



## Quantifying errors in the aerosol mixing-state index based on limited particle sample size

J. T. Gasparik , Q. Ye , J. H. Curtis , A. A. Presto , N. M. Donahue , R. C. Sullivan , M. West & N. Riemer

To cite this article: J. T. Gasparik , Q. Ye , J. H. Curtis , A. A. Presto , N. M. Donahue , R. C. Sullivan , M. West & N. Riemer (2020) Quantifying errors in the aerosol mixing-state index based on limited particle sample size, *Aerosol Science and Technology*, 54:12, 1527-1541, DOI: [10.1080/02786826.2020.1804523](https://doi.org/10.1080/02786826.2020.1804523)

To link to this article: <https://doi.org/10.1080/02786826.2020.1804523>



Published online: 03 Sep 2020.



Submit your article to this journal [↗](#)



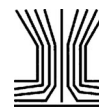
Article views: 237



View related articles [↗](#)



View Crossmark data [↗](#)



# Quantifying errors in the aerosol mixing-state index based on limited particle sample size

J. T. Gasparik<sup>a</sup> , Q. Ye<sup>b,\*</sup> , J. H. Curtis<sup>c</sup> , A. A. Presto<sup>b,d</sup> , N. M. Donahue<sup>b</sup> , R. C. Sullivan<sup>b</sup> , M. West<sup>c</sup> , and N. Riemer<sup>a</sup>

<sup>a</sup>Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; <sup>b</sup>Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; <sup>c</sup>Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; <sup>d</sup>Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

## ABSTRACT

This study evaluates the error that is introduced in quantifying observed aerosol mixing states due to a limited particle sample size. We used the particle-resolved model PartMC-MOSAIC to generate a scenario library that encompasses a large number of reference particle populations with a wide range of mixing states quantified by the mixing-state index  $\chi$ . We stochastically sub-sampled these particle populations using sample sizes of 10 to 10,000 particles and recalculated  $\chi$  based on the sub-samples. The finite sample size led to a consistent overestimation of  $\chi$ , with the 95% confidence intervals ranging from  $-70$  to  $30$  percentage points for sample sizes of 10 particles, and decreasing to  $\pm 10$  percentage points for sample sizes of 10,000 particles. These findings were experimentally confirmed with single-particle measurements from the Pittsburgh area using a soot-particle aerosol mass spectrometer.

## ARTICLE HISTORY

Received 17 April 2020  
Accepted 7 July 2020

## EDITOR

Kihong Park

## 1. Introduction

Atmospheric aerosols are evolving mixtures of different chemical species (Prather, Hatch, and Grassian 2008). The term “aerosol mixing state” is commonly used to describe how different chemical species are distributed throughout a particle population (Winkler 1973; Riemer et al. 2019). Aerosol mixing state influences the particles’ reactivity (Ryder et al. 2014), their optical properties (Moffet and Prather 2009; Lesins, Chylek, and Lohmann 2002), their hygroscopicity (Sullivan et al. 2009; Ching et al. 2017), and their propensity to serve as ice nuclei (Beydoun, Polen, and Sullivan 2017; Knopf and Alpert 2013). Hence, to predict aerosol impacts on atmospheric chemistry and climate, it is important to account for mixing state (Riemer et al. 2019); and this motivates efforts to determine mixing state from ambient observations (Healy et al. 2014; O’Brien et al., 2015; Ye et al. 2018) and via modeling (Riemer et al. 2009; Riemer and West 2013; Ching, Riemer, and West 2016).

The terms “internal” and “external” mixture qualitatively describe mixing state. A population is considered fully internally mixed if each individual particle

consists of the same species mixtures, while an external mixture implies that the different aerosol species reside in separate particles. Most ambient aerosol populations do not fit into either category, but resemble both internal and external mixtures to a degree. Often the term “mixing state” is applied to particles in a given size range, for example the accumulation mode or the coarse mode. This choice depends on the aerosol sampling instrumentation specifications or the science question being addressed. It is also important to be aware that mixing state as defined here does not capture the full potential diversity of particle populations, as the particle morphology (e.g., core-shell, well-mixed) can add additional diversity. In this article, we will only consider mixing state as defined above, which is also termed the “chemical mixing state” in Riemer et al. (2019).

For a quantitative description of aerosol mixing state, the mixing-state index  $\chi$  has been introduced (Riemer and West 2013), which can be calculated based on the particles’ species mass fractions. This scalar quantity ranges from 0 to 100% for fully external to internal mixtures, respectively. Several field studies have used this index to quantify mixing states

**CONTACT** N. Riemer [nriemer@illinois.edu](mailto:nriemer@illinois.edu) Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

\*Current affiliation: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

for different ambient environments using sophisticated single-particle measurement techniques, including electron microscopy and X-ray spectroscopy (O'Brien et al., 2015), and mass spectrometry (Healy et al. 2013). These observations confirm that mixing states in the ambient atmosphere are neither completely internally nor externally mixed but rather exist on a spectrum in between those limiting cases with characteristic temporal and spatial variability. For example, a study by Healy et al. (2014) for the MEGAPOLI campaign in Paris, France, revealed that the aerosol in Paris was estimated to be 59% internally mixed on average, with more external mixtures during the daytime when primary emissions from traffic and woodburning were present, and more internal mixtures during the night when ammonium nitrate formation prevailed. Healy et al. (2014) used single-particle aerosol time-of-flight mass spectrometry (ATOFMS) data for their study. Ye et al. (2018) quantified the spatial variation of  $\chi$  for the city of Pittsburgh on a neighborhood scale with a mobile measurement platform using single-particle measurements from a soot-particle aerosol mass spectrometer (SP-AMS) and found the lowest values (36%) close to an interstate highway and the highest values (76%) in rural or suburban regions.

The mixing-state index  $\chi$  is based on species diversity measures (Riemer and West 2013), which have been extensively used and developed in ecology and related fields. See Daly, Baetens, and Baets (2018) for an excellent overview and Sherwin and Prat I Fornells (2019) for a discussion of the history. Within ecology it is well-known that undersampling can result in inaccurate and biased estimates of diversity measures (Beck and Schwanghart 2010; Beck, Holloway, and Schwanghart 2013) and the performance of different statistical methods has been studied for realistic scenarios (Butturi-Gomes et al. 2017; Brocklehurst, Day, and Fröbisch 2018). In response, better estimators have been proposed that are based on the discovery rate of new species (Chao, Wang, and Jost 2013; Chao and Jost 2015; Haegeman et al. 2013), measures such as pairwise dissimilarities (Marion 2016) have been used as alternatives, and Bayesian estimators have been developed (Marion, Fordyce, and Fitzpatrick 2018).

Just as undersampling is a problem in ecology, measurements of atmospheric aerosols inherently sample a finite number of particles to estimate the mixing state and associated metrics. These finite samples range from a few hundred to many thousands of particles, depending on the instrumentation. For

example, O'Brien et al. (2015) utilized electron microscopy and X-ray spectroscopy methods to analyze particle samples from the 2010 Carbonaceous Aerosols and Radiative Effects study with sample sizes ranging from several hundred to several thousand particles. Ye et al. (2018) used single-particle mass spectrometry with sample sizes on the order of tens of thousand of particles. Unfortunately, we cannot directly apply the improved diversity estimators developed in ecology (Chao and Jost 2015) because they use the fact that species abundance is measured by sampling individuals in the species, a concept that does not readily transfer to chemical measurements.

The question arises of how large a particle sample size must be to adequately represent the mixing state of an atmospheric aerosol. Using a large ensemble of simulated aerosol populations generated with a particle-resolved model, the goal of this article is to quantify errors in determining  $\chi$  introduced by limited-size particle samples. Since the "true" value of  $\chi$  is not known when making observations in practice, we also provide confidence intervals around the measured  $\chi$  values for different sample sizes.

## 2. Methods

### 2.1. Calculating mixing-state index $\chi$

The mixing-state index  $\chi$  by Riemer and West (2013) provides a rigorous definition of aerosol mixing state. It is given by the affine ratio of the diversity metrics  $D_x$  and  $D_y$  as

$$\chi = \frac{D_x - 1}{D_y - 1}. \quad (1)$$

The diversity metrics, in turn, are defined as follows. First, the particle mixing entropies  $H_i$  need to be calculated for each particle based on the particle species mass fractions

$$H_i = \sum_{a=1}^A -p_i^a \ln p_i^a, \quad (2)$$

where  $A$  is the number of distinct aerosol species, and  $p_i^a$  is the mass fraction of species  $a$  in particle  $i$ . The particle diversities  $D_i = \exp(H_i)$  give the *effective number of species* in each particle, which is equal to the number of physical species if they are all present in equal proportions, and less otherwise.

The particle  $H_i$  values are then averaged (mass-weighted) over the entire population to give  $H_x$ , and finally the average particle species diversity  $D_x$ , by

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

$$D_\alpha = e^{H_\alpha}, \quad (4)$$

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.  $D_\alpha$  is the *mean particle diversity*, which gives the *mean effective number of species* over all particles in the population.

Lastly, the bulk diversity  $D_\gamma$  is defined by

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

$$D_\gamma = e^{H_\gamma}. \quad (6)$$

This is the *total diversity* of the population, which is the *effective number of species* in the aerosol bulk.

Importantly, the definition of “species” depends on the application or the instrumentation used to determine mass fractions. In some previous studies, elemental species have been used (O’Brien et al., 2015; Fraund et al. 2017; Bondy et al. 2018), while others used molecular species (Healy et al. 2014; Ye et al. 2018) or species groups (Dickau et al. 2016; Ching et al. 2017; Hughes et al. 2018). To make our results comparable to Ye et al. (2018) (who observed organics, nitrate, sulfate, chloride and black carbon, as detected by the soot-particle aerosol mass spectrometer), we chose the aerosol model species that constitute the dry aerosol mass, such as ammonium, sulfate, nitrate, black carbon, as well as several organic species, with the addition of dust and sodium chloride. Note that the soot-particle aerosol mass spectrometer cannot measure dust and sodium chloride. Aerosol water is excluded from our calculations. While our scenario library includes coarse-mode particles, we only included sub-micron particles in our calculations for  $\chi$ , since this is the relevant size range for ambient measurements using AMS instruments.

