixture. The study shows that for more externally mixed populations (χ below 20%) neglecting mixing state leads to errors up to 150%, whereas for populations with χ larger than 75%, the error vanishes (Figure 3). To establish this relationship, Ching et al. [14] used particle-resolved simulations from a 0-D box model scenario library that represented a suite of idealized urban plume scenarios. Thus far, no studies have calculated spatial distributions of the mixing state parameter. This, however, is important for understanding where global models may need to take mixing state into account.

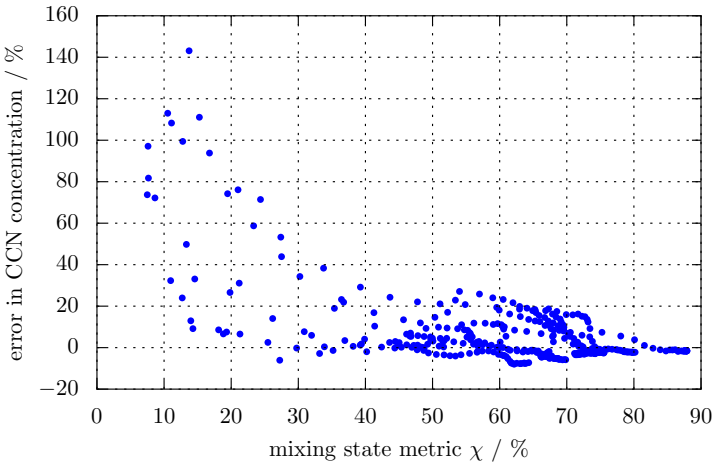

Figure 3. Relative error in CCN concentration when neglecting aerosol mixing state as a function of aerosol mixing state index χ. Each dot represents an aerosol population from Ching et al. [14]. CCN concentration was evaluated at a supersaturation of 0.6%.

The goal of this study is therefore to produce the first global distribution of mixing state parameter χ. This will allow us to map out areas on the globe where low χ values can be expected—these are the areas where we expect large errors in CCN prediction when using a simplified aerosol model that does not or not fully resolve aerosol mixing state. Conversely, it is informative to delineate areas where the mixing state approaches an internal mixture, as for these areas assuming an internal mixture would be appropriate for CCN predictions.

As mentioned before, it is currently not feasible to directly run a particle-resolved aerosol model on a global scale, which would be needed to create a global map of χ directly. We therefore propose an

approach that combines particle-resolved modeling and output from a global chemical transport model with machine learning techniques, as outlined in Figure 4. This involves the construction of a scenario library of particle-resolved simulations using the PartMC-MOSAIC, which cover a wide range of conditions that are expected to be encountered in different environments around the globe. This dataset is then used to train a model of χ using machine learning techniques. Importantly, the features of this model are dictated by the list of variables that are known to the global scale model, in our case GEOS-Chem-TOMAS [17,18].

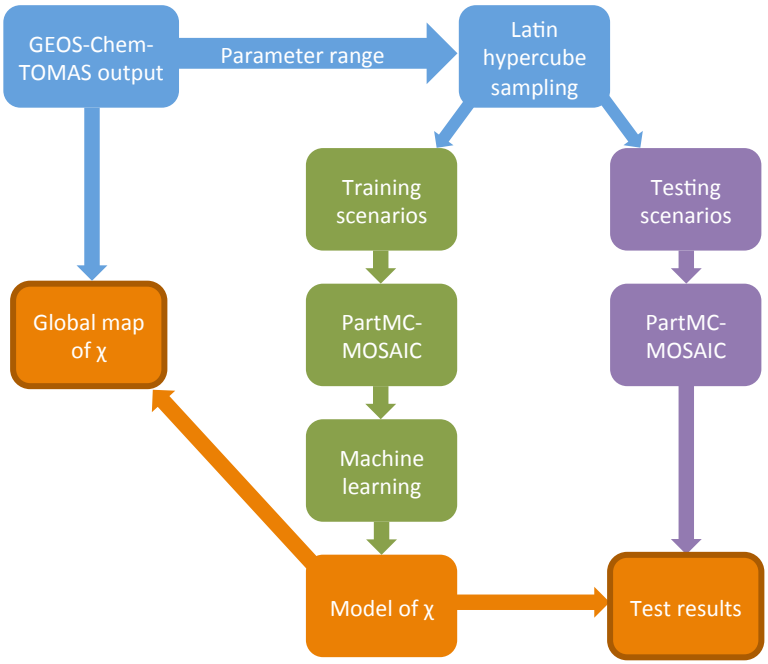

Figure 4. Schematic of the learning architecture used to train, test, and use the machine-learning model.

Many examples exist in the aerosol modeling literature where parameters for coarser models were derived on the basis of box model simulations that capture certain microphysical processes in detail [19]. However, the choice of the explanatory variables (features) and the fitting of the coarse model were typically done "by hand". This approach works well if the relevant parameter space is low-dimensional so that a few features can be identified that govern a certain process. In our case, there are many relevant variables that could potentially influence χ, hence machine learning methods represent an appropriate tool.

The remainder of the paper is structured as follows: Section 2 describes the tools and methods that are used this study, including the mixing state metric χ, the particle-resolved aerosol model PartMC-MOSAIC, the dataset from the global model GEOS-Chem-TOMAS, the simulations that yield the training and testing dataset, and the machine learning methods. Section 3 presents the global maps of mixing state parameter χ as obtained from the machine learning procedure. Section 4 concludes our results and provides a perspective for future work.

2. Methods

2.1. Mixing State Metric χ

We quantified aerosol mixing state with the framework discribed in Riemer and West [13], specifically using the mixing state metric χ. This was inspired by diversity metrics used in other disciplines such as ecology [20], economics [21], neuroscience [22], and genetics [23].

Given a population of N aerosol particles, each consisting of some amounts of A distinct aerosol species, the mixing state metrics can be determined if the masses of species a in particle i are known, denoted by μ_i^a, for $i = 1, \ldots, N$, and $a = 1, \ldots, A$. From this quantity, all other related quantities can be calculated, as described by Riemer and West [13] and here listed in Table 1. The diversity metrics can then be constructed as summarized in Table 2.

Table 1. Aerosol mass and mass fraction definitions and notation, used to construct the diversity metrics shown in Table 2. The number of particles in the population is N, and the number of species is A. This table is taken from Riemer and West [13].

Quantity	Meaning
μ_i^a	Mass of species a in particle i
$\mu_i = \sum_{a=1}^{A} \mu_i^a$	Total mass of particle i
$\mu^a = \sum_{i=1}^{N} \mu_i^a$	Total mass of species a in population
$\mu = \sum_{i=1}^{N} \mu_i$	Total mass of population
$p_i^a = \frac{\mu_i^a}{\mu_i}$	Mass fraction of species a in particle i
$p_i = \frac{\mu_i}{\mu}$	Mass fraction of particle i in population
$p^a = \frac{\mu^a}{\mu}$	Mass fraction of species a in population

Table 2. Definitions of aerosol mixing entropies, particle diversities, and mixing state index. In these definitions, we take $0 \ln 0 = 0$ and $0^0 = 1$. This table is taken from Riemer and West [13].

