Machine Learning Model Evaluation for Estimating Submicron Aerosol Mixing State at the Global Scale

Reetahan Mukhopadhyay, Zhonghua Zheng, Matthew West, Robert Healy, Laurent Poulain, Valérie Gros, Nicole Riemer

September 2022

Abstract. This study evaluates a machine learning model's performance on predicting aerosol mixing state, using observations from a field campaign in Paris, France, and training data from particle-resolved aerosol model simulations. Mixing state is challenging to represent in current aerosol models due to the required computational cost. Previous studies have demonstrated that it is possible to learn emulators for mixing state metrics, which is an efficient way of adding information content to models that do not carry detailed mixing state information. However, to date, these emulators have not been validated with observational data. Our observational dataset is from the MEGAPOLI campaign in Paris, France, containing single-particle data which was used to quantify mixing state, along with many other measurements of gas and aerosol concentrations and meteorological conditions. We created the training data by simulating a large scenario library with PartMC-MOSAIC, a particle-resolved Monte Carlo model for aerosol simulation. We built machine learning models using both the extreme gradient boosting (XGBoost) and automated machine learning (AutoML) frameworks, using the simulated training data followed by testing on the observations. Finally, we performed sensitivity analyses on our models, quantifying the impact of diurnal temperature cycles in the training data, the size of the training data set, and the feature choice. We achieved a 7.64% RMSE and a 10.75% MAPE with the XGBoost model, and overall, gained confidence that it is possible to develop emulators for predicting aerosol mixing state under real-world conditions.

Keywords: aerosol mixing state, machine learning, evaluation
Submitted to: *Environ. Res. Lett.*

1. Introduction

The composition of atmospheric aerosol particles constantly evolves during their transport in the atmosphere due to a variety of particle-level processes including gas-particle partitioning, coagulation, and cloud-processing. Field observations have shown that even within a narrow particle size range, differences in particle composition exist (Prather et al., 2008), and as a result, the aerosol bulk composition is usually not enough to characterize the physical and chemical properties of an aerosol (Riemer et al., 2019).

The distribution of chemical species within the particle population is called the "aerosol mixing state" (Winkler, 1973). The edge cases are the so-called internal

mixture, where all particles have the same composition and the external mixture, where different aerosol species reside in different particles. While the internal and external mixtures are useful idealized concepts, the aerosol mixing state in the ambient atmosphere is frequently somewhere in between these two extremes. Simplifying assumptions about the true aerosol mixing state can introduce errors in predictions of aerosol optical properties and cloud condensation nuclei concentrations, as many closure studies have shown (Cubison et al., 2008; Ervens et al., 2010), and these errors then propagate in predictions of radiative forcing (Fierce et al., 2016, 2017).

Aerosol modules in many state-of-the-art global models represent mixing state to a certain extent. For example, the MATRIX model (Bauer et al., 2008) uses 16 overlapping log-normal modes that differ in their composition, distinguishing between different degrees of mixing of primary and secondary aerosol species. Similarly, but simpler, the MAM4 model (Liu et al., 2016) represents the aerosol by four modes, focusing on the representation of black carbon aging. The choice of modes for any particular modal aerosol model and the extent to which the modes interact introduces structural uncertainty to these approaches.

Our study is motivated by work by Zheng et al. (2021a) and Zheng et al. (2021b), which aimed at developing a framework to quantify this structural uncertainty. The objective of our study is to provide validation of this framework using observational data. The framework is based on two principles. The first principle is the use of a metric that quantifies aerosol mixing state. This scalar quantity called χ (Riemer and West, 2013) is calculated using the particles' species mass fractions. The values of χ range between 0% and 100% for completely external and internal mixtures, respectively. The particle-based composition information to calculate χ can be obtained from particle-resolved model simulations, however, these are computationally too expensive to be performed for the entire globe and for simulation times of several months or years. Therefore, the second principle of the framework is the development of a machine-learned emulator of χ that is trained on particle-resolved box model data and uses global model output data as features.

Using this framework, Zheng et al. (2021a) constructed spatial distributions of the mixing state metric χ using global model output from the Community Earth System Model (CESM). Zheng et al. (2021b) compared the spatial distributions of χ obtained with this approach with predictions of the MAM4 model and found that the two methods yielded very different spatial patterns of the mixing state indices. In some regions, the yearly-averaged mixing state index computed by the MAM4 model differed by up to 70 percentage points from the machine-learned model. These errors tended to be zonally structured, with the MAM4 model predicting a more internally mixed aerosol at low latitudes, and a more externally mixed aerosol at high latitudes.

While these prior studies provided first insights into structural uncertainties of modal aerosol models, a necessary next step is to validate the machine-learning modeling framework for χ. This is the goal of our work in this paper. We used observational data collected during the MEGAPOLI campaign that was conducted in Paris, France,

during January and February 2010. Using single-particle data from an aerosol-time-of-flight mass spectrometer and simultaneously-measured other data products, Healy et al. (2014) determined the mixing state index χ for a site in Paris for the duration of the campaign. Here we investigate if we can develop a machine-learned model that estimates χ during MEGAPOLI, using PartMC-MOSAIC simulations for generating training data and observed variables during the campaign as features (predictors).

The remainder of the paper is structured as follows: Section 2 describes observational data, the development of the training data and the machine-learned model for χ. Section 3 presents the results and discusses the sensitivity of the machine-learned model to a range of model assumptions, including the choice of the ML model, the size of the training dataset, and the feature selection. Section 4 concludes the paper.

2. Methodology

2.1. Observational Data

We used observational data from the MEGAPOLI winter campaign (Healy et al., 2012, 2013), collected at the Laboratoire d'Hygiéne de la Ville de Paris (LHVP), Paris, France from 15 January–11 February 2010. This dataset is suitable because the mixing state metric χ was determined for each hour of the measurement period using an aerosol time-of-flight mass spectrometer (ATOFMS, TSI model 3800) (Healy et al., 2014), and the dataset also provides measurements of many bulk aerosol and gas phase species, which are needed as features for our machine-learning model.