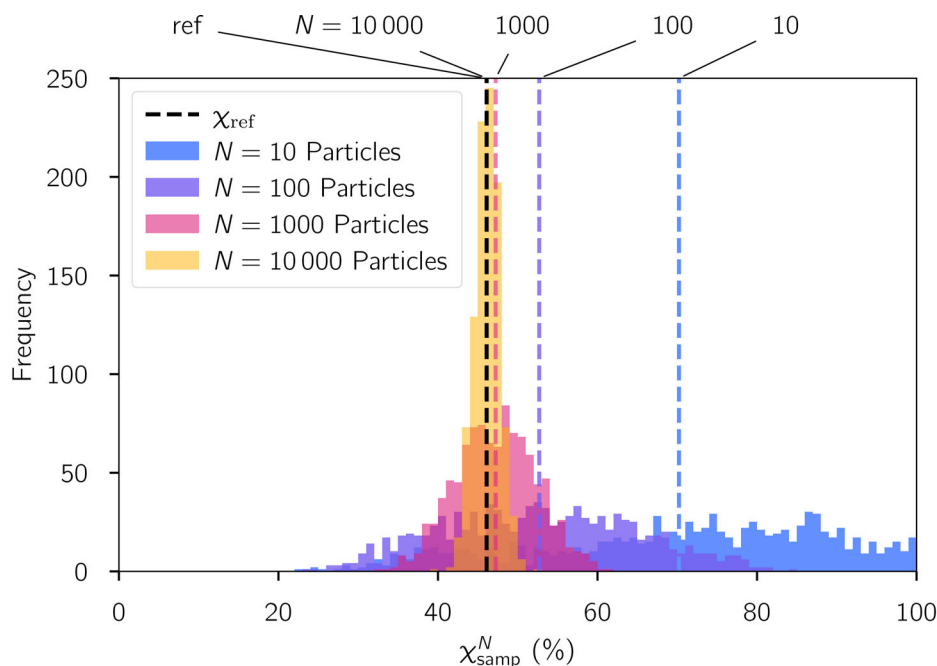
## 2.2. PartMC-MOSAIC model description

PartMC-MOSAIC is a unique modeling tool to simulate aerosol mixing state and its impacts under a wide variety of conditions. With this tool, each particle can be represented explicitly, allowing for accurate calculations of single-particle species mass fraction. This is in contrast to other common aerosol modeling techniques which represent averages of particle composition over certain size ranges rather than per-particle composition (Riemer et al. 2019). A detailed description of PartMC-MOSAIC is given in Riemer et al. (2009). In

brief, PartMC (Particle-resolved Monte Carlo) is a box model that explicitly resolves composition on a per-particle level within a well-mixed computational volume representative of a much larger air parcel. The evolution of the particle population—due to Brownian coagulation, emission and dilution—is tracked throughout the simulation using the Monte Carlo approach. PartMC is coupled to MOSAIC (Model for Simulating Aerosol Interactions and Chemistry) (Zaveri et al. 2008) which models gas-phase chemistry and gas-particle partitioning (condensation processes). MOSAIC consists of four modules: (1) the gas-phase photochemistry mechanism CBM-Z (Zaveri and Peters 1999), (2) the Multicomponent Taylor Expansion Method (MTEM) (Zaveri, Easter, and Wexler 2005b), (3) the Multicomponent Equilibrium Solver for Aerosols (MESA) for intraparticle solid-liquid partitioning (Zaveri, Easter, and Peters 2005a), and (4) the Adaptive Step Time-split Euler Method (ASTEM) for dynamic gas-particle partitioning (Zaveri et al. 2008). To simulate secondary organic aerosol (SOA) the SORGAM scheme is used (Schell et al. 2001). MOSAIC treats all locally and globally important gas and aerosol species including a total of 77 gaseous species and 19 aerosol species including  $\text{SO}_4^{2-}$ ,  $\text{HSO}_4^-$ ,  $\text{NO}_3^-$ ,  $\text{Cl}^-$ ,  $\text{CO}_3^{2-}$ ,  $\text{NH}_4^+$ ,  $\text{Na}^+$ ,  $\text{Ca}^{2+}$ , other inorganic mass (representing crustal material), black carbon (BC), primary organic aerosol (POA) and SOA.

## 2.3. Ensemble of scenarios and sampling technique

To generate confidence intervals for  $\chi$  for a wide range of conditions, we made use of a scenario library comprising 1000 different PartMC-MOSAIC scenarios (Hughes et al. 2018). All scenarios used a simulation time of 24 h, starting at 6:00 AM local time, with output saved every hour. Each simulation was run using 100,000 computational particles, producing a high-resolution representation of aerosol mixing state. Twenty-four input parameters (temperature, relative humidity, latitude, gas phase emission rates, emission rates, size parameters and composition of primary aerosol particles, including carbonaceous particles, sea salt, and mineral dust) were varied between scenarios to allow for large variations in mixing-state evolution. The scenario inputs were generated using Latin hypercube sampling to provide efficient sampling across the high-dimensional input parameter space. The entire scenario library generated 24,000 distinct *reference particle populations* (24 h  $\times$  1000 scenarios) (Gasparik



**Figure 1.** Frequency distribution of  $\chi_{\text{samp}}^N$  for one example population ( $\chi_{\text{ref}} = 46\%$ ) and different sample sizes. For each sample size, the sampling process was repeated 1000 times. The dashed colored lines correspond to the average  $\bar{\chi}_{\text{samp}}^N$  for the specified sample size.

et al. 2020). For each reference population we calculated the reference value  $\chi_{\text{ref}}$  of the mixing-state index.

To mimic the sampling process used in single-particle field measurements, we subsampled each reference population without replacement using different sample sizes ( $N = 10, 100, 1000, 10,000$ ). To determine confidence intervals we repeated each subsampling 1000 times, which resulted in 24,000,000 *sampled particle populations* ( $24\text{ h} \times 1000\text{ scenarios} \times 1000\text{ samples}$ ) for each sample size  $N$ . For each sampled population we recalculated  $\chi$  using only the particles in the sample and we denoted these  $\chi$  values as  $\chi_{\text{samp}}^N$ . Similarly,  $D_{\alpha, \text{samp}}^N$  and  $D_{\gamma, \text{samp}}^N$  are the  $\alpha$ - and  $\gamma$ -diversities computed using only the sampled particles.

It is important to note that even the large number of 100,000 computational particles still represents an—albeit large—subsample of the “true” limiting population with a (near-)infinite number of particles. The key point is that the error between the 100,000 particle sample and the true population is expected to be much smaller than the error between our largest subsample (10,000 particles) and the true population. This is a reasonable assumption as the error scales with  $1/\sqrt{N}$ , meaning that the error for the 100,000-particle sample is a factor of  $\sqrt{10}$  smaller than that of the 10,000-sample. To maintain this  $\sqrt{10}$  factor, we limit the largest sample to 10% of the size of our reference population.

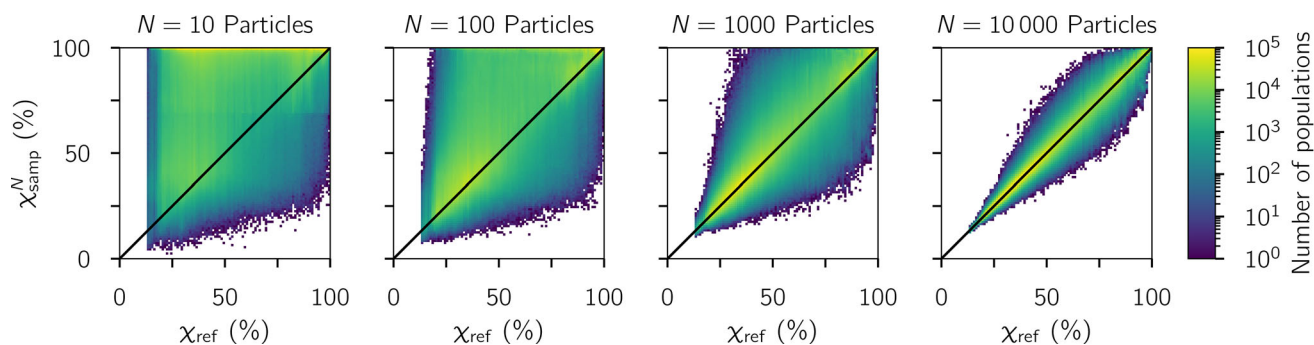
Figure 1 illustrates this process for one single reference population. In this case,  $\chi_{\text{ref}}$  was 46%. Sampling this population with only 10 particles produced a broad range of  $\chi_{\text{samp}}^{10}$  values from 20 to 100%. This range narrowed progressively as  $N$  increased, resulting in a range of 41 to 50% for a sample size of 10,000 particles. Another important result is that for small sample sizes,  $\chi_{\text{samp}}^N$  overestimates  $\chi_{\text{ref}}$ . In Section 3, we will see that this positive bias is a consistent result of the sampling process that can be explained with the fact that on average a sub-population overestimates  $D_{\alpha}$  and underestimates  $D_{\gamma}$ . A rigorous proof is provided in Section 4.

#### 2.4. Experimental determination of $\chi$ using observations in Pittsburgh

To provide experimental confirmation for the particle-resolved modeling results, a similar analysis was conducted using field data from Ye et al. (2018). In this study, aerosol samples were collected using the single-particle mode of a soot-particle aerosol mass spectrometer on a mobile platform throughout the city of Pittsburgh, PA.