Quantity	Name	Units	Range	Meaning
$H_i = \sum_{a=1}^{A} -p_i^a \ln p_i^a$	Mixing entropy of particle i	—	0 to $\ln A$	Shannon entropy of species distribution within particle i
$H_\alpha = \sum_{i=1}^{N} p_i H_i$	Average particle mixing entropy	—	0 to $\ln A$	average Shannon entropy per particle
$H_\gamma = \sum_{a=1}^{A} -p^a \ln p^a$	Population bulk mixing entropy	—	0 to $\ln A$	Shannon entropy of species distribution within population
$D_i = e^{H_i} = \prod_{a=1}^{A} (p_i^a)^{-p_i^a}$	Particle diversity of particle i	Effective species	1 to A	Effective number of species in particle i
$D_\alpha = e^{H_\alpha} = \prod_{i=1}^{N} (D_i)^{p_i}$	Average particle (alpha) species diversity	Effective species	1 to A	Average effective number of species in each particle
$D_\gamma = e^{H_\gamma} = \prod_{a=1}^{A} (p^a)^{-p^a}$	Bulk population (gamma) species diversity	Effective species	1 to A	Effective number of species in the bulk
$\chi = \frac{D_\alpha - 1}{D_\gamma - 1}$	Mixing state index	—	0% to 100%	Degree to which population is externally mixed ($\chi = 0\%$) versus internally mixed ($\chi = 100\%$)

Based on the per-particle mass fractions, the particle diversity D_i can be calculated, which can be interpreted as the number of "effective species" of particle i. For a particle consisting of A species, the particle diversity D_i can be maximally A, which occurs when all A species are present in equal mass fractions. From the D_i values of all particles, we can determine the population-level quantities

D_α and D_γ, with D_α being the average effective number of species in each particle, and D_γ being the effective number of species in the bulk. The mixing state index χ is defined as

$$\chi = \frac{D_\alpha - 1}{D_\gamma - 1}. \quad (1)$$

The mixing state index χ varies from 0% (a fully externally mixed population) to 100% (a fully internally mixed population). Since χ has the intuitive interpretation of the "degree of internal mixing", it can be used as a metric for error quantification, i.e., to determine the magnitude of error that is introduced in estimating aerosol impacts when neglecting mixing state information. This was shown by Ching et al. [14] for the example of CCN concentration, as illustrated in Figure 3.

The definition of "species" for calculating the mass fractions depends on the application. It can refer to individual chemical species, as in the studies by Riemer and West [13], Healy et al. [15], O'Brien et al. [16], Giorio et al. [24], and Fraund et al. [25]. Alternatively, it can refer to species groups, as in Dickau et al. [26] who quantified mixing state with respect to volatile and non-volatile components. Since we are concerned with CCN properties in this paper, we will group the chemical model species according to hygroscopicity, defining two species groups. Black carbon (BC), primary organic aerosol (POA), and freshly emitted mineral dust are combined into one surrogate species, since their hygroscopicities are very low. All other model species (inorganic and secondary organic aerosol species) are combined into a second surrogate species. The mixing state index χ is calculated from these two surrogate species. Note that calculating χ based on the two surrogate species does not bias the value of χ in a systematic way compared to the value based on the individual chemical species. A χ value close to 0% can be interpreted as the hygroscopic and non-hygroscopic species existing in different particles, whereas a χ value close to 100% would correspond to an aerosol population where all particles contain the same amount of hygroscopic and non-hygroscopic species.

2.2. Particle-Resolved Aerosol Modeling

A detailed model description of stochastic particle-resolved aerosol model PartMC-MOSAIC is provided by Riemer et al. [4]. In summary, PartMC (Particle-resolved Monte Carlo) is a zero-dimensional aerosol model, which explicitly tracks the composition of many individual particles within a well-mixed computational volume. This computational volume is assumed to be representative for a much larger air parcel within the planetary boundary layer. The processes of emission, dilution with the background, and Brownian coagulation are simulated with a stochastic Monte Carlo approach. To improve efficiency of the method, we use weighted particles in the sense of DeVille et al. [27] and efficient stochastic sampling methods [28].

PartMC is coupled with the aerosol chemistry model MOSAIC (Model for Simulating Aerosol Interactions and Chemistry) [29]. This includes the gas phase photochemical mechanism CBM-Z [30], the Multicomponent Taylor Expansion Method (MTEM) for estimating activity coefficients of electrolytes and ions in aqueous solutions [31], the multi-component equilibrium solver for aerosols (MESA) for solid–liquid partitioning within particles [32] and the adaptive step time-split Euler method (ASTEM) for dynamic gas–particle partitioning over the size- and composition-resolved aerosol [29]. To simulate secondary organic aerosol (SOA) the SORGAM scheme is used [33]. The CBM-Z gas phase mechanism includes 77 gas species. MOSAIC treats key aerosol species including sulfate (SO_4), nitrate (NO_3), ammonium (NH_4), chloride (Cl), carbonate (CO_3), methanesulfonic acid (MSA), sodium (Na), calcium (Ca), other inorganic mass (OIN), BC, POA, and SOA. The model species OIN represents species such as SiO_2, metal oxides, and other unmeasured or unknown inorganic species. Our SOA model species include reaction products of aromatic precursors, higher alkenes, α-pinene and limonene. In this study, PartMC includes condensation/evaporation of vapors to/from particles and coagulation between particles. It does not included nucleation in this study, and the limitations on our results will be discussed throughout.

PartMC-MOSAIC has been used in the past for process studies of mixing state impacts on aerosol properties in various environments. For example, Tian et al. [34] investigated the aging of aerosol particles in a ship plume. Ching et al. [12] quantified the response of cloud droplet number concentration to changes in emissions of black-carbon-containing particles, and Mena et al. [35] carried out plume-exit modeling to determine cloud condensation nuclei activity of aerosols from residential biofuel combustion.

2.3. GEOS-Chem-TOMAS Dataset

To provide initial concentrations of gas-phase and size-resolved aerosol-phase species in a large-scale global model, we use the Goddard Earth Observing System chemical-transport model, GEOS-Chem, version 10.01 [36] (http://acmg.seas.harvard.edu/geos/) coupled with the TwO Moment Aerosol Sectional (TOMAS) microphysics scheme [17]. We simulated the year 2010 with re-analysis meteorology fields from GEOS5 (http://gmao.gsfc.nasa.gov). Simulations included a horizontal resolution of $2^\circ \times 2.5^\circ$ and 47 vertical layers. GEOS-Chem includes tracers for 52 gas-phase species. Standard emission setup is described in the study by Kodros et al. [18]. We used the 15-bin version of TOMAS, with size sections ranging from approximately 3 nm to 10 µm. TOMAS includes tracers for aerosol number concentration, sulfate, organic aerosol, black carbon, sea salt, and dust. Nucleation in the simulations follows a ternary nucleation scheme involving water, sulfuric acid, and ammonia following the parameterization of Napari et al. [37], scaled with a global tuning factor of 10–5 [38,39]. When ammonia mixing ratios are less than 1 pptv, the model defaults to a binary nucleation scheme (sulfuric acid and water) [40]. Detailed descriptions of aerosol microphysics included in TOMAS can be found in Adams and Seinfeld [17], Lee et al. [41], and Lee and Adams [42]. GEOS-Chem-TOMAS has been evaluated against observed aerosol size distributions [43,44].

2.4. Design of the Training and the Testing Scenarios

At the core of the machine learning framework is the design of a scenario library of particle-resolved simulations to create a large number of aerosol populations with different compositions and different mixing states. Scenario libraries that we developed in previous work [10,12,45] focused on urban environments, and in particular on the aging process of carbonaceous aerosol by coagulation and condensation of secondary aerosol. Here, we expanded the list of aerosol types by including sea salt aerosol and dust emissions.

We did not include the process of particle nucleation in this set of training simulations because there are still significant uncertainties about the treatment of particle-level post-nucleation growth mechanisms [46]. The lack of nucleation in our training library can be expected to introduce errors into our global mixing state predictions in the smaller size bins where particles may be influenced by nucleation and growth. In particular, we expect that true χ values in the Aitken and accumulation modes will generally be lower than our predicted values in areas with pre-existing non-hygroscopic particles (e.g., from combustion) where significant nucleation occurs because freshly nucleated particles will then create a more-externally mixed population.