We used data from a multi-angle absorption photometer for black carbon (BC) mass concentrations (MAAP, Model 5012, Thermo Scientific) (Petzold and Schönlinner, 2004) and from a high-resolution time-of-flight aerosol mass spectrometer (HR-ToF-AMS, Aerodyne Research Inc.) (DeCarlo et al., 2006) for non-refractory aerosol species including ammonium, nitrate, sulfate and organic aerosol. We also used gas phase mixing ratios of CO, O_3, NO, NO_x, and a number of available VOC species (ethane, ethylene, propane, propene, isobutane, n-butane, isopentane, n-pentane, n-hexane, benzene, toluene, xylenes + C8, methanol, acetaldehyde, acetone) (Baudic et al., 2016). We mapped the VOC species to the carbon bond mechanism model species used in PartMC-MOSAIC (CBM-Z) as shown in Table 1. Temperature and relative humidity data measured at the site were also utilized as predictors.

2.2. Mixing state metric calculations

The mixing state index χ measures where an aerosol population is on the continuum of external to internal mixing, that is, how "spread out" the chemical species are across an aerosol population (Riemer and West, 2013). It varies between 0% for a completely external mixture and 100% for a completely internal mixture. As observations show, χ values in the ambient atmosphere range between these two extremes and show characteristic temporal (Healy et al., 2014) and spatial (Ye et al., 2018) variability.

Table 1: Mapping of measured VOC species to model species in CBM-Z mechanism

measured VOC species	model species in CBM-Z
ethane	C2H6
ethylene	ETH
propane	3PAR
propene	OLET+PAR
isobutane, n-butane	4PAR
isopentane, n-pentane	5PAR
n-hexane	6PAR
benzene, toulene	TOL
xylenes+C8	XYL
methanol	CH3OH
acetaldehyde	ALD2
acetone	AONE

Here, we focus on the mixing state of sub-micron aerosols, since the ATOFMS, which is used to observationally determine χ, is limited to this size range.

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

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

Following are the calculations for the diversities D_α and D_γ. First, the per-particle mixing entropies H_i are calculated for each particle by

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

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

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

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

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

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

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

where p^a is the bulk mass fraction of species a in the population. More details on mixing state index calculations can be found at Riemer and West (2013).

2.3. PartMC description

PartMC-MOSAIC (Particle Monte Carlo model—Model for Simulating Aerosol Interactions and Chemistry) (Riemer et al., 2009; Zaveri et al., 2008) is a stochastic particle-resolved aerosol model. A comprehensive description of the model and algorithms can be found in Riemer et al. (2009) and DeVille et al. (2011, 2019) for PartMC, and in Zaveri et al. (2008) for MOSAIC.

The Lagrangian box model PartMC represents the evolution aerosol particles in a fully-mixed computational volume. We simulate the processes of emission, coagulation and dilution stochastically. Gas-phase chemistry and gas–aerosol partitioning are represented deterministically using the MOSAIC model, which includes the carbon-bond-based mechanism CBM-Z for gas-phase photochemical reactions (Zaveri and Peters, 1999), the multicomponent Taylor expansion method (MTEM) for calculating electrolyte activity coefficients in aqueous inorganic mixtures, and the multicomponent equilibrium solver for aerosols (MESA) for calculating the phase states of the particles (Zaveri et al., 2005). The formation of secondary organic aerosol (SOA) is represented by the Secondary Organic Aerosol Model (SORGAM) (Schell et al., 2001).

Since the particle-resolved approach does not make any a priori assumptions about aerosol mixing state, PartMC-MOSAIC can be considered as a benchmark model with respect to the mixing state representation, and we use it here to generate our training and testing data for developing the machine learning model.

2.4. Training Data Generation Process

We followed the procedure described in Zheng et al. (2021a) to generate training and testing data. This entailed creating an ensemble of particle-resolved model scenarios using PartMC-MOSAIC. A "scenario" here is considered the simulation of the evolution of the aerosol phase and gas phase over a 24-hour period given one particular sampling of the input parameters. To create a data set of aerosol populations that encompass a wide range of mixing states, we varied the input parameters for the simulations, including primary emissions of different aerosol types (e.g., carbonaceous aerosol, sea salt, and dust emissions, including contributions from Aitken mode, accumulation mode, and coarse mode size ranges), primary emissions of gas phase species (e.g., SO_2, NO_2, CO, and various volatile organic compounds), and meteorological parameters. The set of parameters and their ranges are listed in Table 2. The parameter combination for each scenario was determined by Latin Hypercube Sampling. For each scenario we used 10,000 computational particles to resolve the aerosol population. For the base case training and testing data used to generate our primary result in Section 3.2, we used a total of 1000 scenarios, composed of 900 scenarios to train and through random selection, held out 100 for testing. Following postprocessing, each scenario contributes to 24 rows of training data and χ labels.

In contrast to the setup used in Zheng et al. (2021a), where the temperature remained constant over the day for simplicity, here we added diurnal temperature

variations. We used hourly ERA5 Reanalysis data at 0.25° by 0.25° resolution from 2009 to 2020 (Hersbach et al., 2021) and sampled longitudes so that 2/3 of the locations were over the ocean and 1/3 over land. With the sampled latitude and longitude, the day of the year, and the year from 2009 to 2020, we searched the ERA5 Reanalysis dataset for the location and date to determine the corresponding temperature profile. The impact of this improvement to the training dataset generation will be evaluated in section 3.3.

2.5. Model Training and Testing

Our machine learning model was built upon the extreme gradient boosting (XGBoost) regression framework (Chen and Guestrin, 2016). The conceptual basis behind XGBoost regression is implementing the ensemble machine learning algorithms known as gradient boosting in an efficient manner.

Boosting refers to the method of iteratively adding weak learners and having them learn, to be added to a final, "strong learner". The higher the accuracy for a weak learner, the higher the weight its predictions are assigned in the final regression model. Additionally, there is weighting on the data itself: when a weak learner is added to the final model, a re-weighting process is applied such that the more mispredicted data is re-assigned to have a higher weight, so the future weak classifiers will know to "focus" on that data. The choice of "weak" learner in our case is a regression decision tree, which is a simple machine learning model that tries to classify data by checking a sequence of properties of the data in a tree-like fashion—in our case, employed for a regression task. Gradient boosting takes this notion of boosting and turns it into a numerical optimization problem, where the step of adding weak classifiers in a fashion to minimize a loss function, which is done with the process of gradient descent. We employ Bayesian Hyperparameter Tuning (Snoek et al., 2012) to deem the optimal hyperparameters for the model. Hyperparameters are parameters typically not learned by the training process, but require manual adjustment. This method essentially builds a surrogate loss function based on the conditional probability of a score given a set of hyperparameters. It finds a set of hyperparameters to optimize this, uses those hyperparameters on the original loss function, updates the surrogate using the results, and it iterates.