The seven major particle clusters or types identified in the Pittsburgh mobile sampling data set were classified as: a sulfate-rich inorganic class that also contained OA and nitrate measured in the summer, a





**Figure 2.** Distribution of sampled population mixing-state index  $\chi_{\text{samp}}^N$  and reference mixing-state index  $\chi_{\text{ref}}$  for increasing sample sizes based on the simulated scenario library described in Section 2.3. The solid black line is the 1:1 line.

nitrate-dominated class with small amounts of sulfate and little OA measured in the winter, less-oxidized hydrocarbon-like organics (HOA) associated with vehicle emissions, less-oxidized cooking-like organic aerosol (COA) associated with restaurant emissions, black carbon-dominated with small amounts of OA or inorganics, more-oxygenated OA rich, and less-oxygenated OA rich (Ye et al. 2018). The later two OA-rich classes contained small amounts of inorganics. Differences in particle composition and mixing state and of these properties as a function of particle size were observed in different sampled environments that included: in highly trafficked tunnels, on highways, an urban area with high traffic density, and on a road through a large park. Other specific environments that produced unique particle mixing states included: inside a restaurant plume (COA dominated), downtown with high restaurant density (mixture of COA, inorganics, and HOA), and a suburban residential area with low restaurant density (diverse mixture of inorganics, OA, COA, and HOA) (Ye et al. 2018).

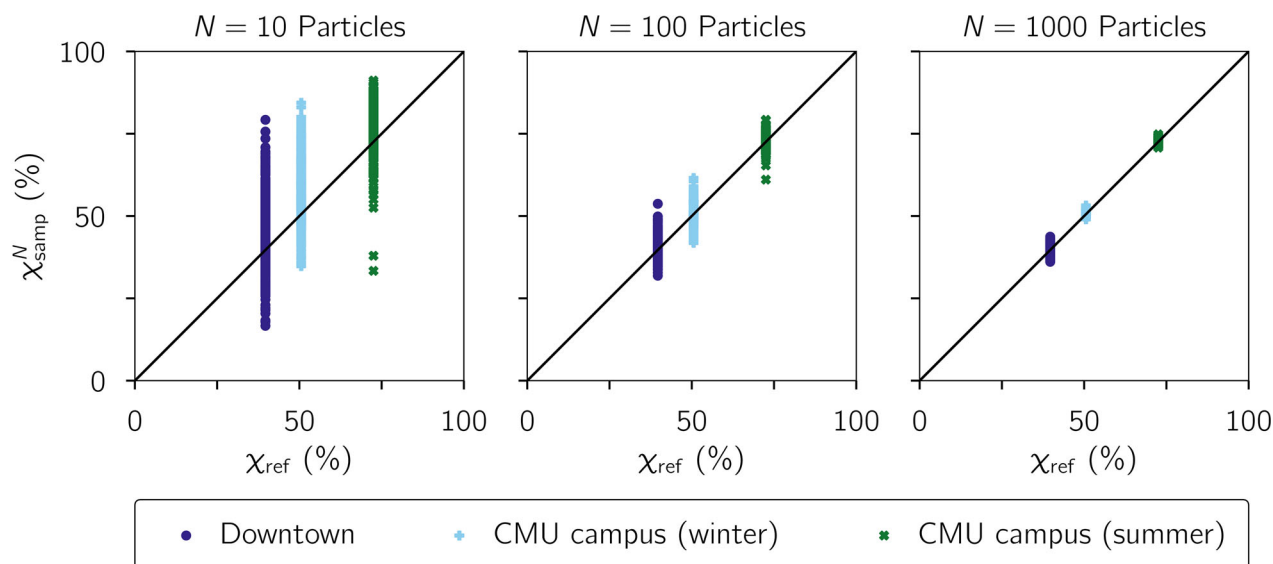
Three aerosol samples were used for the analysis here including the sample collected in Pittsburgh downtown (~11,000 particles), at the Carnegie Mellon University (CMU) urban campus in the summer (~15,000 particles) and at the CMU campus in the winter (~47,000 particles). Aerosols collected in Pittsburgh downtown were chosen to represent samples from regions that are in close proximity to sources of primary particles, while the CMU campus represents an urban background. Aerosol populations collected in Pittsburgh downtown were a combination of six visits to downtown, and the sampling time for each visit ranged between 30 min and 1 h. Aerosol populations collected on CMU campus were aggregated over several hours to a day. Based on the single-particle spectra of the three populations, per-particle mass fractions were determined and mixing state indices  $\chi_{\text{ref}}$  were calculated for each population. Five

species are considered for calculations of the mixing state index: organics, nitrate, sulfate, chloride and black carbon. Ammonium is not considered due to the large interference from the water signal in the mass spectra in the single-particle aerosol mass spectrometer. For more details about the method of using a soot-particle aerosol mass spectrometer to determine per-particle mass fractions and the mixing state index, please see Ye et al. (2018). Populations of 10, 100, and 1000 particles were stochastically sampled from the full particle samples from each location. Since the full particle samples are on the order of tens of thousand particles, our largest subsample is 1000, with the rationale explained in Section 2.3. The mixing state indices  $\chi_{\text{samp}}$  were calculated for each of the limited sample sizes and used for comparative analysis in this article.

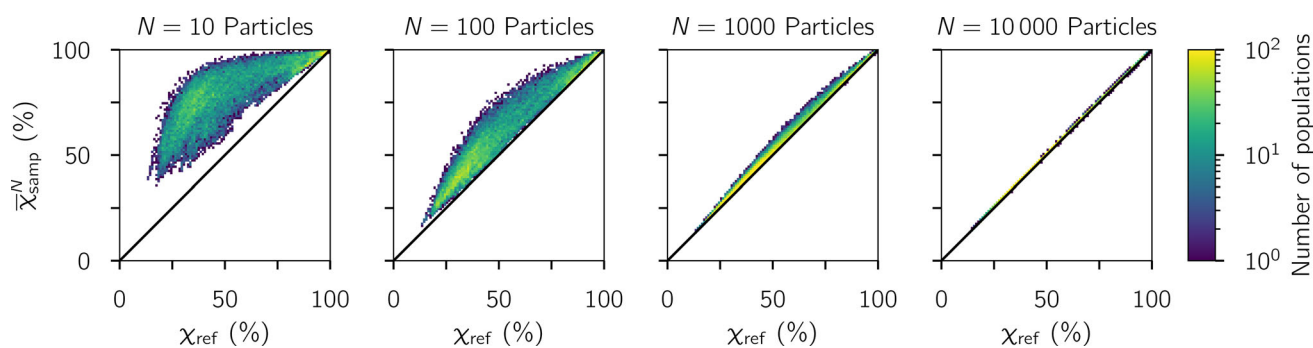
### 3. Numerical and observational results

The sampling results are presented as two-dimensional histograms that, for each sample size  $N$ , include all 1000 samples of the 24,000 simulated reference populations (resulting in 24,000,000 data points for each sample size). Figure 2 shows  $\chi_{\text{samp}}^N$  versus  $\chi_{\text{ref}}$  for the four sample sizes. Estimating  $\chi_{\text{ref}}$  based on a particle sample of only 10 particles led to results that can greatly overestimate or underestimate  $\chi_{\text{ref}}$ . As the sample size increased from 10 to 10,000 particles the points converged on the one-to-one line, meaning that sampled mixing-state values more accurately approximate the associated reference values.

Figure 3 shows the corresponding plot using the field data from Pittsburgh. Each vertical cluster of points corresponds to data from a location or time period where the sampling occurred. Each of these clusters has a different  $\chi_{\text{ref}}$  value, calculated from the full samples discussed in Section 2.4. For the downtown location, the  $\chi_{\text{ref}}$  value was lowest (40%),



**Figure 3.** Sampled  $\chi_{\text{samp}}^N$  as a function of  $\chi_{\text{ref}}$  from data obtained in Pittsburgh, PA. The one-to-one line is drawn for reference.



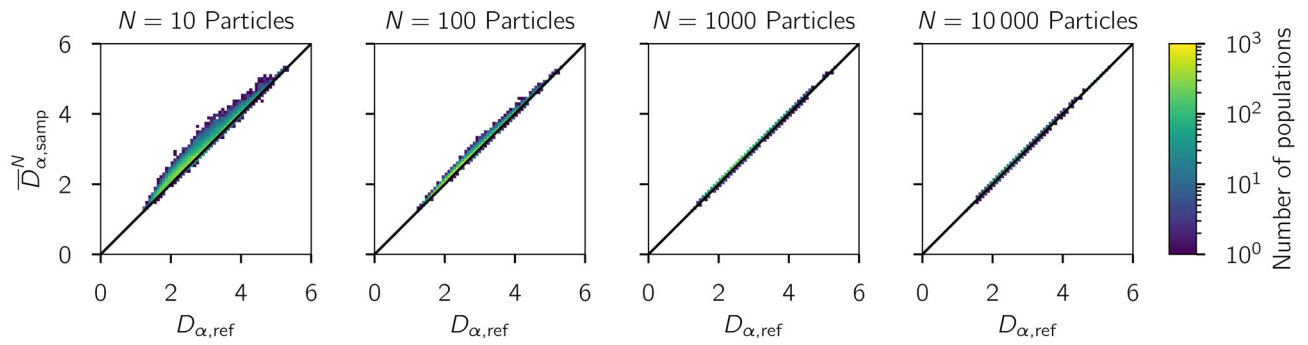
**Figure 4.** Distribution of average sampled  $\bar{\chi}_{\text{samp}}^N$  (averaged over the 1000 repeats) and  $\chi_{\text{ref}}$  for increasing sample sizes based on the simulated scenario library described in Section 2.3. The one-to-one line is drawn for reference.

consistent with the expectation for an area where fresh emissions mix with more aged aerosol. The aerosol at the CMU campus during winter ( $\chi_{\text{ref}} = 51\%$ ) was more externally mixed compared to the summer period ( $\chi_{\text{ref}} = 72\%$ ). This can be explained by the more extensive photochemical oxidation that drives the production of condensable secondary components, which condense onto preexisting particles and thus homogenize aerosol composition. A more detailed discussion on the spatial and temporal variability of aerosol mixing state in the Pittsburgh area can be found in Ye et al. (2018). Similar to the procedure used to analyze the modeled results, each  $\chi_{\text{ref}}$  population was stochastically sub-sampled, which resulted in a range of  $\chi_{\text{samp}}^N$  values that converged to the reference value as  $N$  increased.