All scenarios used a simulation time of 24 h, starting at 6:00 a.m. local time, with output being saved every 10 min. We used 10,000 computational particles for each simulation. The initial conditions for aerosol and gas phase were the same for all scenarios and are identical to Zaveri et al. [8]. Specifically, the aerosol initial condition consisted of Aitken and Accumulation mode with internally mixed ammonium sulfate, secondary organic aerosol, and trace amounts of black carbon, as listed in Table 3. Although the initial conditions were fixed in these scenarios, these particles generally evolved substantially over the course of the simulations. However, we cannot rule out that this choice influenced our results, and we will address this in future work by introducing more variability to the design of the initial condition.

Table 3. Number concentration, N_a, of the initial aerosol population. The aerosol size distributions are assumed to be lognormal and defined by the geometric mean diameter, D_g, and the geometric standard deviation, σ_g.

Initial/Background	N_a/cm^{-3}	D_g/μm	σ_g	Composition by Mass
Aitken mode	1800	0.02	1.45	49.64% $(NH_4)_2SO_4$ + 49.64% SOA + 0.72% BC
Accumulation mode	1500	0.116	1.65	49.64% $(NH_4)_2SO_4$ + 49.64% SOA + 0.72% BC

Twenty-five input parameters were varied between scenarios to represent a range of environmental conditions with different levels of gas phase emissions and emissions of primary aerosol particles to allow for large variations in the mixing state evolution. Latin hypercube sampling [47] was used to provide an efficient sampling across this high-dimensional space. The details of our setup are listed in Table 4. The input parameter space was sampled so that the resulting distributions of simulated variables, such as gas phase and bulk aerosol concentrations, were similar to that of the corresponding distribution in the output data of GEOS-Chem-TOMAS. The distributions need not be identical, but they must be similar enough that the model that is trained from the PartMC library is not required to extrapolate far outside the parameter range on which it was trained.

Table 4. List of input parameters and their sampling ranges and procedures to construct the scenario library. See the main text for details.

	Range	Sampling Method
Environmental Variable		
RH	10–100%	uniform within specified ranges (1)
Latitude	70° S–70° N	uniform
Day of Year	1–365	uniform
Temperature	based on latitude and day of year (2)	uniform
Dilution rate	1.5×10^{-5} s^{-1}	constant
Mixing height	400 m	constant
Gas phase emissions		
SO_2, NO_x, NH_3, VOC	0–100% of emissions in Riemer et al. [4]	non-uniform (3)
Carbonaceous Aerosol Emissions (one mode) (4)		
D_g	25–250 nm	uniform
σ_g	1.4–2.5	uniform
BC/OC mass ratio	0–100%	non-uniform (3)
E_a	0–1.6×10^7 m^{-2} s^{-1}	non-uniform (3)
Sea Salt Emissions (two modes) (5)		
$D_{g,1}$	180–720 nm	uniform
$\sigma_{g,1}$	1.4–2.5	uniform
$E_{a,1}$	0–1.69×10^5 m^{-2} s^{-1}	non-uniform (3)
$D_{g,2}$	1–6 μm	uniform
$\sigma_{g,2}$	1.4–2.5	uniform
$E_{a,2}$	0–2380 m^{-2} s^{-1}	non-uniform (3)
OC fraction	0–20%	uniform
Dust Emissions (two modes) (6)		
$D_{g,1}$	80–320 nm	uniform
$\sigma_{g,1}$	1.4–2.5	uniform
$E_{a,1}$	0–586,000 m^{-2} s^{-1}	non-uniform (3)
$D_{g,2}$	1–6 μm	uniform
$\sigma_{g,2}$	1.4–2.5	uniform
$E_{a,2}$	0–2380 m^{-2} s^{-1}	non-uniform (3)
hygroscopicity (κ)	0.001–0.031	uniform

Specific information about the individual variables listed in Table 4 is as follows. (1) Relative humidity was sampled from a range of 10% to 100%, using two uniform distributions: The range 10% to 60% comprised 25% of the sampled RH values, while the range 60% to 100% made up the remaining 75%. The 60%-to-100% range was sampled more heavily because the global average RH is 73% [48]. (2) We obtained monthly global temperatures from the NCEP/NCAR reanalysis data [48] for the years 1981–2010. For each latitude ϕ and month m, we determined the mean temperature $\bar{T}(\phi, m)$ and the standard deviation $\sigma(\phi, m)$, taken over all longitudes and all 30 years in the dataset. The temperature was then uniformly sampled from a range of $\bar{T}(\phi, m) \pm 3\sigma(\phi, m)$, if $3\sigma > 8$ K, or $\bar{T}(\phi, m) \pm 8$ K otherwise. For simplicity, the sampled temperature was kept constant for the duration of the 24-h simulation. (3) The emission fluxes of aerosol and gases were sampled from a non-uniform distribution by multiplying the maximum emission rate with a random number between 0 and 1 raised to the fourth power. This ensured that our sampling space was skewed towards the lower emission rates, while still retaining some scenarios that represent highly polluted conditions. (4) The aerosol distributions for the emitted carbonaceous particles were prescribed as log-normal, with geometric mean diameter D_g and geometric standard deviation σ_g. (5) Sea salt particles were emitted wet rather than dry. For composition, a simplified mixture of 53.89% Cl^-, 38.56% Na^+, and 7.55% SO_4^{2-} by mass was, based on the mass ratio of Cl^- to SO_4^{2-} of 7.15 in seawater ([49], p. 384) and adding enough Na^+ to balance the charges. Additionally, because organic species are a substantial but variable component of sea salt aerosols Vignati et al. [50], a variable amount OC is added, making up 0% to 20% of the mass of the particles. One third of all scenarios had no sea salt emissions. (6) One third of all scenarios had no dust emissions.

A total of 1000 scenarios were created in this fashion to make up the training library. Since we are saving the output every 10 min of each 24-h simulations, this yields 144,000 particle populations for our training dataset. For testing purposes, a second library of 240 scenarios (34,560 populations) was created in the same manner to gauge the accuracy of the model, using the same distributions, but with different combinations of parameters. This provides a check against overfitting, in which the model that is learned has been fit to the stochastic noise in the training set, resulting in poor predictive performance for any other data set.

2.5. Machine Learning as Applied to PartMC

Machine learning refers to a variety of algorithms that are used to identify and model patterns in large datasets, and then use these models to make predictions. It has proven to be a diverse set of tools in the atmospheric sciences. Past applications have included interpreting remote sensing data [51], estimating uncertainty in aerosol optical depth data [52], prediction of aerosol-induced health impacts [53], and forecasting solar radiation for energy generation [54].

Our model predicts χ in a single global-model grid cell, given inputs of the GEOS-Chem-TOMAS variables in that grid cell. We present two variants of this model, one that predicts χ for the bulk aerosol population, and one that predicts χ for each size bin of the global model. A total of 34 input feature variables were used, including gas concentrations, aerosol mass concentrations, aerosol number concentration, solar zenith angle, and latitude. Note that the mass concentrations of the different aerosol species are not lumped into hygroscopic and non-hygroscopic species for this purpose, but are used individually. MOSAIC species were mapped to TOMAS species when training the model. At each horizontal location, we computed the average predicted χ over grid layers up to 840 mb.

We used gradient-boosted regression trees ([55], Chapter 10) as the machine-learning algorithm for this study, because this is a well-understood algorithm that offers good predictive accuracy with moderate computational cost and is able to perform automatic feature selection during training. Gradient boosting methods [56,57] form a prediction model as a sequence of weak prediction models, each of which fits the residual of the previous predictors in the sequence and thus serves to slightly improve the overall prediction accuracy.