Additionally, we explored the Fast Lightweight Automated Machine Learning (FLAML), an implementation of the automated machine learning (AutoML) framework for modeling (Wang et al., 2021). Both AutoML (Sun et al., 2021) and XGBoost (Zamani Joharestani et al., 2019; Ma et al., 2020) have precedented use as machine learning frameworks of choice in the geosciences community. An automated machine learning framework aims to automate the pipeline of steps typically required for machine learning, provided the data, a task, and an error metric. In our use, the framework notably takes care of hyperparameter tuning and model search. It uses cost-frugal optimization (CFO), a method built atop FLOW^2, which is a randomized direct

Table 2: List of input parameters and their sampling ranges to construct the training and testing scenarios. The variables D_g, σ_g, E_a refer to geometric mean diameter, geometric standard deviation, and number emission flux, respectively.

Parameters	**Range**
Environmental Variable	
Relative humidity (RH)	[0.1, 1) or [0.4, 1)
Latitude	(70°S, 70°N) or (90°S, 90°N)
Day of Year	[1, 365]
Temperature	Varies with time of day and location‡
Gas Phase Emissions Scaling Factor	
SO_2, NO_2, NO, NH_3, CO, CH_3OH, ALD2 (Acetaldehyde), ANOL (Ethanol), AONE (Acetone), DMS (Dimethyl sulfide), ETH (Ethene), HCHO (Formaldehyde), ISOP (Isoprene), OLEI (Internal olefin carbons), OLET (Terminal olefin carbons), PAR (Paraffin carbon), TOL (Toluene), XYL (Xylene)	[0, 200%]
Carbonaceous Aerosol Emissions (one mode)	
D_g	[25 nm, 250 nm]
σ_g	[1.4, 2.5]
BC/OC mass ratio	[0, 100%]
E_a	[0, 1.6×10^7 m^{-2} s^{-1}]
Sea Salt Emissions (two modes)	
$D_{g,1}$	[180 nm, 720 nm]
$\sigma_{g,1}$	[1.4, 2.5]
$E_{a,1}$	[0, 1.69×10^5 m^{-2} s^{-1}]
$D_{g,2}$	[1 μm, 6 μm]
$\sigma_{g,2}$	[1.4, 2.5]
$E_{a,2}$	[0, 2380 m^{-2} s^{-1}]
OC fraction	[0, 20%]
Dust Emissions (two modes)	
$D_{g,1}$	[80 nm, 320 nm]
$\sigma_{g,1}$	[1.4, 2.5]
$E_{a,1}$	[0, 5.86×10^5 m^{-2} s^{-1}]
$D_{g,2}$	[1 μm, 6 μm]
$\sigma_{g,2}$	[1.4, 2.5]
$E_{a,2}$	[0, 2380 m^{-2} s^{-1}]
Restart Timestamp	
Timestamp	[0, 24 hours]

[1] See Section 2.4 to see how this is implemented

search method (Wu et al. (2021)), which can systematically pass through the space of hyperparameters, and also tries out options for learners, provided a preset list of types of candidate learners. These include their own implementation of XGBoost, but also learners like Extra Trees, LGBM, etc. The process for tuning involves a randomized search and a re-sampling process.

We trained on our generated training data, using the data of all the features as the input, and their associated χ values as the labels. The features were: ozone (O_3), temperature (T), relative humidity (RH), carbon monoxide (CO), nitrogen oxides (NO_x), nitrous oxide (NO), ammonium (NH_4), sulfate (SO_4), nitrate (NO_3), black carbon (BC), organic aerosols (OA), xylene (XYL), ethene (ETH), methanol (CH_3OH), ethane (C_2H_6), acetone (AONE), toluene (TOL), paraffin carbon (PAR), internal olefin carbons (OLET), and acetaldehyde (ALD2). Following training, we evaluated the model using the testing data by making predictions upon χ. Finally, we used the model to make predictions on the observed data, taking in the observed values of the features at each timestamp and outputting a prediction for the χ value at that timestamp, which we then compared to the χ values from observations.

3. Results and Discussion

We present model performance on the testing dataset in Section 3.1, under what we will refer to as "base case" conditions. These conditions include that a diurnal cycle is applied to the temperature in the scenario simulations, using 900 scenarios for the training process and 100 for testing, and using all aforementioned features. We present the main results of our study in Section 3.2, which evaluates model performance on the observed dataset, under base case conditions with one addendum–the predictions of χ were restricted to the marine period (i.e., the period when the air mass observed at the station came from the ocean as opposed from the continent). We will justify this choice in Section **??** below. In addition to the base case results, we present sensitivity analyses in Section 3.3 where we explored the effects of (1) using constant temperature scenarios for training data, (2) varying the number of scenarios used for training, and (3) dropping features from the model.

Note that AutoML itself is not a model but is a framework that develops models and includes a step for selecting types of models. Thus, while we henceforth refer to the model generated as "AutoML", the true model is what AutoML selected, which we note was lightweight gradient boosting (LGBM) for both the base case and all the sensitivity analyses of AutoML generating a model. LGBM is similar to XGBoost, but some key differences include that LGBM has a leaf-wise growth of its tree learners while XGBoost is level-wise. XGBoost uses a simple histogram-based approach to getting optimal splits, while LGBM uses Gradient-based One Sided Sampling (GOSS).

3.1. Model performance of XGBoost and AutoML models

Figure 1 presents the model performance of the XGBoost model and the AutoML model. Each data point represents a pair (χ_{ref}, χ_{pred}). Our 100 testing scenarios, which each simulate a 24-hour day, correspond to 2400 testing points. With this much data, we present the testing performance as density heatmaps for clarity. Ideally, the points should be concentrated around the identity ($y = x$) line, indicating that the model is adept at predicting the held out simulated testing χ_{ref} values given the simulated testing feature data, following training of the model.

We chose root mean squared error (RMSE) and mean absolute percentage error (MAPE) as model performance metrics as they are common and widely used in regression studies (de Myttenaere et al., 2017; Dua et al., 2016). Note that the units for MAPE are always a percentage, however, the units for RMSE take the units of the quantity of focus. In our case, this is χ, which is reported in percentage. Therefore, comparing RMSE values considers percentage points, while comparing MAPE values compares percentage errors. As given in Table 3 we obtain a RMSE of 6.88% and 6.21% percentage points for XGBoost and AutoML, respectively, and a MAPE of 7.33% and 5.65%.