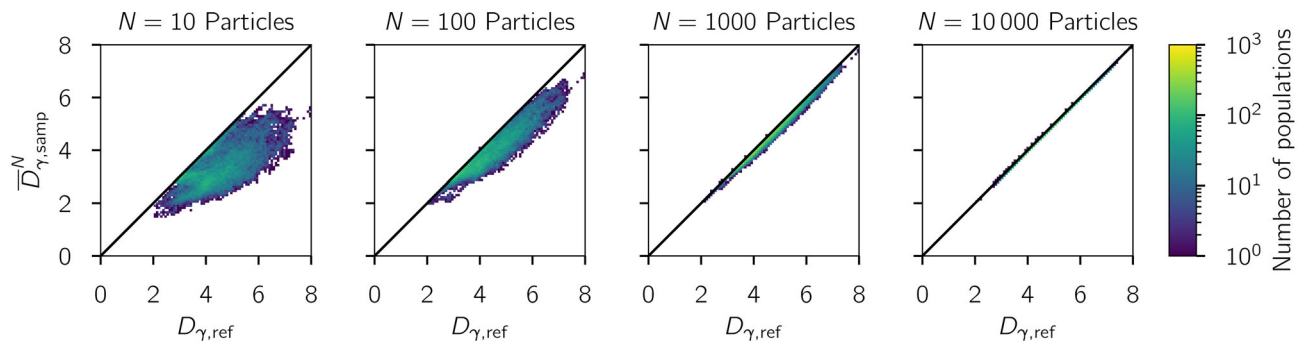
Both the simulation (Figure 2) and field data (Figure 3) results show that the sampled  $\chi_{\text{samp}}^N$  has a positive bias. That is, it tends to be larger than the reference  $\chi_{\text{ref}}$ . To investigate this more precisely, we computed the average  $\bar{\chi}_{\text{samp}}^N$  for the simulation data, where the average is taken over all 1000 repeats and is

mass-weighted (see Section 4.2 for details). This quantity is plotted versus  $\chi_{\text{ref}}$  in Figure 4. In contrast to Figure 2 which displayed all individual  $1000 \times 24,000$  data points, Figure 4 shows the averages over the 1000 repeats. This lets us see the patterns more clearly and confirms that  $\chi_{\text{samp}}^N$  is positively biased (above the one-to-one line). As expected, the bias vanished as the sample size increased from 10 to 10,000 particles. The question arises of how this bias can be explained. In particular, since  $\chi$  is the affine ratio of the average particle species diversity  $D_\alpha$  and the bulk diversity  $D_\gamma$  (Equation (1)), we need to determine the effect of sub-sampling on estimating these quantities individually before calculating the ratio.

The positive bias is a result of biases in the average sampled diversity metrics,  $\bar{D}_{\alpha, \text{samp}}^N$  and  $\bar{D}_{\gamma, \text{samp}}^N$ . Figures 5 and 6 show the averaged sampled diversities versus reference diversities for the simulated populations. That is, Figure 5 shows the quantity needed for the numerator in Equation (1), and Figure 6 shows the quantity needed for the denominator in Equation (1).



**Figure 5.** Distribution of average sampled  $\bar{D}_{\alpha,\text{samp}}^N$  (mean particle diversity, averaged over the 1000 repeats) and  $D_{\alpha,\text{ref}}$  (reference mean particle diversity) for increasing sample sizes based on the simulated scenario library described in Section 2.3. The one-to-one line is drawn for reference.



**Figure 6.** Distribution of average sampled  $\bar{D}_{\gamma,\text{samp}}^N$  (total diversity, averaged over the 1000 repeats) and  $D_{\gamma,\text{ref}}$  (reference total diversity) for increasing sample sizes based on the simulated scenario library described in Section 2.3. The one-to-one line is drawn for reference.

These show that, on average, sampling *overestimates* the mean particle diversity ( $D_{\alpha}$ ) and *underestimates* the total diversity ( $D_{\gamma}$ ). Said another way, using a sample of particles will lead us to think that the average particle has *more* effective species than are really present, but that the bulk has *less* effective species than it actually does. Because  $\chi$  is the affine ratio of mean particle to total diversity (see Equation (1)),  $\chi$  is thus *overestimated*. Writing these numerical observations in equations, we have

$$\bar{D}_{\alpha,\text{samp}}^N \geq D_{\alpha,\text{ref}},$$

[sampled populations *overestimate* mean particle diversity]

(7)

$$\bar{D}_{\gamma,\text{samp}}^N \leq D_{\gamma,\text{ref}},$$

[sampled populations *underestimate* total diversity]

(8)

$$\bar{\chi}_{\text{samp}}^N \geq \chi_{\text{ref}}.$$

[sampled populations *overestimate* mixing-state index]

(9)

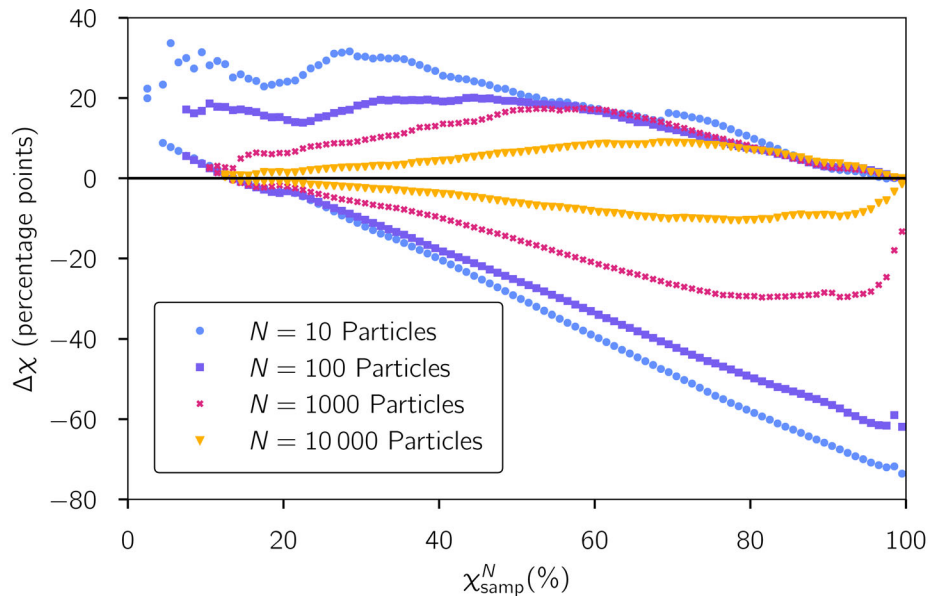
Section 4 provides rigorous definitions of the sampling and averaging procedures and proves certain aspects of the above results.

To understand how many particles we should sample to be confident that the error is likely below some threshold, it is helpful to think about confidence intervals for the reference  $\chi$ . The 95% confidence intervals for reference values,  $\chi_{\text{ref}}$ , are shown in Figure 7 as a range  $\Delta\chi$  about the sampled values  $\chi_{\text{samp}}^N$ . This means that, if we measure a  $\chi_{\text{samp}}^N$  value from a sampled particle population, 95% of the time the true  $\chi_{\text{ref}}$  value will fall within the range  $\chi_{\text{samp}}^N + \Delta\chi$ .

For small sample sizes  $N$ , the confidence interval is highly asymmetric and larger for populations that appear more internally mixed. For example, assuming that a sample of 10 particles is used to compute a sampled  $\chi_{\text{samp}}^{10}$  value of 20%, the 95% confidence interval for the reference  $\chi_{\text{ref}}$  for the whole population extends from 15% to 45% (i.e.,  $\Delta\chi$  ranges from  $-5$  to  $+25$  percentage points). In contrast, for a  $\chi_{\text{samp}}^{10}$  value of 90%, the confidence interval extends from 20% to 95% ( $\Delta\chi$  from  $-70$  to  $+5$  points).

For large sample sizes (e.g.,  $N=10,000$ ), the confidence interval is narrow for large and small sampled  $\chi$  values. It broadens for intermediate  $\chi$  values, but remains within  $\pm 10$  percentage points. It is reasonable that both highly diverse (low  $\chi$ ) and highly homogeneous (high  $\chi$ ) populations can have their mixing state





**Figure 7.** 95%-Confidence intervals for sampled  $\chi_{\text{samp}}^N$  values for sample sizes of  $N = 10, 100, 1000,$  and  $10,000$  particles based on the simulated scenario library described in Section 2.3.

measured well from a reasonably small sample. More complex distributions with intermediate  $\chi$  require the sampling of more particles to obtain accurate estimates of the mixing state index.

By the central limit theorem, we expect that the sampling error should decrease proportionally to the square root of the number of particles. We can observe this in Figure 7, where going from  $N = 1000$  to  $N = 10,000$  particles reduced the 95% confidence interval bound by a factor of 3 (from 30 percentage points to 10 points), which is approximately  $\sqrt{10}$ , the square root of the increase in the number of samples.

Considering that the true  $\chi$  value of a population is not known a priori, our results suggest that a sample size of 1000 particles is needed to obtain an estimate of  $\chi$  within 30 percentage points and 10,000 particles are needed to determine  $\chi$  within 10 percentage points for any mixing state.

### 4. Mathematical proofs

In Section 3 we have seen that  $\bar{\chi}_{\text{samp}}^N$  is positively biased, which was caused by a positive bias in  $\bar{D}_{\alpha, \text{samp}}^N$ , a negative bias in  $\bar{D}_{\gamma, \text{samp}}^N$ , and the fact that  $\chi = (D_\alpha - 1)/(D_\gamma - 1)$ . In this section, we will show that the overestimation of  $D_\alpha$  and underestimation of  $D_\gamma$  are both consequences of the entropy averaging procedures combined with convexity of the exponential function (for  $D_\alpha$ ) and concavity of the entropy function (for  $D_\gamma$ ). To do this, we will start by precisely

defining what we mean by sampling and averaging, and then we will prove the results themselves.