For gradient-boosted regression trees, the weak prediction models are regression trees ([55], Section 9.2.2), which predict an output value as a tree of decisions on input values. For example, a single depth-2 regression tree for χ might have a first decision of "(latitude > 50°)?", and if this is true it might have a second decision of "([SO_2] < 30 ppb)?", and if this is false then it outputs $\chi = 0.8$. A depth-n tree allows up to n-way interactions between feature variables.

We used the implementation of gradient-boosted regression trees from scikit-learn [58]. The model was trained on the training data set and then its performance was evaluated on the testing data set (see Section 2.4). We used a least-squares loss function and all of our gradient-boosted models used 400 decision trees as submodels, as this was sufficient to obtain the best performance on the testing data set. We tested different tree depths, as shown in Figure 5 (left). Similar to many applications ([55], Chapter 10) we found that tree depths between 4 and 8 worked well, and we used depth 8 for the final model used in the remainder of the paper to give good prediction accuracy with reasonable computational speed.

The performance of our final model is shown in Figure 5 (right). In this figure, a perfect model would be the red 1:1 line. Our model has $R^2 = 0.94$ and a mean error of 1.67%. The maximum error for any testing scenario is 13.02%.

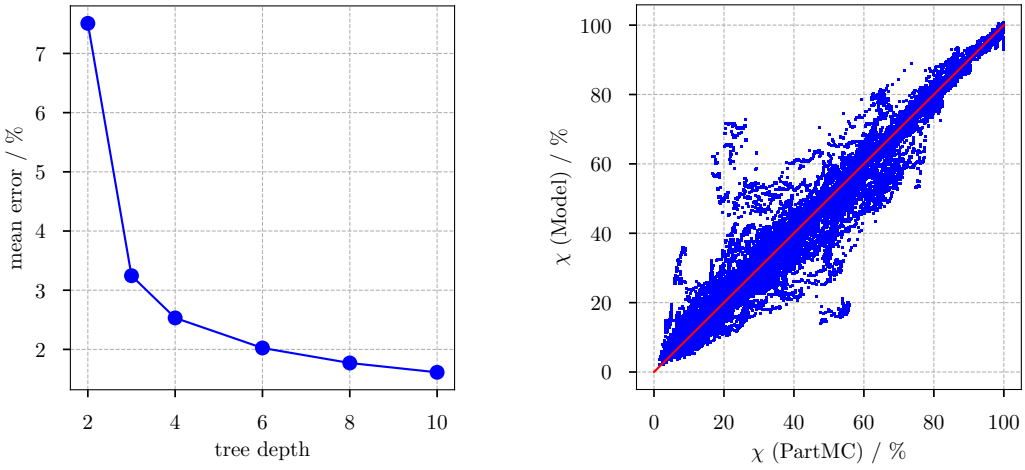

Figure 5. (**Left**) Mean error in the predicted χ values from the testing data set as a function of tree depth for the gradient boosted regression tree model; and (**Right**) true χ values versus model-predicted values for our final model (corresponding to depth 8 in the left panel).

3. Results

3.1. Predicting χ for the Bulk Aerosol Population

Using output from GEOS-Chem-TOMAS and the model for χ that was trained on particle-resolved data, we can now produce global distributions of χ. Figure 6 shows examples of such distributions using six-hourly output from GEOS-Chem-TOMAS and comparing two different dates, 06:00 UTC on 1 January 2010, and 1 June 2010. Note that χ was calculated based on the entire size range of aerosol particles and hence if coarse-mode particles and fine-mode particles have different compositions, this would result in a lower χ value (more externally mixed), even if the course and fine modes each had higher χ values (more internally mixed). Because χ is a mass-weighted quantity, the χ values for all sizes are dominated by the coarse mode mixing state.

We determined χ only for grid cells that contained between 5% and 95% hygroscopic material, hence excluding areas where essentially only one surrogate species (either hygroscopic or non-hygroscopic) was present. We see from Figure 6 that these excluded areas cover much of the oceans, and much of the Sahara and other deserts. This exclusion is because it is meaningless to discuss

the mixing state between hygroscopic and non-hygroscopic material when there is essentially only a single type present.

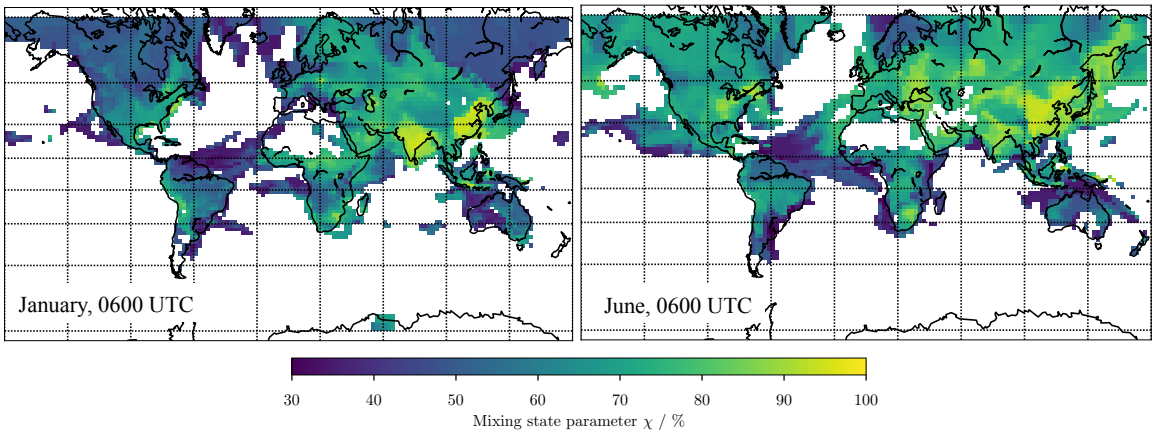

Figure 6. Global distribution of χ from the machine-learning model, at 06:00 UTC on January 1 (**left**), and June 1 (**right**), 2010. The model used to predict χ here is trained on the PartMC output that includes the entire particle population.

For both dates, the predicted χ varied from 30% to 97%. High χ values existed over industrial source regions including East China, India, and the Eastern and Midwestern United States, with χ approaching 100%. This result can be interpreted that in these regions non-hygroscopic (mainly freshly emitted carbonaceous aerosol) and hygroscopic (mainly secondary) aerosol species are mixed together within the same particle. This prediction is consistent with the fact that highly polluted areas have extremely short aging timescales for carbonaceous emissions [9,59], and so—at least on the scale of the grid resolution used here in GEOS-Chem-TOMAS—assuming an internal mixture of non-hygroscopic and hygroscopic species is appropriate. However, we note that the nucleation is frequently observed in many of these regions, and hence our training data that omitted nucleation may be overestimating χ in some of these regions.

Plumes of aerosol with relatively high χ values of around 80% can also be seen to be transported over the oceans in the outflow of continents, e.g., east of China. This was more prominent for 1 June over the Northern Hemisphere, which is consistent with a larger availability of photochemically produced secondary species that can condense on the originally non-hygroscopic carbonaceous particles, thereby moving the population towards a more internal mixture.

3.2. Predicting χ for Individual Size Bins

Rather than including the entire PartMC particle populations for the machine-learning process, we can also group the PartMC output according to particle size first, and then train a separate model for χ for each individual size category. This altered the input feature variables for the model from bulk aerosol mass concentrations and total number concentration to the mass aerosol mass concentrations and number concentration within the size range.