In both figures we see the expected result, indicating that both the XGBoost and AutoML models were capable of predicting their testing data, which is a prerequisite before considering applying them to the observed data. AutoML is mainly concentrated around this line, compared to XGBoost, but the XGBoost testing performance is still satisfactory.

3.2. Base case results

Figure 2 presents the time series over the entire time period comparing the model predictions with observed data. During the period between January 28, 2010 at 12:00AM local time and February 7, 2010 at 12:00AM local time the air mass present over the observation site originated from over the ocean, while before and after this time period, the air came from continental Europe (Healy et al., 2014). While the overall model performance is fairly satisfactory over the entire time period, there is an evident degradation in performance for the continental period. The RMSE and MAE under XGBoost for the continental time period was 11.44% and 14.93%, respectively; the RMSE and MAPE for AutoML landed at 8.32% and 11.13%, respectively. Meanwhile, the marine period achieved an RMSE and MAPE of 7.64% and 10.75%, respectively for XGBoost, and 6.29% and 9.09% for AutoML.

This indicates that the model's performance is influenced by the source of the air mass in the area. A possible reason for this behavior is that during the marine period the observations at the site are dominated by local pollution. In contrast, during the continental period, aged aerosol originating from long-range transport is present at the site in addition to local pollution. However, our ML model only uses local features to predict the aerosol mixing state at the site, and therefore we cannot expect that

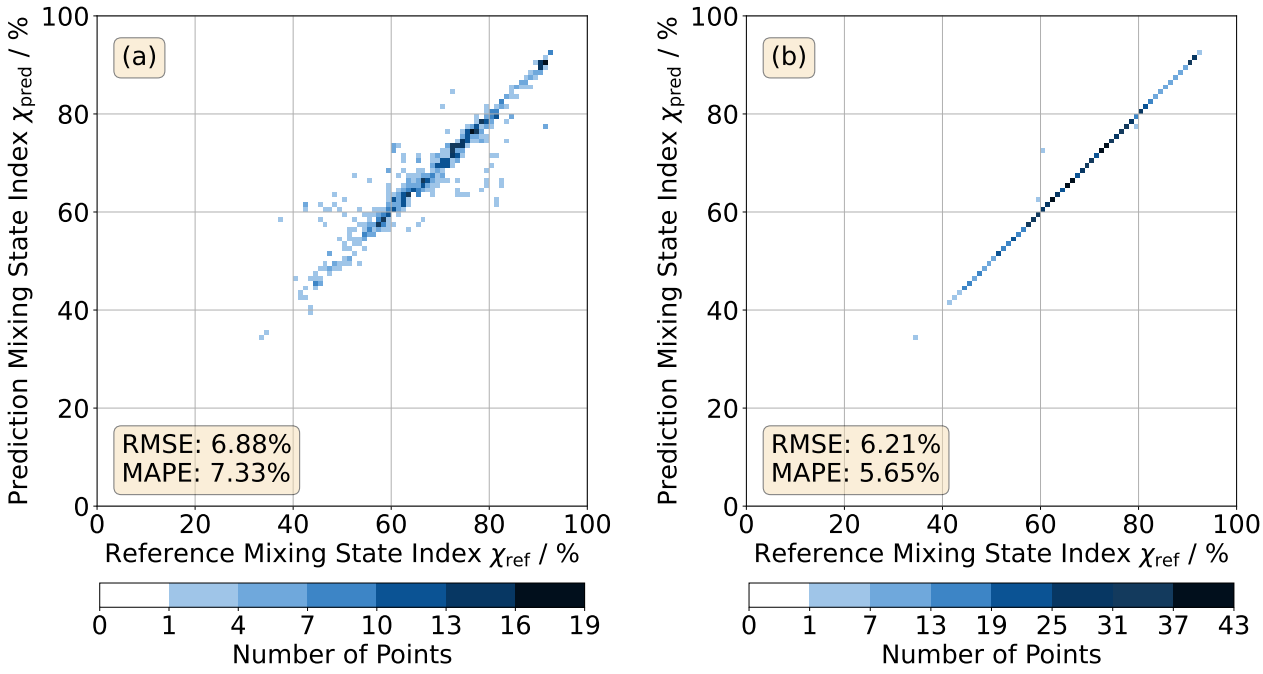

Figure 1: 2-D histogram of the (a) XGBoost model and the (b) AutoML model predicting χ on the testing dataset