#### 4.1. Notation for sampled particle populations

We consider a *reference population* of particles to be a set  $\pi = \{\vec{\mu}_1, \dots, \vec{\mu}_N\}$ . We use uppercase letters  $I, J$  to denote reference particle indices in  $\pi$ . Each particle is a vector  $\vec{\mu}_I \in \mathbb{R}^A$  with coordinates  $\vec{\mu}_I = (\mu_I^1, \dots, \mu_I^A)$ , where each coordinate  $\mu_I^a$  is the mass of species  $a$  in particle  $I$ . We use superscripts for species indexes and subscripts for particle indexes.

Consider *sampled populations*  $\{\pi_1, \dots, \pi_S\}$  from  $\pi$ . Each sampled population  $\pi_s = \{\vec{\mu}_{s,1}, \dots, \vec{\mu}_{s,N_s}\}$  has  $N_s$  particles corresponding to reference particle indices  $I_{s,1}, \dots, I_{s,N_s}$ . We use lowercase letters  $i, j$  for sampled particle indices and we write  $\vec{\mu}_{s,i}$  for the  $i$ -th particle in sample  $s$ . The mass of species  $a$  in particle  $i$  of sample  $s$  is thus  $\mu_{s,i}^a$ . The sampled particle  $\vec{\mu}_{s,i}$  is equal to reference particle  $\vec{\mu}_I$  with  $I = I_{s,i}$ , so the set of all reference particle indexes in a given sampled population is  $\mathcal{I}_s = \{I_{s,i} | i = 1, \dots, N_s\}$ .

Given a per-particle quantity  $X_I$ , we write  $X_{s,i} = X_I$  when the reference index matches the sampled particle index:  $I = I_{s,i}$ . To understand sampled diversities versus reference-population diversities, we want to compare mass-weighted averages computed over the reference population, and over sampled populations. To do this, we will now introduce the Iverson bracket and the key averaging lemma (Lemma 1).

The *Iverson bracket* gives a binary indicator of set membership:

$$[I \in \mathcal{I}^s] = \begin{cases} 1 & \text{if } I \in \mathcal{I}^s, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Using the Iverson bracket we can translate between local and global indexes:

$$\sum_{i=1}^{N_s} X_{s,i} = \sum_{I=1}^N [I \in \mathcal{I}^s] X_I. \quad (11)$$

In this article we consider particle populations to be sets, which means that all particles in the population must have at least one species with a different mass. In particular, this excludes monodisperse populations. To overcome this limitation we could consider populations to be multisets in the sense of (Knuth 1998, p. 473). Roughly speaking, a multiset is an extension of a set to allow elements to appear multiple times and for which the set union  $\uplus$  and set union  $\setminus$  operators have been appropriately extended. For multisets, the equivalent to the Iverson bracket is the multiplicity operator which gives the integer number of occurrences of any particle in the population. All the theoretical results in this article carry through for multisets, but we restrict ourselves to regular sets for convenience.

#### 4.2. Mass fractions and mass-weighted probability distributions

Given a reference population  $\pi = \{\vec{\mu}_1, \dots, \vec{\mu}_N\}$  and sampled populations  $\pi_s = \{\vec{\mu}_{s,1}, \dots, \vec{\mu}_{s,N_s}\}$  for  $s = 1, \dots, S$ , as described in Section 4.1, we define the total masses:

$$\mu_I = \sum_{a=1}^A \mu_I^a, \quad [\text{total mass of reference particle } I] \quad (12)$$

$$\mu_{\text{tot}} = \sum_{I=1}^N \mu_I, \quad [\text{total mass of reference population}] \quad (13)$$

$$\mu_{s,i} = \sum_{a=1}^A \mu_{s,i}^a, \quad [\text{total mass of particle } i \text{ in sampled populations}] \quad (14)$$

$$\mu_{s,\text{tot}} = \sum_{i=1}^{N_s} \mu_{s,i}, \quad [\text{total mass of sampled populations}] \quad (15)$$

$$\mu_{\text{samp,tot}} = \sum_{s=1}^S \mu_{s,\text{tot}}. \quad [\text{total mass of all sampled populations}] \quad (16)$$

From this we define the mass fractions (or probabilities):

$$p_I = \frac{\mu_I}{\mu_{\text{tot}}}, \quad [\text{mass fraction of reference particle } I] \quad (17)$$

$$p_{s,i} = \frac{\mu_{s,i}}{\mu_{s,\text{tot}}}, \quad [\text{mass fraction of particle } i \text{ in sampled population } s] \quad (18)$$

$$p_{s,\text{tot}} = \frac{\mu_{s,\text{tot}}}{\mu_{\text{samp,tot}}}. \quad [\text{mass fraction of sampled population } s] \quad (19)$$

Interpreting these mass fractions as probabilities gives mass-weighted probability distributions over particle populations. For example, we can define the distribution  $p_I$  to be the distribution over reference particles so that the probability  $I \sim p_I$  of reference particle  $I$  is  $p_I$ . Doing this similarly for  $p_{s,i}$  and  $p_{s,\text{tot}}$  gives us probability distributions over sampled particles and the set of sampled populations. Note that we are using roman-letter subscripts to denote probability distributions.

Consider a quantity  $X$  that can be indexed by either the reference particle index,  $X_I$ , the sampled particle indexes,  $X_{s,i}$ , or the sampled population index,  $X_{s,\text{tot}}$ . Then we can compute expected values with respect to the mass-weighted probability distributions by averaging over the corresponding sets:

$$\mathbb{E}_{I \sim p_I}[X_I] = \sum_{I=1}^N p_I X_I = \sum_{I=1}^N \frac{\mu_I}{\mu_{\text{tot}}} X_I, \quad [\text{reference average}] \quad (20)$$

$$\begin{aligned} \mathbb{E}_{i \sim p_{s,i}}[X_{s,i}] &= \sum_{i=1}^{N_s} p_{s,i} X_{s,i} \\ &= \sum_{i=1}^{N_s} \frac{\mu_{s,i}}{\mu_{s,\text{tot}}} X_{s,i}, \quad [\text{sample average}] \end{aligned} \quad (21)$$

$$\begin{aligned} \mathbb{E}_{s \sim p_{s,\text{tot}}}[X_{s,\text{tot}}] &= \sum_{s=1}^S p_{s,\text{tot}} X_{s,\text{tot}} \\ &= \sum_{s=1}^S \frac{\mu_{s,\text{tot}}}{\mu_{\text{samp,tot}}} X_{s,\text{tot}}. \quad [\text{average over samples}] \end{aligned} \quad (22)$$

#### 4.3. Entropy and diversity measures of mixing state

The entropy  $H$  associated with a vector  $\vec{p}$  of mass fractions (equivalently, probabilities) is

$$H(\vec{p}) = - \sum_{a=1}^A p^a \log(p^a). \quad (23)$$

The diversity  $D$  is the exponential of the entropy:

$$D(\vec{p}) = \exp(H(\vec{p})). \quad (24)$$

Importantly, entropy is a concave function, which is fundamental to our results on over- and under-estimation of

mixing-state measures. In fact, entropy is also log-concave in low dimensions (Alirezaei and Mathar 2018), which will be used in the proof of Theorem 4.

The mass fractions in a particle (equivalently, species probability vector) can be indexed by either reference or sampled particle indexes. These are thus:

$$\vec{p}_I = \frac{\vec{\mu}_I}{\mu_I}, \text{ [species mass fractions in reference particle } I\text{]} \tag{25}$$

$$\vec{p}_{s,i} = \frac{\vec{\mu}_{s,i}}{\mu_{s,i}}. \text{ [species mass fractions in particle } i \text{ in sampled population } s\text{]} \tag{26}$$

Using the particle species mass fraction vectors, we denote the particle entropy by  $H_I = H(\vec{p}_I)$  with reference particle indexes and  $H_{s,i} = H(\vec{p}_{s,i})$  with sampled particle indexes. Similarly, we write  $D_I$  and  $D_{s,i}$  for the corresponding diversities.

The  $\alpha$ -entropy of a particle population is the mass-weighted average of the particle entropies, with the corresponding  $\alpha$ -diversity. That is,  $\alpha$ -entropies and  $\alpha$ -diversities for our populations are

$$\begin{aligned} H_{\alpha,\text{ref}} &= \sum_{I=1}^N p_I H_I \\ &= \mathbb{E}_{I \sim p_I} [H_I], \text{ [}\alpha\text{-entropy of reference population]} \end{aligned} \tag{27}$$

$$\begin{aligned} H_{\alpha,s} &= \sum_{i=1}^{N_s} p_{s,i} H_{s,i} \\ &= \mathbb{E}_{i \sim p_{s,i}} [H_{s,i}], \text{ [}\alpha\text{-entropy of sampled populations]} \end{aligned} \tag{28}$$

$$D_{\alpha,\text{ref}} = \exp(H_{\alpha,\text{ref}}), \text{ [}\alpha\text{-diversity of reference population]} \tag{29}$$

$$D_{\alpha,s} = \exp(H_{\alpha,s}). \text{ [}\alpha\text{-diversity of sampled populations]} \tag{30}$$

The  $\gamma$ -entropy of a population is the entropy of the mass-weighted average composition vector, with the corresponding  $\gamma$ -diversity. That is, while  $\alpha$ -entropy is the average of entropies,  $\gamma$ -entropy is the entropy of the average. This gives:

$$\begin{aligned} H_{\gamma,\text{ref}} &= H\left(\sum_{I=1}^N p_I \vec{p}_I\right) \\ &= H(\mathbb{E}_{I \sim p_I} [\vec{p}_I]), \text{ [}\gamma\text{-entropy of reference population]} \end{aligned} \tag{31}$$

$$\begin{aligned} H_{\gamma,s} &= H\left(\sum_{i=1}^{N_s} p_{s,i} \vec{p}_{s,i}\right) \\ &= H(\mathbb{E}_{i \sim p_{s,i}} [\vec{p}_{s,i}]), \text{ [}\gamma\text{-entropy of sampled populations]} \end{aligned} \tag{32}$$

$$D_{\gamma,\text{ref}} = \exp(H_{\gamma,\text{ref}}), \text{ [}\gamma\text{-diversity of reference population]} \tag{33}$$

$$D_{\gamma,s} = \exp(H_{\gamma,s}). \text{ [}\gamma\text{-diversity of sampled populations]} \tag{34}$$