Choosing the TOMAS size bins, we obtained results for the testing data set, as shown in Table 5. The R^2 values are generally lower than for the case without size resolution, which is expected since for each size bin a smaller set of particles is available for learning the model. In fact, for size bins 1–6 (corresponding to dry diameters from ~3–30 nm), the R^2 value were very low, so that we only discuss the results for size bins 7 and larger (dry diameters above ~30 nm). In future work, we plan to refine these results by increasing the particle samples in the smaller size bins.

Table 5. Error statistics for prediction error in size-resolved χ.

Bin Number	7	8	9	10	11	12	13	14	15
Bin median diameter (nm)	56.3	89.4	142	225.3	357.7	567.8	901.4	2024	6424
R^2	19.68%	65.31%	79.68%	87.63%	90.87%	89.45%	81.87%	70.94%	36.42%
Mean error	12.55%	9.13%	7.21%	5.99%	5.16%	5.51%	6.91%	8.64%	11.86%

Figure 7 shows the global maps of size-resolved χ, based on GEOS-Chem-TOMAS output fields averaged for the months of January and July for size bin 8 (χ_8, bin median diameter of ~90 nm) and size bin 14 (χ_{14}, bin median diameter of ~2 µm). Other months had very similar distributions and are not shown.

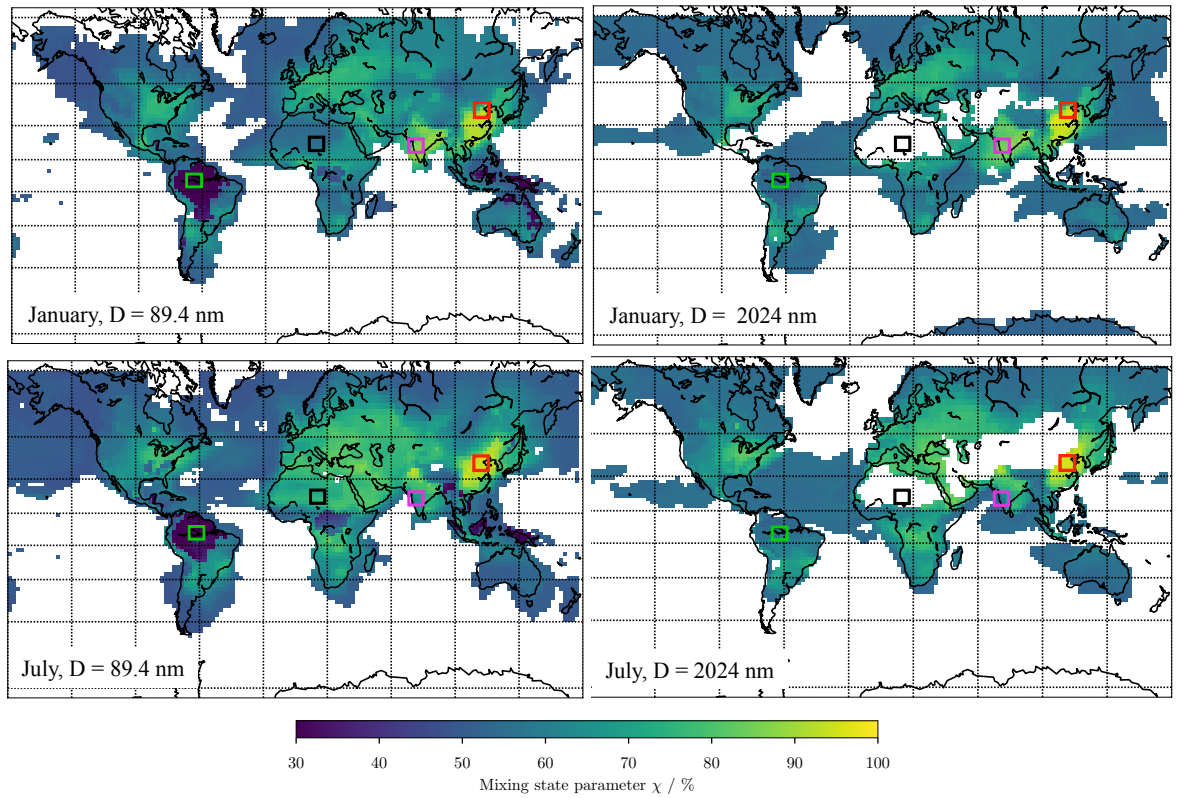

Figure 7. Global distribution of size-resolved χ values from the machine-learning model based GEOS-Chem-TOMAS inputs for the months of: January (**top**); and July (**bottom**). (**Left**) χ for size bin 8, bin median diameter is 89.4 nm. (**Right**) χ for size bin 14, bin median diameter is 2024 nm. The colored boxes show the regions over which data were averaged for display in Figure 9.

The distribution for χ_8 shows low values of approximately 20% in the Amazon basin, central Africa, and Indonesia. These are areas with large contribution of carbonaceous aerosol from biomass burning. The low χ_8 value in this size range means that the carbonaceous material is externally mixed from other (more hygroscopic) aerosol in these areas. In contrast, internally mixed aerosol is predicted for East Asia and India. For January, plumes with internally mixed aerosol extend from India into the Arabian Sea (winter Monsoon), while, for July, this is not the case (summer Monsoon).

Due to the setup of our scenario library, we need to be aware of some biases that we might introduce with our choices. By using the same initial condition for all simulations, we may underestimate χ in locations where the local emissions are relatively small but different to the initial conditions and where the conditions are not conducive for secondary aerosol formation. Conversely, the χ values for 90-nm particles in the polluted regions (Eastern US, India, Europe,

and China) are likely biased high, since our scenario library does not include nucleation, as mentioned in Section 2.4. Nucleation events and growth to 90 nm are routinely observed in these areas, along with primary carbonaceous emissions at these sizes, which may be fresh or aged. Overall, nucleation events are likely to decrease χ in this size range, since a more external mixture would be created. We plan to quantify the impacts of both the initial condition choice as well as the impacts of nucleation on the machine learning procedure in future work.

Figure 8 shows the composition of the aerosol in these two size bins as a fraction of hygroscopic species. This figure confirms that for large areas over the oceans the aerosol consists of only hygroscopic material, and the 2 µm bin over the desert areas contains only non-hygroscopic material, which is the reason why χ was not determined for those areas.