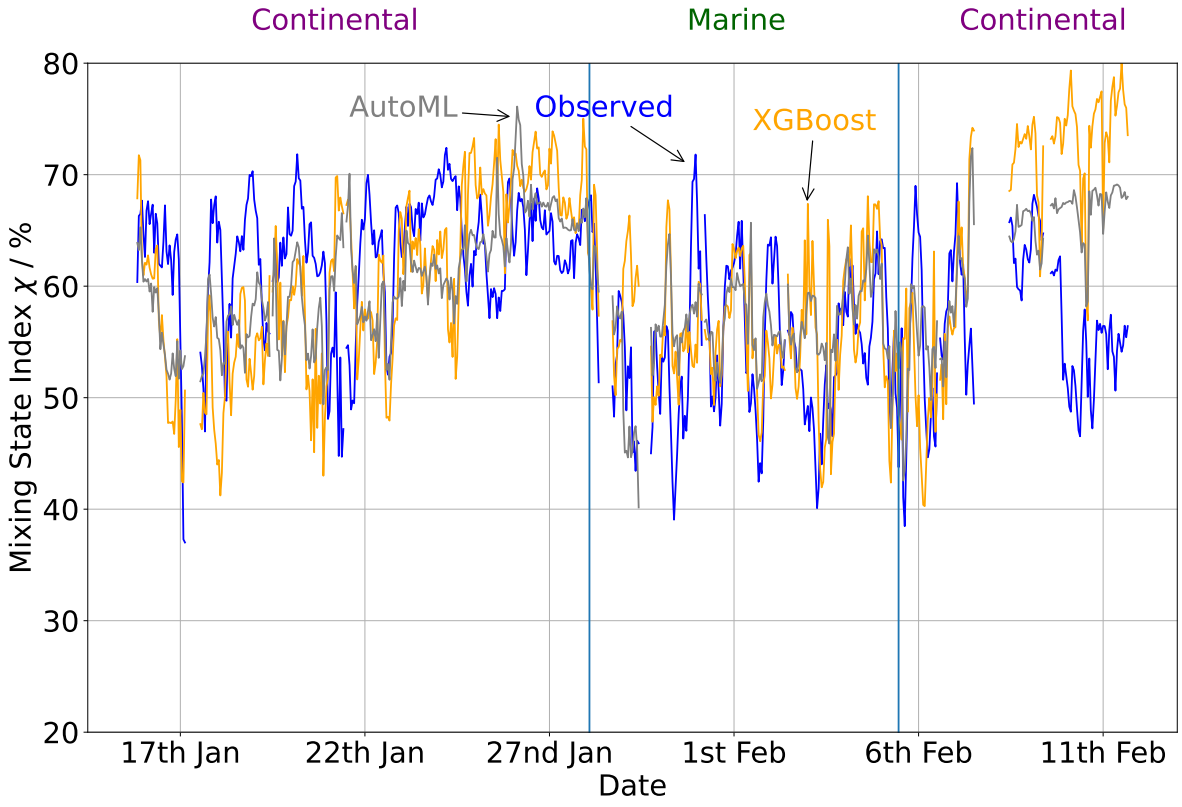

Figure 2: Time series of XGBoost and AutoML models predicting χ on the observed dataset over the marine versus the observed χ over the full-time period.

	RMSE (%)	MAPE (%)
Testing - XGBoost	6.88	7.33
Testing - AutoML	6.21	5.65
Observed - XGBoost	7.64	10.75
Observed - AutoML	6.29	9.09

Table 3: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) on the testing set and the observed set under base case conditions, using both the XGBoost model and the AutoML model.

it captures the mixing state during the continental period where long-range transport plays a role. Because of this caveat, we present all further prediction results limited to the marine period.

Table 3 provides our base case performance, which demonstrates that machine learning models are capable of making useful predictions on the real-world mixing state index χ while being trained on simulated data from software such as PartMC-MOSAIC. Table 3 displays the expected drop in performance going from the testing data, which is simulated, to the real-world, observed data. The performance metrics displayed for both models far exceed the performance of a baseline linear regression model. The baseline performance was so poor that it predicted values outside the valid range of the mixing state index χ. We therefore do not report the results of a linear regression baseline model. Given that our predictions on the observations are not significantly worse than predictions on the testing data, our models do not experience overfitting. As according to Lewis (1982), our MAPE values of 10.75% and 9.09% can be considered quite good, thus allowing the models to be potentially fit for further use.

In working with observations, the issue of missing data amongst features deserves additional discussion. The observation dataset contains a total of 645 timesteps, but 55 of them do not have observed χ values, leaving us with 590 observed χ values. Of those 590 timesteps, 466 have all features, thus 124 have at least one missing feature. The features that sometimes have missing values are the VOC species: PAR (119 timesteps with missing features), TOL (62), XYL (61), CH3OH (61), ALD2 (61), AONE (61), OLET (60), C2H6 (57), ETH (57). Frequently, multiple features would be missing at the same time, and those times would last several timesteps.

While we could simply discard these timesteps from our prediction, our machine learning frameworks can deal with the missing feature data, so we present our results including these timesteps. Algorithms like XGBoost attempt to address missing feature data while making forecasts by using a default set direction when encountering a missing value for the split node. The default direction is determined during the training process, by trying both directions and seeing which option minimizes the loss function. When excluding the timesteps with missing features for the marine period, the performance is an RMSE of 7.13% and MAPE of 9.43% for XGBoost, and an RMSE of 5.79%

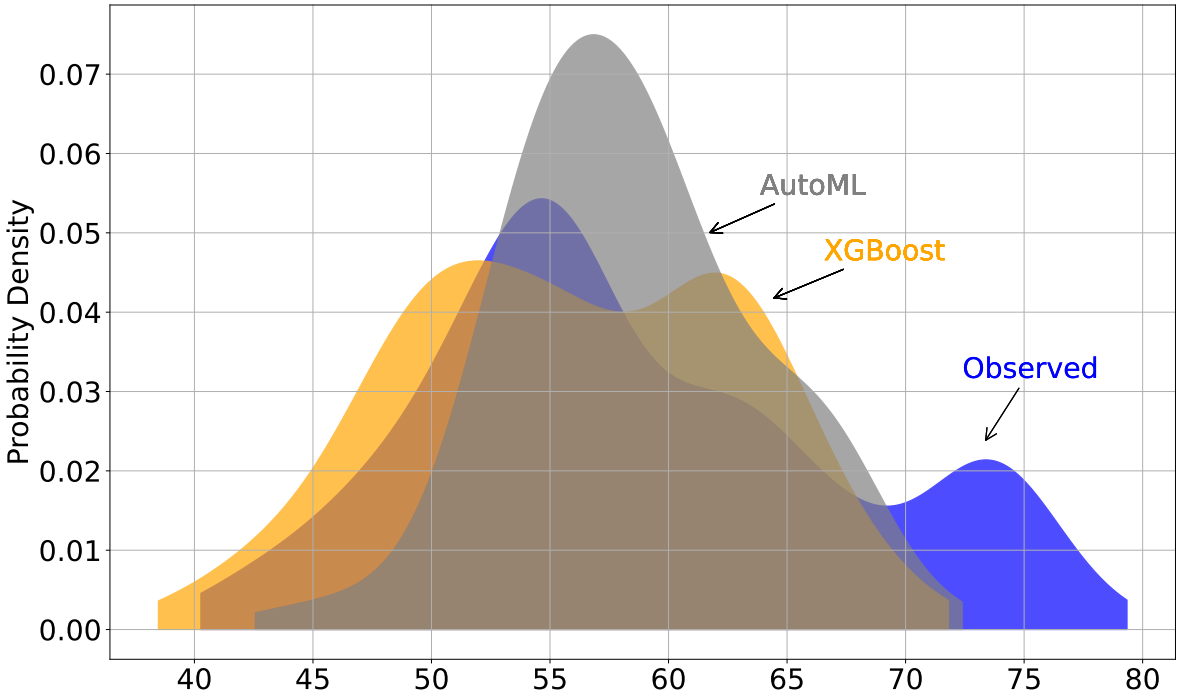

Figure 3: Kernel Density Estimators (KDEs) of XGBoost and AutoML models' predicted χ values on the observed dataset over the target (marine) time period, and the KDE of the observed χ over the same time period.

and MAPE of 7.43% for AutoML. As expected, we see a performance improvement when excluding the timesteps with missing features as those timesteps attempt to make predictions with less information, thus should be less accurate than the timesteps with all features present. However, this performance loss when including timesteps with missing features is not too large This suggests the frameworks were robust against the extent of the missing data in our case.