From the population diversities we define the overall mixing-state index of a population to be the affine ratio:

$$\chi_{\text{ref}} = \frac{D_{\alpha,\text{ref}} - 1}{D_{\gamma,\text{ref}} - 1}, \text{ [mixing-state index of reference population]} \tag{35}$$

$$\chi_s = \frac{D_{\alpha,s} - 1}{D_{\gamma,s} - 1}. \text{ [mixing-state index of sampled populations]} \tag{36}$$

We are particularly interested in the average sampled entropies and diversities, where the mass-weighted average is taken over all sampled populations. This gives

$$\bar{H}_{\alpha,\text{samp}} = E_{s \sim p_s} [H_{\alpha,s}], \text{ [average sampled } \alpha\text{-entropy]} \tag{37}$$

$$\bar{H}_{\gamma,\text{samp}} = E_{s \sim p_s} [H_{\gamma,s}], \text{ [average sampled } \gamma\text{-entropy]} \tag{38}$$

$$\bar{D}_{\alpha,\text{samp}} = E_{s \sim p_s} [D_{\alpha,s}], \text{ [average sampled } \alpha\text{-diversity]} \tag{39}$$

$$\bar{D}_{\gamma,\text{samp}} = E_{s \sim p_s} [D_{\gamma,s}], \text{ [average sampled } \gamma\text{-diversity]} \tag{40}$$

$$\bar{\chi}_{\text{samp}} = E_{s \sim p_s} [\chi_s]. \text{ [average sampled mixing-state index]} \tag{41}$$

#### 4.4. Fair sampling and a fundamental lemma

The sampled populations may all contain the same number of particles, in which case  $N_s$  is independent of  $s$ , or they may be of different sizes. Our theoretical results apply in either case, so long as we assume that the sampled populations sample all particles fairly. To make this precise, we define  $N_I$  to be the number of sampled populations containing particle  $I$ :

$$N_I = \sum_{s=1}^S [I \in \mathcal{I}_s]. \tag{42}$$

Using this, we make our assumption precise as follows.

**Assumption 1 (Uniform sampling).** *Each particle in the population appears in the same number of sampled populations, so  $N_I = N_J$  for any  $I, J = 1, \dots, N$ .*

**Lemma 1.** Given a per-particle quantity  $X$  indexed both by the reference index,  $X_I$ , and sampled population indexes,  $X_{s,i}$ , the mass-weighted average of  $X$  over all particles in all sampled populations is the same as the mass-weighted average over all reference particles:

$$\mathbb{E}_{s \sim p_s, \text{tot}} \left[ \underbrace{\mathbb{E}_{i \sim p_{s,i}} [X_{s,i}]}_{X_{s, \text{tot}}} \right] = \mathbb{E}_{I \sim p_I} [X_I]. \quad (43)$$

**Proof.** Starting from the left hand side (LHS) of (43), we compute:

$$\text{LHS} = \sum_{s=1}^S \frac{\mu_{s, \text{tot}}}{\mu_{\text{samp}, \text{tot}}} \sum_{i=1}^{N_s} \frac{\mu_{s,i}}{\mu_{s, \text{tot}}} X_{s,i} \quad [\text{Equations (21) and (22)}] \quad (44)$$

$$= \frac{1}{\mu_{\text{samp}, \text{tot}}} \sum_{s=1}^S \sum_{i=1}^{N_s} \mu_{s,i} X_{s,i} \quad (45)$$

$$= \frac{1}{\sum_{s=1}^S \sum_{i=1}^{N_s} \mu_{s,i}} \sum_{s=1}^S \sum_{i=1}^{N_s} \mu_{s,i} X_{s,i} \quad [\text{Equations (15) and (16)}] \quad (46)$$

$$= \frac{1}{\sum_{s=1}^S \sum_{j=1}^N [\mathcal{I}_s] \mu_j} \sum_{s=1}^S \sum_{I=1}^N [\mathcal{I}_s] \mu_I X_I \quad [\text{Equation (11)}] \quad (47)$$

$$= \frac{1}{\sum_{j=1}^N \mu_j \sum_{s=1}^S [\mathcal{I}_s]} \sum_{I=1}^N \mu_I X_I \sum_{s=1}^S [\mathcal{I}_s] \quad (48)$$

$$= \frac{1}{\sum_{j=1}^N \mu_j N_j} \sum_{I=1}^N \mu_I X_I N_I \quad [\text{Equation (42)}] \quad (49)$$

$$= \frac{1}{\sum_{j=1}^N \mu_j} \sum_{I=1}^N \mu_I X_I \quad [\text{Assumption 1}] \quad (50)$$

$$= \mathbb{E}_{I \sim p_I} [X_I]. \quad [\text{Equations (13) and (20)}] \quad (51)$$

#### 4.5. Theoretical sampling results for mixing-state measures

We are now ready to prove our main results, which describe how the average sampled entropies and diversities relate to the reference entropies and diversities. To summarize, we will show that

$$\begin{aligned} \bar{H}_{\alpha, \text{samp}} &= H_{\alpha, \text{ref}}, \quad [\text{exactly estimate } \alpha \\ &\quad \text{-entropy from samples}] \end{aligned} \quad (52)$$

$$\begin{aligned} \bar{H}_{\gamma, \text{samp}} &\leq H_{\gamma, \text{ref}}, \quad [\text{underestimate } \gamma \\ &\quad \text{-entropy from samples}] \end{aligned} \quad (53)$$

$$\begin{aligned} \bar{D}_{\alpha, \text{samp}} &\geq D_{\alpha, \text{ref}}, \quad [\text{overestimate } \alpha\text{-diversity from samples}] \\ & \end{aligned} \quad (54)$$

$$\begin{aligned} \bar{D}_{\gamma, \text{samp}} &\leq D_{\gamma, \text{ref}}, \quad [\text{underestimate } \gamma\text{-diversity from samples} \\ &\quad \text{(only with 2 or 3 species)}] \end{aligned} \quad (55)$$

See Figures 5 and 6 for numerical simulations that confirm (54) and (55), respectively.

The mixing-state index  $\chi$  is the affine ratio of  $D_\alpha$  to  $D_\gamma$ , so overestimating  $D_\alpha$  and underestimating  $D_\gamma$  makes it plausible that the average sampled  $\bar{\chi}_{\text{samp}}$  values will overestimate the reference  $\chi_{\text{ref}}$ . As shown in Figure 4, numerical simulations on atmospherically relevant particle populations show that  $\bar{\chi}_{\text{samp}}$  values do indeed overestimate  $\chi_{\text{ref}}$ . However, because  $D_{\alpha,s}$  is correlated with  $D_{\gamma,s}$ , it is not straightforward to prove a precise relationship between  $\bar{\chi}_{\text{samp}}$  and  $\chi_{\text{ref}}$ .

In all of the following results we are using mass-weighted averages, as defined in Section 4.2, which is natural because  $\alpha$ - and  $\gamma$ -entropy are mass-weighted quantities. We begin by showing that mass-weighted averaging results in reference and sampled  $\alpha$ -diversities being equal on average.

**Theorem 1** (Sampled  $\alpha$ -entropy). *If the sampled populations are drawn uniformly from the reference population (Assumption 1) then the average sampled  $\alpha$ -entropy is equal to the reference  $\alpha$ -entropy:*

$$\bar{H}_{\alpha, \text{samp}} = H_{\alpha, \text{ref}}. \quad (56)$$

**Proof.** We compute

$$\bar{H}_{\alpha, \text{samp}} = E_{s \sim p_s} [H_{\alpha, s}] \quad [\text{Definition of } \bar{H}_{\alpha, \text{samp}}] \quad (57)$$

$$= E_{s \sim p_s} [E_{i \sim p_{s,i}} [H(\vec{p}_{s,i})]] \quad [\text{Definition of } H_{\alpha, s}] \quad (58)$$

$$= E_{I \sim p_I} [H(\vec{p}_I)] \quad [\text{Lemma 1}] \quad (59)$$

$$= H_{\alpha, \text{ref}}. \quad [\text{Definition of } H_{\alpha, \text{ref}}] \quad (60)$$

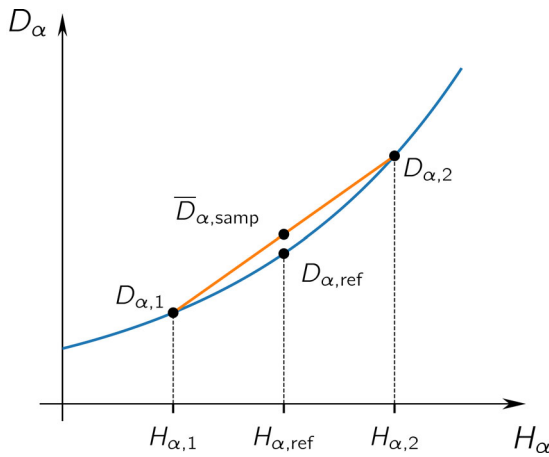
□

Next, we show that concavity of entropy means that  $\gamma$ -entropy is consistently underestimated from sampled populations.