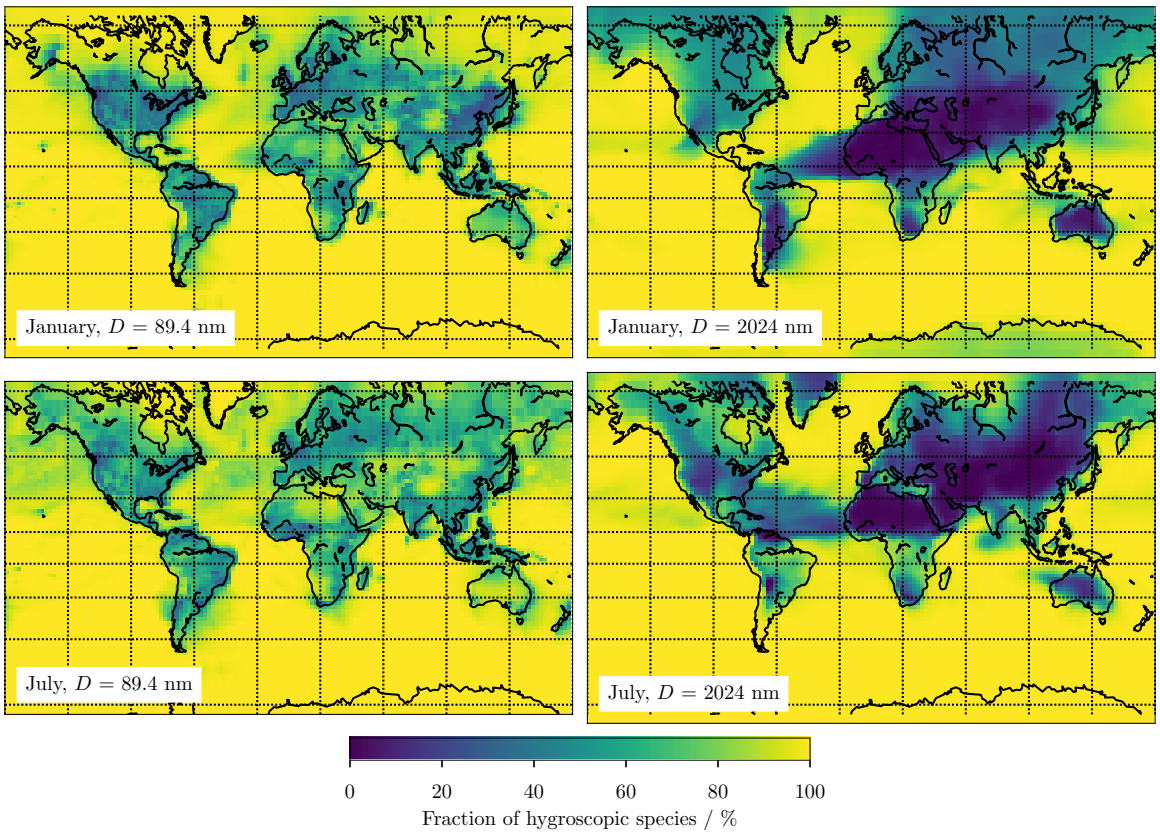

Figure 8. Global distribution of fraction of hygroscopic species as simulated by GEOS-Chem-TOMAS for the months of: January (**top**); and July (**bottom**). (**Left**) χ for size bin 8, bin median diameter is 89.4 nm. (**Right**) χ for size bin 14, bin median diameter is 2024 nm.

Figure 9 shows the size-resolved χ for selected regions, which are indicated in Figure 7 as colored boxes. This figure confirms the strong size dependence for the Amazon region while the other regions do not show a pronounced size dependence. Differences between summer and winter are noticeable for North East China (χ is higher in July), Sahara (χ is higher in July), and Central India (χ is lower in July). A possible explanation for the higher χ values in summer for North East China and the Sahara is a generally larger production of condensable gases during the Northern Hemisphere summer, which help creating a more internal mixture. The lower χ values over India during summer might be related to the Monsoon, which removes both condensable gases as well as aged aerosol.

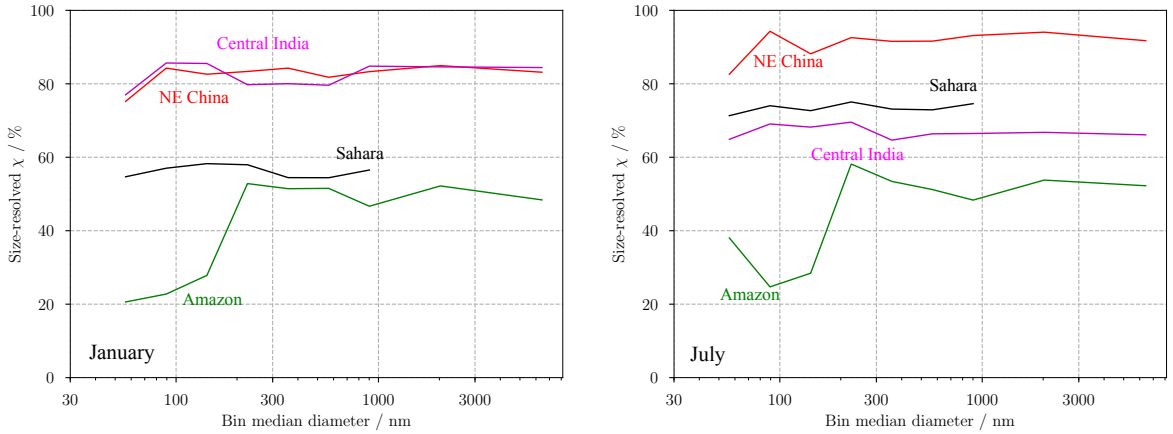

Figure 9. Size-resolved values of χ forselected regions in: January (**left**); and July (**right**). See Figure 7 for the location of these regions.

It is interesting to note that the smallest χ value is about 20%, so the classic "external mixture" with χ approaching zero is not found anywhere in these examples. We want to emphasize again that the χ values for smaller sizes in polluted regions such as North East China and Central India may be overestimated because our training data set does not include the process of nucleation. Including nucleation is generally expected to create a more external mixture, since freshly nucleated (hygroscopic) particles would co-exist with carbonacous particles in these environments.

4. Conclusions

This paper presents the first estimate of spatial distribution of aerosol mixing state over the globe as quantified by the mixing state metric χ. We defined this metric to estimate the degree to which hygroscopic and non-hygroscopic species are mixed on a per-particle basis, with $\chi = 0\%$ being completely externally mixed and $\chi = 100\%$ being completely internally mixed. We obtained this global estimate by training a machine-learning model of χ on detailed particle-resolved box model data, and then applying the model to GCM output to predict χ globally.

In some parts of the globe, the aerosol appeared to be quite externally mixed, with χ values as low as 20%, suggesting that an external-mixing assumption is likely to be valid there. This was the case for the size range below 150 nm in regions where biomass burning aerosol dominated, such as the Amazon Basin, Central Africa, and Indonesia. In contrast, the mixing state index χ reached values of 90% for polluted regions in East Asia in July, indicating that an internally-mixed assumption is appropriate for those regions, at least for the spatial resolution of the GCM that was used here. In much of the globe, however, the aerosol mixing state was not clearly internally or externally mixed, which may indicate that assuming either limiting case could lead to significant errors. Previous work by Ching et al. [14] can be used to link the global maps of χ values from this study with estimated errors for CCN concentrations. For the χ values between 30% and 100% found in this study, assuming an internal mixture would introduce an overestimation in CCN concentrations of up to 50%, with the error decreasing to a few percent for χ larger than 80%. For χ values lower than 20%, errors in CCN concentration of up to 100% can occur, but these χ values did not occur in our study. The scenarios in the study by Ching et al. [14] were focused on the aging of carbonaceous aerosol and and therefore did not encompass the full range of conditions that might be encountered around the globe. Nevertheless, they provide guidance of how the predicted distribution of χ values relates to expected errors in CCN predictions when assuming an internal mixture.

While the methodology used in this paper is effective at extrapolating high-detail simulation output to the global scale, it is important to understand the limitations of such a method. Roughly speaking, our model takes the GCM output variables in each grid cell and infers the

mixing state χ value from particle-resolved box model simulations with similar corresponding state variables. This could deliver inaccurate χ estimates if there are no similar box model scenarios, if there are multiple box model scenarios that differ significantly in their χ predictions; if the comparison is inexact due to differences in the microphysics/chemistry models between the GCM and PartMC-MOSAIC; or if the matching box model scenarios had significantly different histories and therefore have misleading mixing states. For example, the lack of nucleation in the box model scenarios may well lead to somewhat overpredicted χ values in the sizes up to 90 nm in polluted regions. Additionally, we assumed a composition of our pre-existing particles in our training simulations, which may influence our results presented here.