Figure 3 shows the distribution of the predictions for all timesteps. The distribution of the XGBoost model predictions better matches the observed data than AutoML, especially by having a more pronounced pair of local maxima in its distribution. Note that both models fail to capture the upper extreme observed χ values.

Table 3 shows that merely based on the RMSE and MAPE metrics, the AutoML selected model performed slightly better than XGBoost. However, in addition to a single performance metric (Table 3), we also included in our assessment the distribution of predictions (the KDE, Figure 3), and the time series predictions (Figure 2). Using this three-layered method of assessment yields a result that each method has its strengths and weaknesses. Figure 2 shows that the AutoML predictions tended to have a more limited range of predictions that on average are closer to the observed data than XGBoost, but XGBoost was better at capturing the amplitude of maxima and minima. Both models follow the general trend of the observed data, which is consistent with the single performance metrics in Table 3. Combining these three views of the results, both models

Case	RMSE (%)	MAPE (%)
Base case	7.64	10.75
Constant Temperature	10.80	14.30
$N = 200$ scenarios	13.72	17.78
$N = 600$ scenarios	8.53	11.86
VOC Features Dropped	8.71	14.09
Only Top 4 Features Kept	11.71	16.64

Table 4: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) on the observed set using the XGBoost model over the marine period for our various sensitivity cases.

provide satisfactory performance towards the objective of predicting the observed χ. We will present the XGBoost model for the results of the sensitivity analyses. The AutoML results show qualitatively the same behavior and can be found in the supplement.

3.3. Sensitivity Analysis

Figure 4 and Table 4 summarize the results for the sensitivity analyses. For each sensitivity case, the parameters that are not being discussed are all held to their value in the base case. Additionally, we explored the use of both the XGBoost and AutoML-based models, but for brevity we only present the results for XGBoost as the patterns presented from the XGBoost model also held for AutoML.

Figure 4a shows the effect of including a diurnal cycle for temperature in the training scenarios compared to keeping temperature constant during one scenario (although temperature can be different for different scenarios). Using constant temperatures for the training data results in an overestimation of χ prediction. Using a diurnal cycle for temperature in the training data improved our model—a 29% improvement in the RMSE, corresponding to 3.2 percentage points. The time series appears more in line with the observed data, and both single performance metrics improved as well. Applying a daily diurnal cycle in temperature allows the training data to be more realistic of real-world conditions, thus allowing for a training set more similar to the observed set which results in better model performance.

Figure 4b shows the impact of the size of the training set. We compare using a total of 200, 600, and 1000 scenarios – note that the testing size in each case remains 100 scenarios, so it is actually 100 training scenarios, 500, and 900. Both the metrics in Table 4 and the time series in Figure 4 paint a similar picture in that model performance improves with larger training sets. There may be some diminishing returns in those improvements: the difference in performance going from 200 to 600 scenarios (38% RMSE improvement, 5.2 percentage points) is more significant than going from 600 to 1000 scenarios (a 10% improvement in the RMSE, and 0.9 percentage points). As additional scenarios, while can be theoretically created on-demand, require more