**Theorem 2** (Sampled  $\gamma$ -entropy). *If the sampled populations are drawn uniformly from the reference population (Assumption 1) then the average sampled  $\gamma$ -entropy is less than or equal to the reference  $\gamma$ -entropy:*

$$\bar{H}_{\gamma, \text{samp}} \leq H_{\gamma, \text{ref}}. \quad (61)$$

**Proof.** Similarly to Theorem 1, but with expectations and entropy reversed, we compute



**Figure 8.** Schematic to illustrate Theorem 3. Here we consider two sampled populations with  $\alpha$ -diversities of  $H_{\alpha,1}$  and  $H_{\alpha,2}$ , which we assume have average exactly equal to the reference value  $H_{\alpha,ref}$  (this is true on average, as we saw from Theorem 1). The exponential function maps entropies  $H$  to diversities  $D$ , and because it is convex the average sampled value,  $\bar{D}_{\alpha,samp}$ , will be greater than the reference value,  $D_{\alpha,ref}$ .

$$\bar{H}_{\gamma,samp} = E_{s \sim p_s} [H_{\gamma,s}] \quad [\text{Definition of } \bar{H}_{\gamma,samp}] \quad (62)$$

$$= E_{s \sim p_s} [H(E_{i \sim p_{s,i}} [\vec{p}_{s,i}])] \quad [\text{Definition of } H_{\gamma,s}] \quad (63)$$

$$\leq H(E_{s \sim p_s} [E_{i \sim p_{s,i}} [\vec{p}_{s,i}]]) \quad [\text{Jensen's inequality and concavity of } H] \quad (64)$$

$$= H(E_{I \sim p_I} [\vec{p}_I]) \quad [\text{Lemma 1}] \quad (65)$$

$$= H_{\gamma,ref}. \quad [\text{Definition of } H_{\gamma,ref}] \quad (66)$$

□

Having established the average behavior of sampled entropies, we now turn our attention to sampled diversities. We begin by showing that the  $\alpha$ -diversity is consistently overestimated from sampled populations, as we saw in Figure 5. The reason for this overestimation is that the exponential function is convex, as illustrated in Figure 8.

**Theorem 3** (Sampled  $\alpha$ -diversity). If the sampled populations are drawn uniformly from the reference population (Assumption 1) then the average sampled  $\alpha$ -diversity is greater than or equal to the reference  $\alpha$ -diversity:

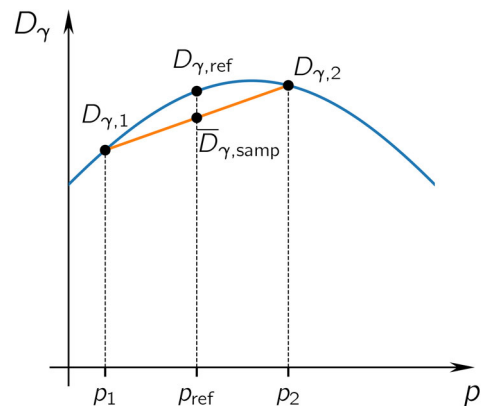
$$\bar{D}_{\alpha,samp} \geq D_{\alpha,ref}. \quad (67)$$

**Proof.** Using Theorem 1 gives

$$\bar{D}_{\alpha,samp} = E_{s \sim p_s} [D_{\alpha,s}] \quad [\text{Definition of } \bar{D}_{\alpha,samp}] \quad (68)$$

$$= E_{s \sim p_s} [\exp(H_{\alpha,s})] \quad [\text{Diversity is the exponential of entropy}] \quad (69)$$

$$\geq \exp(E_{s \sim p_s} [H_{\alpha,s}]) \quad [\text{Jensen's inequality and convexity of exp}] \quad (70)$$



**Figure 9.** Schematic to illustrate Theorem 4 in the case of a two-species aerosol population, where  $p$  is the mass fraction of the first species. We consider two sampled populations with first-species mass-fractions of  $p_1$  and  $p_2$ , which we assume have average exactly equal to the first-species reference mass fraction of  $p_{ref}$  (this is exactly true on average by (43)). The diversity function is concave (for 2 or 3 species) so the average sampled value,  $\bar{D}_{\gamma,samp}$ , will be less than the reference value,  $D_{\gamma,ref}$ .

$$= \exp(H_{\alpha,ref}) \quad [\text{Theorem 1}] \quad (71)$$

$$= D_{\alpha,ref}. \quad [\text{Diversity is the exponential of entropy}] \quad (72)$$

□

Finally, we consider the sampled  $\gamma$ -diversity. Because entropy is only log-concave in dimensions 2 and 3, we are only able to prove a consistent relationship when we have these number of species. As shown in Figure 6, however, even in higher dimensions we see from numerical simulations that sampled populations tend to underestimate  $D_{\gamma}$ . Figure 9 illustrates how concavity of the diversity function leads to underestimation.

**Theorem 4** (Sampled  $\gamma$ -diversity). If the sampled populations are drawn uniformly from the reference population (Assumption 1) and the number of species is 2 or 3, then the average sampled  $\gamma$ -diversity is less than or equal to the reference  $\gamma$ -diversity:

$$\bar{D}_{\gamma,samp} \leq D_{\gamma,ref}. \quad (73)$$

**Proof.** This proof is almost identical to that of Theorem 2, except we use the fact that  $D$  is concave in dimension 2 or 3. Because  $D(\cdot) = \exp(H(\cdot))$ , or equivalently  $H(\cdot) = \log(D(\cdot))$ , concavity of  $D$  is equivalent to log-concavity of  $H$ . As shown in Alirezaei and Mathar (2018, Theorem 16),  $H$  is log-concave if and only if the dimension (number of species) is 2 or 3.



Assuming we have 2 or 3 species and thus concave  $D$ , we compute

$$\bar{D}_{\gamma, \text{samp}} = E_{s \sim p_s} [D_{\gamma, s}] \quad [\text{Definition of } \bar{D}_{\gamma, \text{samp}}] \quad (74)$$

$$= E_{s \sim p_s} \left[ D(E_{I \sim p_{s,i}} [\vec{p}_{s,i}]) \right] \quad [\text{Definition of } D_{\gamma, s}] \quad (75)$$

$$\leq D(E_{s \sim p_s} [E_{I \sim p_{s,i}} [\vec{p}_{s,i}]]) \quad [\text{Jensen's inequality and concavity of } D] \quad (76)$$

$$= D(E_{I \sim p_I} [\vec{p}_I]) \quad [\text{Lemma 1}] \quad (77)$$

$$= D_{\gamma, \text{ref}}. \quad [\text{Definition of } D_{\gamma, \text{ref}}] \quad (78)$$

□

As we see from the above proofs, we now understand the source of the over/under-estimation of the average sampled mixing state index and diversity metrics that we observed from simulations and experimental data in Section 3. The consistent overestimation of  $D_\alpha$  is a consequence of the exact averaging of  $H_\alpha$  (Theorem 1) combined with the convexity of the exponential function (Theorem 3). The underestimation of  $D_\gamma$  (and  $H_\gamma$ ), on the other hand, is due to the concavity of the (log-)entropy function (Theorems 2 and 4). By the central limit theorem, all of these over/under-estimations will decrease at a rate of  $1/\sqrt{N}$  as the number of sampled particles,  $N$ , increases.

## 5. Conclusions

Single-particle instruments necessarily use finite particle samples to determine population-level quantities, with the sample size being determined by practical considerations of data acquisition. In this study we developed a method to determine confidence intervals for a population-level quantity, the mixing-state index  $\chi$ , that is determined from particle-level information. We accomplished this by using model-generated particle populations as a reference, which were sequentially subsampled. Both numerical and mathematical analyses revealed that finite particle samples introduce a positive bias in the estimation of the diversity metric  $D_\alpha$  (the average particle species diversity), and a negative bias in  $D_\gamma$  (the bulk diversity), which overall results in a positive bias in the estimation of  $\chi$ . These results are consistent with the measurements using the aerosol samples of the Ye et al. (2018) study of Pittsburgh. The confidence interval for  $\chi$ , not surprisingly, depends on the mixing state itself.

A sample size of 1000 particles allows an estimate of  $\chi$  within 30 percentage points, and 10,000 particles

are needed to determine  $\chi$  within 10 percentage points, for any mixing state. This approach could be extended to the measurement of other population-level quantities that are estimated based on particle samples. Furthermore, it may be important in practice to consider measurement error (not just undersampling) when calculating  $\chi$ , or to extend  $\chi$  to include species similarity (Leinster and Cobbold 2012) or functional diversity (Scheiner et al. 2017).






## Data availability

The output of the particle-resolved modeling scenario library can be accessed at [https://doi.org/10.13012/B2IDB-2774261\\_V1](https://doi.org/10.13012/B2IDB-2774261_V1).

## Funding

We acknowledge funding from NSF AGS-1254428, NSF AGS-1543786, NSF CHE-1807530, and DOE ASR DE-SC0019192. This publication was developed under assistance agreement RD83587301 awarded by the U.S. Environmental Protection Agency as part of the Center for Air, Climate, and Energy Solutions. It has not been formally reviewed by EPA. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. This research is part of the Blue Waters sustained petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

## ORCID

J. T. Gasparik  <http://orcid.org/0000-0001-6335-8753>  
 Q. Ye  <http://orcid.org/0000-0003-3797-8988>  
 J. H. Curtis  <http://orcid.org/0000-0002-1447-2127>  
 A. A. Presto  <http://orcid.org/0000-0002-9156-1094>  
 N. M. Donahue  <http://orcid.org/0000-0003-3054-2364>  
 R. C. Sullivan  <http://orcid.org/0000-0003-0701-7158>  
 M. West  <http://orcid.org/0000-0002-7605-0050>  
 N. Riemer  <http://orcid.org/0000-0002-3220-3457>

## References

- Alirezai, G., and R. Mathar. 2018. On exponentially concave functions and their impact in information theory. In *Proceedings of the 2018 Information Theory and Applications Workshop (ITA)*. doi:10.1109/ITA.2018.8503202.
- Beck, J., and W. Schwanghart. 2010. Comparing measures of species diversity from incomplete inventories: An update. *Methods Ecol. Evol.* 1 (1):38–44. 00003.x. doi:10.1111/j.2041-210X.2009.