An important issue that should be addressed in future work is the question of end-to-end verification and validation of the χ predictions. This could be accomplished by performing single-particle measurements in different locations, similar to what has been done in Healy et al. [15] for a single location in Paris during the MEGAPOLIS campaign. Another possibility would be to perform particle-resolved aerosol simulations within a 3D chemical transport model (at great computational expense) to calculate χ directly over small regions, and to compare these explicitly calculated χ values to χ predicted with machine learning.

It will be straightforward to adapt our model training to predict χ based on aerosol optical properties, rather than hygroscopicity. This would answer the question of how absorbing and non-absorbing aerosol species are mixed on a per-particle basis, which is important to capture the absorption enhancement of black-carbon-containing aerosol [60,61]. The approach presented in this paper could be generalized to other problems where particle-scale processes cannot directly be simulated within the large-scale modeling framework, but for which accurate small-scale models exist.

Acknowledgments: M.H. and N.R. were supported by NSF AGS-1254428. J.K.K. and J.R.P. acknowledge funding from NSF AGS-1559607. M.W. acknowledges funding from NSF CMMI-1150490.

Author Contributions: M.H. and N.R. designed the PartMC-MOSAIC simulations and analyzed the results; M.H. performed the PartMC model runs and the machine learning; M.W. guided the machine learning procedures; N.R. wrote the paper with contributions of all authors; and J.K.K. and J.R.P. performed the GEOS-Chem-TOMAS simulations. All authors contributed to the interpretation of the results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Noble, C.A.; Prather, K.A. Real-time single particle mass spectrometry: A historical review of a quarter century of the chemical analysis of aerosols. *Mass Spectrom. Rev.* **2000**, *19*, 248–274.
2. Bein, K.J.; Zhao, Y.; Wexler, A.S.; Johnston, M.V. Speciation of size-resolved individual ultrafine particles in Pittsburgh, Pennsylvania. *J. Geophys. Res.* **2005**, *110*, D07S05.
3. Johnson, K.S.; Zuberi, B.; Molina, L.T.; Molina, M.J.; Iedema, M.J.; Cowin, J.P.; Gaspar, D.J.; Wang, C.; Laskin, A. Processing of soot in an urban environment: Case study from the Mexico City Metropolitan Area. *Atmos. Chem. Phys.* **2005**, *5*, 3033–3043.
4. Riemer, N.; West, M.; Zaveri, R.; Easter, R. Simulating the evolution of soot mixing state with a particle-resolved aerosol model. *J. Geophys. Res. Atmos.* **2009**, *114*, D09202.
5. Seigneur, C.; Hudischewskyj, A.B.; Seinfeld, J.H.; Whitby, K.T.; Whitby, E.R.; Brock, J.R.; Barnes, H.M. Simulation of aerosol dynamics: A comparative review of mathematical models. *Aerosol Sci. Technol.* **1986**, *5*, 205–222.
6. Wexler, A.S.; Lurmann, F.W.; Seinfeld, J.H. Modelling urban aerosols—I. Model development. *Atmos. Environ.* **1994**, *28*, 531–546.
7. Whitby, E.R.; McMurry, P.H. Modal aerosol dynamics modeling. *Aerosol Sci. Technol.* **1997**, *27*, 673–688.
8. Zaveri, R.; Barnard, J.; Easter, R.; Riemer, N.; West, M. Effect of aerosol mixing-state on optical and cloud activation properties. *J. Geophys. Res. Atmos.* **2010**, *115*, D17210.
9. Fierce, L.; Riemer, N.; Bond, T.C. Toward reduced representation of mixing state for simulating aerosol effects on climate. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 971–980.

10. Fierce, L.; Bond, T.C.; Bauer, S.E.; Mena, F.; Riemer, N. Black carbon absorption at the global scale is affected by particle-scale diversity in composition. *Nat. Commun.* **2016**, *7*, 12361.
11. Ching, J.; Riemer, N.; West, M. Impacts of black carbon mixing state on black carbon nucleation scavenging: Insights from a particle-resolved model. *J. Geophys. Res. Atmos.* **2012**, *117*, doi:10.1029/2012JD018269.
12. Ching, J.; Riemer, N.; West, M. Impacts of black carbon particles mixing state on cloud microphysical properties: Sensitivity to environmental conditions. *J. Geophys. Res. Atmos.* **2016**, *121*, 5990–6013.
13. Riemer, N.; West, M. Quantifying aerosol mixing state with entropy and diversity measures. *Atmos. Chem. Phys.* **2013**, *13*, 11423–11439.
14. Ching, J.; Fast, J.; West, M.; Riemer, N. Metrics to quantify the importance of mixing state for CCN activity. *Atmos. Chem. Phys.* **2017**, *17*, 7445–7458.
15. Healy, R.; Riemer, N.; Wenger, J.; Murphy, M.; West, M.; Poulain, L.; Wiedensohler, A.; O'Connor, I.; McGillicuddy, E.; Sodeau, J.; et al. Single particle diversity and mixing state measurements. *Atmos. Chem. Phys.* **2014**, *14*, 6289–6299.
16. O'Brien, R.E.; Wang, B.; Laskin, A.; Riemer, N.; West, M.; Zhang, Q.; Sun, Y.; Yu, X.Y.; Alpert, P.; Knopf, D.A.; et al. Chemical imaging of ambient aerosol particles: Observational constraints on mixing state parameterization. *J. Geophys. Res.* **2015**, *120*, 9591–9605.
17. Adams, P.J.; Seinfeld, J.H. Predicting global aerosol size distributions in general circulation models. *J. Geophys. Res. Atmos.* **2002**, *107*, 4370.
18. Kodros, J.K.; Cucinotta, R.; Ridley, D.A.; Wiedinmyer, C.; Pierce, J.R. The aerosol radiative effects of uncontrolled combustion of domestic waste. *Atmos. Chem. Phys.* **2016**, *16*, 6771–6784.
19. Ghan, S.J.; Abdul-Razzak, H.; Nenes, A.; Ming, Y.; Liu, X.; Ovchinnikov, M.; Shipway, B.; Meskhidze, N.; Xu, J.; Shi, X. Droplet nucleation: Physically-based parameterizations and comparative evaluation. *J. Adv. Model. Earth Syst.* **2011**, *3*, doi:10.1029/2011MS000074.
20. Whittaker, R.H. Evolution and Measurement of Species Diversity. *Taxon* **1972**, *21*, 213–251.
21. Drucker, J. Industrial Structure and the Sources of Agglomeration Economies: Evidence from Manufacturing Plant Production. *Growth Chang.* **2013**, *44*, 54–91.
22. Strong, S.; Koberle, R.; de Ruyter van Steveninck, R.; Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200.
23. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol. Ecol. Notes* **2007**, *7*, 574–578.
24. Giorio, C.; Tapparo, A.; Dall'Osto, M.; Beddows, D.C.; Esser-Gietl, J.K.; Healy, R.M.; Harrison, R.M. 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. *Environ. Sci. Technol.* **2015**, *49*, 3330–3340.
25. Fraund, M.; Pham, D.Q.; Bonanno, D.; Harder, T.H.; Wang, B.; Brito, J.; de Sá, S.S.; Carbone, S.; China, S.; Artaxo, P.; et al. Elemental Mixing State of Aerosol Particles Collected in Central Amazonia during GoAmazon2014/15. *Atmosphere* **2017**, *8*, 173, doi:10.3390/atmos8090173.
26. Dickau, M.; Olfert, J.; Stettler, M.E.J.; Boies, A.; Momenimovahed, A.; Thomson, K.; Smallwood, G.; Johnson, M. Methodology for quantifying the volatile mixing state of an aerosol. *Aerosol Sci. Technol.* **2016**, *50*, 759–772.
27. DeVille, R.E.L.; Riemer, N.; West, M. Weighted Flow Algorithms (WFA) for stochastic particle coagulation. *J. Comput. Phys.* **2011**, *230*, 8427–8451.
28. Michelotti, M.D.; Heath, M.T.; West, M. Binning for efficient stochastic multiscale particle simulations. *Multiscale Model. Simul.* **2013**, *11*, 1071–1096.
29. Zaveri, R.A.; Easter, R.C.; Fast, J.D.; Peters, L.K. Model for simulating aerosol interactions and chemistry (MOSAIC). *J. Geophys. Res. Atmos. (1984–2012)* **2008**, *113*, doi:10.1029/2007JD008782.
30. Zaveri, R.A.; Peters, L.K. A new lumped structure photochemical mechanism for large-scale applications. *J. Geophys. Res. Atmos. (1984–2012)* **1999**, *104*, 30387–30415.
31. Zaveri, R.A.; Easter, R.C.; Wexler, A.S. A new method for multicomponent activity coefficients of electrolytes in aqueous atmospheric aerosols. *J. Geophys. Res. Atmos. (1984–2012)* **2005**, *110*, doi:10.1029/2004JD004681.
32. Zaveri, R.A.; Easter, R.C.; Peters, L.K. A computationally efficient multicomponent equilibrium solver for aerosols (MESA). *J. Geophys. Res. Atmos. (1984–2012)* **2005**, *110*, doi:10.1029/2004JD005618.
33. Schell, B.; Ackermann, I.J.; Binkowski, F.S.; Ebel, A. Modeling the formation of secondary organic aerosol within a comprehensive air quality model system. *J. Geophys. Res.* **2001**, *106*, 28275–28293.