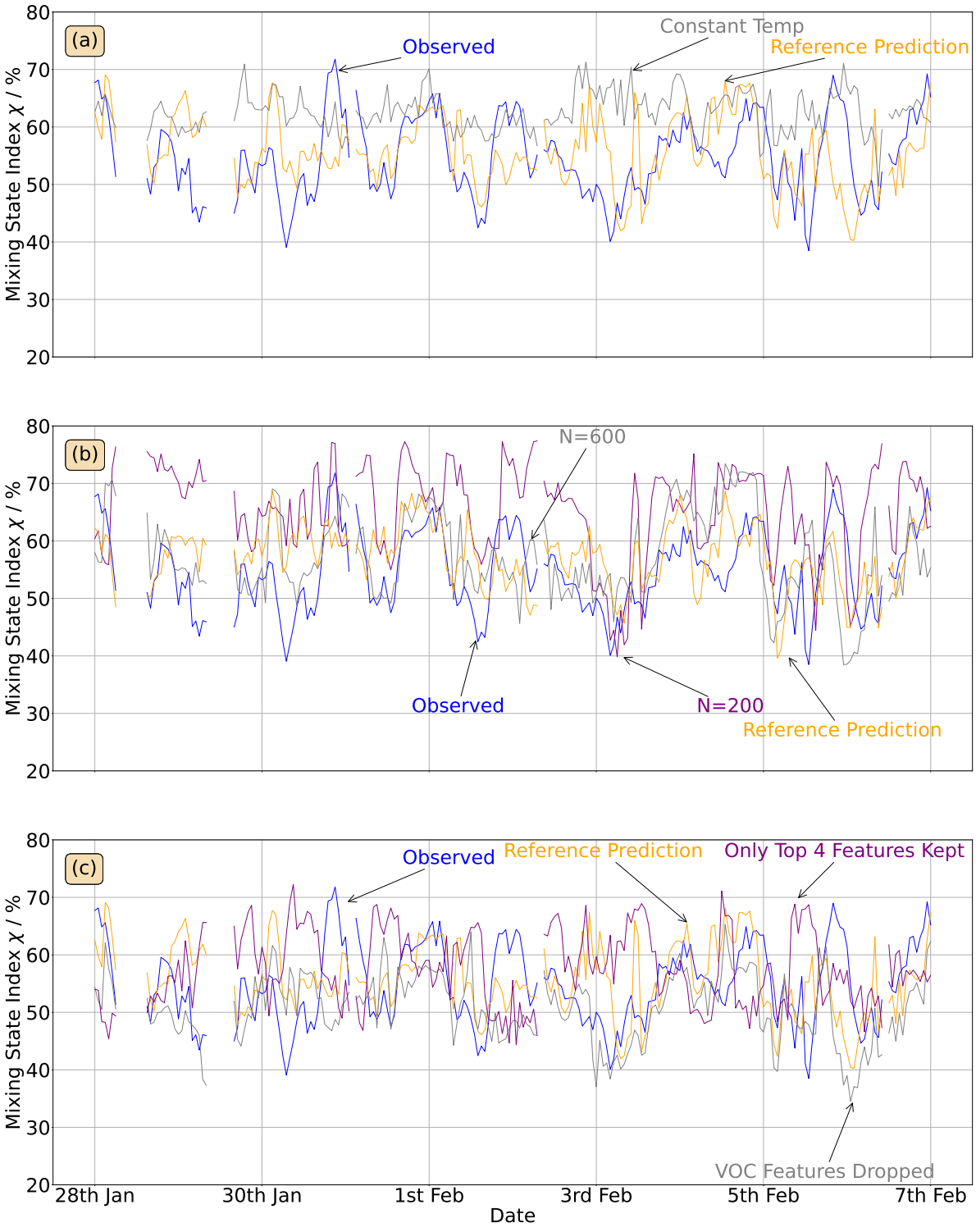

Figure 4: Sensitivity cases (grey and purple lines) compared to the base case (orange line) and observed data (blue line) (a) using constant temperature in PartMC scenarios to generate training data; (b) varying the training data volume ($N = 200, 600, 1000$); (c) varying the features included. See text for more details. All results are for the XGBoost model.

time and computational power, diminishing returns on performance would suggest the existence of a practical upper bound on the choice of the number of training scenarios.

Lastly, Figure 4c shows the effect of dropping features on model performance. In general, models with more features tend to be larger and possibly require more time or computational resources, so it may be worthwhile to let go of features with negligible impact on performance in turn for reduced training time and resources required. In our case, we note that the measurement process of our VOC features is more challenging than features such as temperature or trace gases that are routinely measured such as O_3 and NO_2. If the performance drop from dropping these features is negligible, it would suggest that there would be little need to exert the effort required to collect the observed data on these features.

As Table 4 and Figure 4 suggest, the impact of dropping the VOC features is not negligible; it does noticeably worsen performance as we see a 14% drop in the RMSE (1.1 percentage points). However, this drop may still be in an acceptable range considering that VOC species are not routinely measured.

Additionally, we show the effects of dropping all but "the top 4 features", for reference. These features are O_3, relative humidity, temperature, and CO. The ordering of the features is done using the notion of feature importance. For XGBoost, the feature importance score is the number of times the feature was used as the splitting variable for a tree in the training process. Once the features are ordered in this manner, and we drop all features from the model except for the top 4 features, we obtain the result we present. As it appears, this significantly degrades model performance, to the tune of a 53% drop in the RMSE (4.1 percentage points). A minor performance drop would suggest the vast majority of features are not useful to the model, which would not be expected, so this result is reasonable.

4. Conclusion

This work used observations from the MEGAPOLI field campaign in Paris, France, to evaluate a machine learning model's performance on predicting the aerosol mixing state metric χ. The training data was produced using the particle-resolved aerosol model PartMC-MOSAIC. The features were determined by the observed quantities of gas phase concentrations, bulk aerosol concentrations, relative humidity, and temperature.

We applied XGBoost and AutoML frameworks and showed that they yield comparable results in terms of RMSE (7.6% and 6.3% for XGBoost and AutoML, respectively and MAPE (10.8% and 9.1% for XGBoost and AutoML, respectively). However, the XGBoost model achieved a better agreement with the observed distribution of χ.

We conducted sensitivity analyses to assess the effect of key model contributors such as the amount of training data, the effect of adding diurnal temperature variation for scenario generation, and dropping certain features from the model altogether. Increasing the number of training scenarios from 200 to 600 scenarios gave significant improvement

of RMSE, while increasing the number to 1000 scenarios improved the RMSE further, but to a lesser extent. Adding the diurnal temperature variation to the training data was an important improvement to the workflow. Dropping the VOC gas concentration features led to an increase in RMSE from 7.6% to 8.7%, but this could be considered acceptable considering the difficulties in collecting VOC measurements on a routine basis. Only considering the top-four features (O_3, relative humidity, temperature, and CO) increased the RMSE to 11.7%. While still reasonable, this tells us that it is worth carrying additional information of nitrogen oxides and bulk aerosol concentrations.

While our results apply to the data collected from Paris in the wintertime, more observations coming from different locations around the world with varying distributions of aerosols would be valuable to ensure our findings are generalizable.

References

Baudic, A., V. Gros, S. Sauvage, N. Locoge, O. Sanchez, R. Sarda-Estève, C. Kalogridis, J.-E. Petit, N. Bonnaire, D. Baisnée, O. Favez, A. Albinet, J. Sciare, and B. Bonsang. Seasonal variability and source apportionment of volatile organic compounds (vocs) in the paris megacity (france). *Atmospheric Chemistry and Physics*, 16(18):11961–11989, 2016. doi:10.5194/acp-16-11961-2016.

Bauer, S., D. Wright, D. Koch, E. Lewis, R. McGraw, L.-S. Chang, S. Schwartz, and R. Ruedy. MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models. *Atmospheric Chemistry and Physics*, 8(20):6003–6035, 2008.

Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi:10.1145/2939672.2939785.

Cubison, M., B. Ervens, G. Feingold, K. Docherty, I. Ulbrich, L. Shields, K. Prather, S. Hering, and J. Jimenez. The influence of chemical composition and mixing state of Los Angeles urban aerosol on CCN number and cloud properties. *Atmospheric Chemistry and Physics*, 8(18):5649–5667, 2008.

de Myttenaere, A., B. Golden, B. Le Grande, and F. Rossi. Mean absolute percentage error for regression models. *Selected papers from the 23rd European Symposium on Artificial Neural Networks (ESANN 2015)*, 2017. doi:10.1016/j.neucom.2015.12.114.

DeCarlo, P. F., J. R. Kimmel, A. Trimborn, M. J. Northway, J. T. Jayne, A. C. Aiken, M. Gonin, K. Fuhrer, T. Horvath, K. S. Docherty, et al. Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer. *Analytical chemistry*, 78(24):8281–8289, 2006.

DeVille, L., N. Riemer, and M. West. Convergence of a generalized weighted flow algorithm for stochastic particle coagulation. *Journal of Computational Dynamics*, pages 1–18, 2019. doi:10.3934/jcd.2019003.

DeVille, R. E. L., N. Riemer, and M. West. Weighted Flow Algorithms (WFA) for stochastic particle coagulation. *J. Computational Phys.*, 230(23):8427–8451, 2011. doi:10.1016/j.jcp.2011.07.027.