- Beck, J., J. D. Holloway, and W. Schwanghart. 2013. Undersampling and the measurement of beta diversity. *Methods Ecol. Evol.* 4 (4):370–82. doi:10.1111/2041-210X.12023.
- Beydoun, H., M. Polen, and R. C. Sullivan. 2017. A new multicomponent heterogeneous ice nucleation model and its application to snomax bacterial particles and a snomax–illite mineral particle mixture. *Atmos. Chem. Phys.* 17 (22):13545–57. doi:10.5194/acp-17-13545-2017.
- Bondy, A. L., D. Bonanno, R. C. Moffet, B. Wang, A. Laskin, and A. P. Ault. 2018. The diverse chemical mixing state of aerosol particles in the southeastern United States. *Atmos. Chem. Phys.* 18 (16):12595–612. doi:10.5194/acp-18-12595-2018.
- Brocklehurst, N., M. O. Day, and J. Fröbisch. 2018. Accounting for differences in species frequency distributions when calculating beta diversity in the fossil record. *Methods Ecol. Evol.* 9 (6):1409–20. doi:10.1111/2041-210X.13007.
- Butturi-Gomes, D., M. Petrere, H. C. Giacomini, and S. S. Zocchi. 2017. Statistical performance of a multicomparison method for generalized species diversity indices under realistic empirical scenarios. *Ecol. Indic.* 72:545–52. doi:10.1016/j.ecolind.2016.08.054.
- Chao, A., and L. Jost. 2015. Estimating diversity and entropy profiles via discovery rates of new species. *Methods Ecol. Evol.* 6 (8):873–82. doi:10.1111/2041-210X.12349.
- Chao, A., Y. T. Wang, and L. Jost. 2013. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* 4 (11):1091–100. doi:10.1111/2041-210X.12108.
- Ching, J., J. Fast, M. West, and N. Riemer. 2017. Metrics to quantify the importance of mixing state for CCN activity. *Atmos. Chem. Phys.* 17 (12):7445–58. doi:10.5194/acp-17-7445-2017.
- Ching, J., N. Riemer, and M. West. 2016. Black carbon mixing state impacts on cloud microphysical properties: Effects of aerosol plume and environmental conditions. *J. Geophys. Res. Atmos.* 121 (10):5990–6013. doi:10.1002/2016JD024851.
- Daly, A. J., J. M. Baetens, and B. D. Baets. 2018. Ecological diversity: Measuring the unmeasurable. *Mathematics* 6 (7):119. doi:10.3390/math6070.
- Dickau, M., J. Olfert, M. E. J. Stettler, A. Boies, A. Momenimovahed, K. Thomson, G. Smallwood, and M. Johnson. 2016. Methodology for quantifying the volatile mixing state of an aerosol. *Aerosol. Sci. Technol.* 50 (8):759–72. doi:10.1080/02786826.2016.1185509.
- Fraund, M., D. Q. Pham, D. Bonanno, T. H. Harder, B. Wang, J. Brito, S. S. De Sá, S. Carbone, S. China, P. Artaxo, et al. 2017. Elemental mixing state of aerosol particles collected in Central Amazonia during GoAmazon2014/15. *Atmosphere* 8 (12):173. doi:10.3390/atmos8090173.
- Gasparik, J. T., Q. Ye, J. H. Curtis, A. A. Presto, N. M. Donahue, R. C. Sullivan, M. West, and N. Riemer. 2020. Data from: Quantifying errors in the aerosol mixing-state index based on limited particle sample size [dataset]. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-2774261\_V1.
- Haegeman, B., J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, and J. S. Weitz. 2013. Robust estimation of microbial diversity in theory and in practice. *ISME J.* 7 (6):1092–101. doi:10.1038/ismej.2013.10.
- Healy, R. M., N. Riemer, J. C. Wenger, M. Murphy, M. West, L. Poulain, A. Wiedensohler, I. P. O'Connor, E. McGillicuddy, J. R. Sodeau, et al. 2014. Single particle diversity and mixing state measurements. *Atmos. Chem. Phys.* 14 (12):6289–99. doi:10.5194/acp-14-6289-2014.
- Healy, R., J. Sciare, L. Poulain, M. Crippa, A. Wiedensohler, A. Prévôt, U. Baltensperger, R. Sarda-Estève, M. McGuire, C.-H. Jeong, et al. 2013. Quantitative determination of carbonaceous particle mixing state in Paris using single-particle mass spectrometer and aerosol mass spectrometer measurements. *Atmos. Chem. Phys.* 13 (18):9479–96. doi:10.5194/acp-13-9479-2013.
- Hughes, M., J. K. Kodros, J. R. Pierce, M. West, and N. Riemer. 2018. Machine learning to predict the global distribution of aerosol mixing state metrics. *Atmosphere* 9 (1):15. doi:10.3390/atmos9010015.
- Knopf, D. A., and P. A. Alpert. 2013. A water activity based model of heterogeneous ice nucleation kinetics for freezing of water and aqueous solution droplets. *Faraday Discuss.* 165:513–34. doi:10.1039/c3fd00035d.
- Knuth, D. E. 1998. *The art of computer programming, volume 2: Seminumerical Algorithms*. 3rd ed. Boston, MA: Addison Wesley.
- Leinster, T., and C. A. Cobbold. 2012. Measuring diversity: The importance of species similarity. *Ecology* 93 (3):477–89. doi:10.1890/10-2402.1.
- Lesins, G., P. Chylek, and U. Lohmann. 2002. A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing. *J. Geophys. Res.* 107 (D10):AAC 5-1–4106. doi:10.1029/2001JD000973.
- Marion, Z. H. 2016. On the quantification of complexity and diversity from phenotypes to ecosystems. Ph.D. thesis, University of Tennessee.
- Marion, Z. H., J. A. Fordyce, and B. M. Fitzpatrick. 2018. A hierarchical Bayesian model to incorporate uncertainty into methods for diversity partitioning. *Ecology* 99 (4):947–56. doi:10.1002/ecy.2174.
- Moffet, R. C., and K. A. Prather. 2009. In-situ measurements of the mixing state and optical properties of soot with implications for radiative forcing estimates. *Proc. Natl. Acad. Sci. USA.* 106 (29):11872–7. doi:10.1073/pnas.0900040106.
- O'Brien, R. E., B. Wang, A. Laskin, N. Riemer, M. West, Q. Zhang, Y. Sun, X. -Y. Yu, P. Alpert, D. A. Knopf, et al. 2015. Chemical imaging of ambient aerosol particles: Observational constraints on mixing state parameterization. *J. Geophys. Res. Atmos.* 120 (18):9591–605. doi:10.1002/2015JD023480.
- Prather, K. A., C. D. Hatch, and V. H. Grassian. 2008. Analysis of atmospheric aerosols. *Annu Rev Anal Chem (Palo Alto Calif)* 1:485–514. doi:10.1146/annurev.anchem.1.031207.113030.
- Riemer, N., A. Ault, M. West, R. Craig, and J. Curtis. 2019. Aerosol mixing state: Measurements, modeling, and impacts. *Rev. Geophys.* 57 (2):187–249. doi:10.1029/2018RG000615.
- Riemer, N., and M. West. 2013. Quantifying aerosol mixing state with entropy and diversity measurements. *Atmos.*

- Chem. Phys.* 13 (22):11423–39. doi:10.5194/acp-13-11423-2013.
- Riemer, N., M. West, R. Zaveri, and R. Easter. 2009. Simulating the evolution of soot mixing state with a particle-resolved aerosol model. *J. Geophys. Res.* 114 (D9): D09202. doi:10.1029/2008JD011073.
- Ryder, O. S., A. P. Ault, J. F. Cahill, T. L. Guasco, T. P. Riedel, L. A. Cuadra-Rodriguez, C. J. Gaston, E. Fitzgerald, C. Lee, K. A. Prather, et al. 2014. On the role of particle inorganic mixing state in the reactive uptake of N<sub>2</sub>O<sub>5</sub> to ambient aerosol particles. *Environ. Sci. Technol.* 48 (3):1618–27. doi:10.1021/es4042622.
- Scheiner, S. M., E. Kosman, S. J. Presley, and M. R. Willig. 2017. Decomposing functional diversity. *Methods Ecol. Evol.* 8 (7):809–20. doi:10.1111/2041-210X.12696.
- Schell, B., I. J. Ackermann, H. Hass, F. S. Binkowski, and A. Ebel. 2001. Modeling the formation of secondary organic aerosol within a comprehensive air quality modeling system. *J. Geophys. Res.* 106 (D22):28275–93. doi:10.1029/2001JD000384.
- Sherwin, W. B., and N. Prat I Fornells. 2019. The introduction of entropy and information methods to ecology by Ramon Margalef. *Entropy* 21 (8):794. doi:10.3390/e21080794.
- Sullivan, R., M. Moore, M. Petters, S. Kreidenweis, G. Roberts, and K. Prather. 2009. Effect of chemical mixing state on the hygroscopicity and cloud nucleation properties of calcium mineral dust particles. *Atmos. Chem. Phys.* 9 (10):3303–16. doi:10.5194/acp-9-3303-2009.
- Winkler, P. 1973. The growth of atmospheric aerosol particles as a function of the relative humidity-II. An improved concept of mixed nuclei. *J. Aerosol Sci.* 4 (5): 373–87. doi:10.1016/0021-8502(73)90027-X.
- Ye, Q., P. Gu, H. Z. Li, E. S. Robinson, E. Lipsky, C. Kaltsonoudis, A. K. Y. Lee, J. S. Apte, A. L. Robinson, R. C. Sullivan, et al. 2018. Spatial variability of sources and mixing state of atmospheric particles in a metropolitan area. *Environ. Sci. Technol.* 52 (12):6807–15., doi:10.1021/acs.est8b01011.
- Zaveri, R., and L. Peters. 1999. A new lumped structure photochemical mechanism for large-scale applications. *J. Geophys. Res.* 104 (D23):30387–415. doi:10.1029/1999JD900876.
- Zaveri, R., R. Easter, and A. Wexler. 2005b. A new method for multicomponent activity coefficients of electrolytes in aqueous atmospheric aerosols. *J. Geophys. Res.* 110 (D2), D02201. doi:10.1029/2004JD004681.
- Zaveri, R., R. Easter, and L. Peters. 2005a. A computationally efficient multicomponent equilibrium solver for aerosols (MESA). *J. Geophys. Res.* 110 (D24), D24203. doi:10.1029/2004JD005618.
- Zaveri, R., R. Easter, J. Fast, and L. Peters. 2008. Model for simulating aerosol interactions and chemistry (MOSAIC). *J. Geophys. Res.* 113 (D13), D13204. doi:10.1029/2007JD008782.