34. Tian, J.; Riemer, N.; West, M.; Pfaffenberger, L.; Schlager, H.; Petzold, A. Modeling the evolution of aerosol particles in a ship plume using PartMC-MOSAIC. *Atmos. Chem. Phys.* **2014**, *14*, 5327–5347.
35. Mena, F.; Bond, T.C.; Riemer, N. Plume-exit modeling to determine cloud condensation nuclei activity of aerosols from residential biofuel combustion. *Atmos. Chem. Phys.* **2017**, *17*, 9399–9415.
36. Bey, I.; Jacob, D.J.; Yantosca, R.M.; Logan, J.A.; Field, B.D.; Fiore, A.M.; Li, Q.; Liu, H.Y.; Mickley, L.J.; Schultz, M.G. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *J. Geophys. Res. Atmos.* **2001**, *106*, 23073–23095.
37. Napari, I.; Noppel, M.; Vehkamäki, H.; Kulmala, M. Parametrization of ternary nucleation rates for H_2SO_4-NH_3-H_2O vapors. *J. Geophys. Res. Atmos.* **2002**, *107*, 4381.
38. Jung, J.; Fountoukis, C.; Adams, P.J.; Pandis, S.N. Simulation of in situ ultrafine particle formation in the eastern United States using PMCAMx-UF. *J. Geophys. Res. Atmos.* **2010**, *115*, doi:10.1029/2009JD012313.
39. Westervelt, D.; Pierce, J.; Riipinen, I.; Trivitayanurak, W.; Hamed, A.; Kulmala, M.; Laaksonen, A.; Decesari, S.; Adams, P. Formation and growth of nucleated particles into cloud condensation nuclei: model—Measurement comparison. *Atmos. Chem. Phys.* **2013**, *13*, 7645–7663.
40. Vehkamäki, H.; Kulmala, M.; Napari, I.; Lehtinen, K.E.; Timmreck, C.; Noppel, M.; Laaksonen, A. An improved parameterization for sulfuric acid–water nucleation rates for tropospheric and stratospheric conditions. *J. Geophys. Res. Atmos.* **2002**, *107*, AAC 3-1–AAC 3-10.
41. Lee, Y.; Pierce, J.; Adams, P. Representation of nucleation mode microphysics in a global aerosol model with sectional microphysics. *Geosci. Model Dev.* **2013**, *6*, 1221–1232.
42. Lee, Y.; Adams, P. A fast and efficient version of the TwO-Moment Aerosol Sectional (TOMAS) global aerosol microphysics model. *Aerosol Sci. Technol.* **2012**, *46*, 678–689.
43. Kodros, J.; Pierce, J. Important global and regional differences in aerosol cloud-albedo effect estimates between simulations with and without prognostic aerosol microphysics. *J. Geophys. Res. Atmos.* **2017**, *122*, 4003–4018.
44. Pierce, J.; Croft, B.; Kodros, J.; D'Andrea, S.; Martin, R. The importance of interstitial particle scavenging by cloud droplets in shaping the remote aerosol size distribution and global aerosol-climate effects. *Atmos. Chem. Phys.* **2015**, *15*, 6147–6158.
45. Fierce, L.; Riemer, N.; Bond, T.C. Explaining variance in black carbon's aging timescale. *Atmos. Chem. Phys.* **2015**, *15*, 3173–3191.
46. Kulmala, M.; Petäjä, T.; Ehn, M.; Thornton, J.; Sipilä, M.; Worsnop, D.; Kerminen, V.M. Chemistry of atmospheric nucleation: On the recent advances on precursor characterization and atmospheric cluster composition in connection with atmospheric new particle formation. *Annu. Rev. Phys. Chem.* **2014**, *65*, 21–37.
47. McKay, M.D.; Beckman, R.J.; Conover, W.J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **1979**, *21*, 239–245.
48. Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; et al. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 437–471.
49. Seinfeld, J.; Pandis, S. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
50. Vignati, E.; Facchini, M.; Rinaldi, M.; Scannell, C.; Ceburnis, D.; Sciare, J.; Kanakidou, M.; Myriokefalitakis, S.; Dentener, F.; O'Dowd, C. Global scale emission and distribution of sea-spray aerosol: Sea-salt and organic enrichment. *Atmos. Environ.* **2010**, *44*, 670–677.
51. Camps-Valls, G.; Bruzzone, L. *Kernel Methods for Remote Sensing Data Analysis*; John Wiley & Sons: Chichester, UK, 2009.
52. Ristovski, K.; Vucetic, S.; Obradovic, Z. Uncertainty analysis of neural-network-based aerosol retrieval. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 409–414.
53. Lary, D.J.; Lary, T.; Sattler, B. Using machine learning to estimate global PM2.5 for environmental health studies. *Environ. Health Insights* **2015**, *9*, 41–52.
54. Lauret, P.; Voyant, C.; Soubdhan, T.; David, M.; Poggi, P. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Sol. Energy* **2015**, *112*, 446–457.
55. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
56. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378.

57. Mason, L.; Baxter, J.; Bartlett, P.L.; Frean, M.R. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems*; Solla, S.A., Leen, T.K., Müller, K., Eds.; MIT Press: Cambridge, MA, USA, 2000; Volume 12, pp. 512–518.
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
59. Riemer, N.; West, M.; Zaveri, R.; Easter, R. Estimating black carbon aging time-scales with a particle-resolved aerosol model. *J. Aerosol Sci.* **2010**, *41*, 143–158.
60. Schnaiter, M.; Horvath, H.; Möhler, O.; Naumann, K.H.; Saathoff, H.; Schöck, O. UV-VIS-NIR spectral optical properties of soot and soot-containing aerosols. *J. Aerosol Sci.* **2003**, *34*, 1421–1444.
61. Peng, J.; Hu, M.; Guo, S.; Du, Z.; Zheng, J.; Shang, D.; Zamora, M.L.; Zeng, L.; Shao, M.; Wu, Y.S.; et al. Markedly enhanced absorption and direct radiative forcing of black carbon under polluted urban environments. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 4266–4271.