Dua, R., M. S. Ghotra, and N. Pentreath. *Machine learning with spark - second edition*. Packt Publishing, 2016.

Ervens, B., M. Cubison, E. Andrews, G. Feingold, J. Ogren, J. Jimenez, P. Quinn, T. Bates, J. Wang, Q. Zhang, et al. CCN predictions using simplified assumptions of organic aerosol composition and mixing state: a synthesis from six different locations. *Atmospheric Chemistry and Physics*, 10(10):4795–4807, 2010.

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

Fierce, L., N. Riemer, and T. C. Bond. Toward Reduced Representation of Mixing State for Simulating Aerosol Effects on Climate. *Bull. Amer. Meteor. Soc.*, 98(5):971–980, 2017. doi:10.1175/BAMS-D-16-0028.1.

Healy, R., N. Riemer, J. Wenger, M. Murphy, M. West, L. Poulain, A. Wiedensohler, I. O'Connor, E. McGillicuddy, J. Sodeau, et al. Single particle diversity and mixing state measurements. *Atmospheric Chemistry and Physics*, 14(12):6289–6299, 2014.

Healy, R. M., J. Sciare, L. Poulain, M. Crippa, A. Wiedensohler, A. S. Prévôt, U. Baltensperger, R. Sarda-Estève, M. L. McGuire, C.-H. Jeong, et al. Quantitative determination of carbonaceous particle mixing state in paris using single-particle mass spectrometer and aerosol mass spectrometer measurements. *Atmospheric Chemistry and Physics*, 13(18):9479–9496, 2013.

Healy, R. M., J. Sciare, L. Poulain, K. Kamili, M. Merkel, T. Müller, A. Wiedensohler, S. Eckhardt, A. Stohl, R. Sarda-Estève, et al. Sources and mixing state of size-resolved elemental carbon particles in a european megacity: Paris. *Atmospheric Chemistry and Physics*, 12(4):1681–1700, 2012.

Hersbach, H., B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1959 to present. Technical report, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2021. doi:10.24381/cds.adbb2d47.

Lewis, C. *Industrial and Business Forecasting Methods*. Butterworth-Heinemann, 1982.

Liu, X., P.-L. Ma, H. Wang, S. Tilmes, B. Singh, R. Easter, S. Ghan, and P. Rasch. Description and evaluation of a new four-mode version of the modal aerosol module (MAM4) within version 5.3 of the Community Atmosphere Model. *Geoscientific Model Development*, 9:505–522, 2016.

Ma, J., Z. Yu, Y. Qu, J. Xu, and Y. Cao. Application of the xgboost machine learning method in pm2.5 prediction: A case study of shanghai. *Aerosol Air Qual. Res*, 20:138 – 148, 2020. doi:10.4209/aaqr.2019.08.0408.

Petzold, A. and M. Schönlinner. Multi-angle absorption photometry—a new method for the measurement of aerosol light absorption and atmospheric black carbon. *Journal of Aerosol Science*, 35(4):421–441, 2004.

Prather, K. A., C. D. Hatch, and V. H. Grassian. Analysis of atmospheric aersols. *Annual Review of Analytical Chemistry*, 1:485–514, 2008. doi: 10.1146/annurev.anchem.1.031207.113030.

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

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

Riemer, N., M. West, R. A. Zaveri, and R. C. Easter. Simulating the evolution of soot mixing state with a particle-resolved aerosol model. *J. Geophys. Res. Atmos.*, 114(D9):D09202, 2009. doi:10.1029/2008JD011073.

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

Snoek, J., H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2:2951 – 2959, 2012. doi: 10.5555/2999325.2999464.

Sun, A. Y., B. R. Scanlon, H. Save, and A. Rateb. Reconstruction of grace total water storage through automated machine learning. *Water Resources Research*, 57, 2021. doi:https://doi.org/10.1029/2020WR028666.

Wang, C., Q. Wu, M. Weimer, and E. Zhu. Flaml: A fast and lightweight automl library. *MLSys '21: Proceedings of the Fourth Conference on Machine Learning and Systems Conference*, 3:434 – 447, 2021. doi:10.1145/2939672.2939785.

Winkler, P. The growth of atmospheric aerosol particles as a function of the relative humidity—II. An improved concept of mixed nuclei. *Journal of Aerosol Science*, 4(5):373–387, 1973.

Wu, Q. W., C. Wang, and S. Huang. Frugal optimization for cost-related hyperparameters. *Proceedings Of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021. doi:10.48550/arXiv.2005.01571.

Ye, Q., P. Gu, H. Z. Li, E. S. Robinson, E. Lipsky, C. Kaltsonoudis, A. K. Lee, J. S. Apte, A. L. Robinson, R. C. Sullivan, et al. Spatial variability of sources and mixing state of atmospheric particles in a metropolitan area. *Environmental Science & Technology*, 52(12):6807–6815, 2018.

Zamani Joharestani, M., C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani. Pm2.5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, 10, 2019. doi:10.3390/atmos10070373.

Zaveri, R. A., R. C. Easter, J. D. Fast, and L. K. Peters. Model for Simulating Aerosol Interactions and Chemistry (MOSAIC). *J. Geophys. Res. Atmos.*, 113(D13), 2008. doi:10.1029/2007JD008782.

Zaveri, R. A., R. C. Easter, and L. K. Peters. A computationally efficient multicomponent equilibrium solver for aerosols (mesa). *Journal of Geophysical Research: Atmospheres*, 110(D24), 2005.

Zaveri, R. A. and L. K. Peters. A new lumped structure photochemical mechanism for large-scale applications. *Journal of Geophysical Research: Atmospheres*, 104:30387–30415, 1999.

Zheng, Z., J. H. Curtis, Y. Yao, J. T. Gasparik, V. G. Anantharaj, L. Zhao, M. West, and N. Riemer. Estimating submicron aerosol mixing state at the global scale with machine learning and earth system modeling. *Earth and Space Science*, 8(2), 2021a. doi:https://doi.org/10.1029/2020ea001500.

Zheng, Z., M. West, L. Zhao, P.-L. Ma, X. Liu, and N. Riemer. Quantifying the structural uncertainty of the aerosol mixing state representation in a modal model. *Atmospheric Chemistry and Physics*, 21(23):17727–17741, 2021b.