ACS ES&T Air

Open Access

This article is licensed under CC-BY 4.0

Integrating Simulations and Observations: A Foundation Model for Estimating the Aerosol Mixing State Index

Published as part of ACS ES&T Air special issue "Elevating Atmospheric Chemistry Measurements and Modeling with Artificial Intelligence".

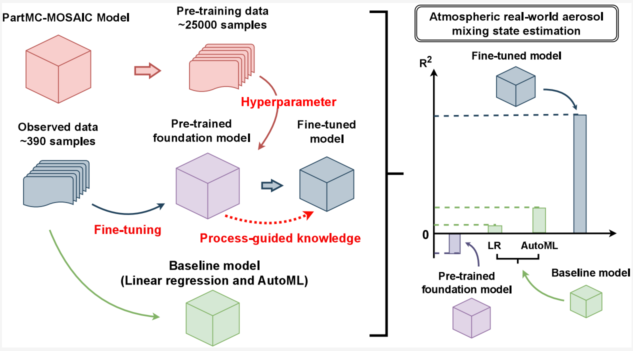
Fei Jiang, Zhonghua Zheng,* Hugh Coe, Robert M. Healy, Laurent Poulain, Valérie Gros, Hao Zhang, Weijun Li, Dantong Liu, Matthew West, David Topping,* and Nicole Riemer*

Cite This: *ACS EST Air* 2025, 2, 877−890

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurately predicting aerosol mixing states in real-world environments is crucial for understanding their impacts on climate change and human health. However, observational data inherently exhibit spatiotemporal gaps, and high costs and equipment requirements further exacerbate these limitations, particularly for in situ measurements. While particle-resolved models can simulate individual particle composition and size changes and serve as benchmarks, they face challenges in real-world applications due to a combination of factors. One of the major challenges is the limited availability of detailed input data (e.g., emission inventories) that accurately reflect actual environmental conditions. In this study, we frame the emulation of aerosol simulation as a general task and treat the estimation of real-world mixing states as a downstream task. We developed a foundation model pretrained on particle-resolved simulations and fine-tuned it using observational data from the field campaign. The fine-tuned model consistently outperformed baseline models, showing greater stability and robustness across various data sets. Permutation feature importance and sensitivity analyses revealed that aerosol species concentrations were the most critical factors for the foundation model. This approach, which involves pretraining on particle-resolved simulations and fine-tuning on limited observational data, offers a viable solution to challenges posed by limited observational data.


PartMC-MOSAIC Model
Pre-training data ~25000 samples
Hyperparameter
Observed data ~390 samples
Pre-trained foundation model
Fine-tuned model
Fine-tuning
Process-guided knowledge
Baseline model (Linear regression and AutoML)
Atmospheric real-world aerosol mixing state estimation
Fine-tuned model
R^2
0
LR
AutoML
Pre-trained foundation model
Baseline model

KEYWORDS: *Aerosol Mixing State, Foundation Model, Transfer Learning, Atmospheric Aerosols, Particle-resolved Model, Machine Learning, Deep Learning, Aerosol Mixing State Index*

■ INTRODUCTION

Atmospheric aerosols are complex mixtures with variable chemical compositions that evolve through processes such as condensation, coagulation, and evaporation.[1−3] The differences in the composition among aerosol particles are described by "aerosol mixing state".[4,5] In a "fully internally mixed" state, all particles share the same chemical composition, while in a "fully externally mixed" state, each particle consists of a single species.[5] In reality, most aerosols exhibit mixing states that fall between these two extremes.[6−10] For instance, aerosols are typically more externally mixed near emission sources,[5,7,11] and aging processes such as coagulation and gas condensation shift the aerosols toward a more internally mixed state.[7,12]

Recent studies have shown that misrepresenting the aerosol mixing state introduces notable errors in estimating aerosol properties, which can impact the evaluation of aerosol effects on climate and human health. For example, assuming aerosols are internally mixed may lead to increased estimates of aerosol absorption of solar radiation compared to external mixing, even with the same bulk aerosol species concentration, potentially resulting in overestimating radiative forcing.[13,14] The aerosol mixing state also influences studies on aerosol optical properties,[15−18] cloud condensation nuclei (CCN) activity,[19−22] cloud droplet activation,[23] and deposition of aerosols in the human respiratory system.[24,25] Therefore, accurately quantifying the aerosol mixing state is essential for improving our understanding of aerosols' environmental and health impacts.

Riemer and West[26] defined the aerosol mixing state index χ, ranging from 0% (completely external mixtures) to 100%

Received: November 25, 2024
Revised: March 28, 2025
Accepted: March 28, 2025
Published: April 11, 2025

https://doi.org/10.1021/acsestair.4c00329

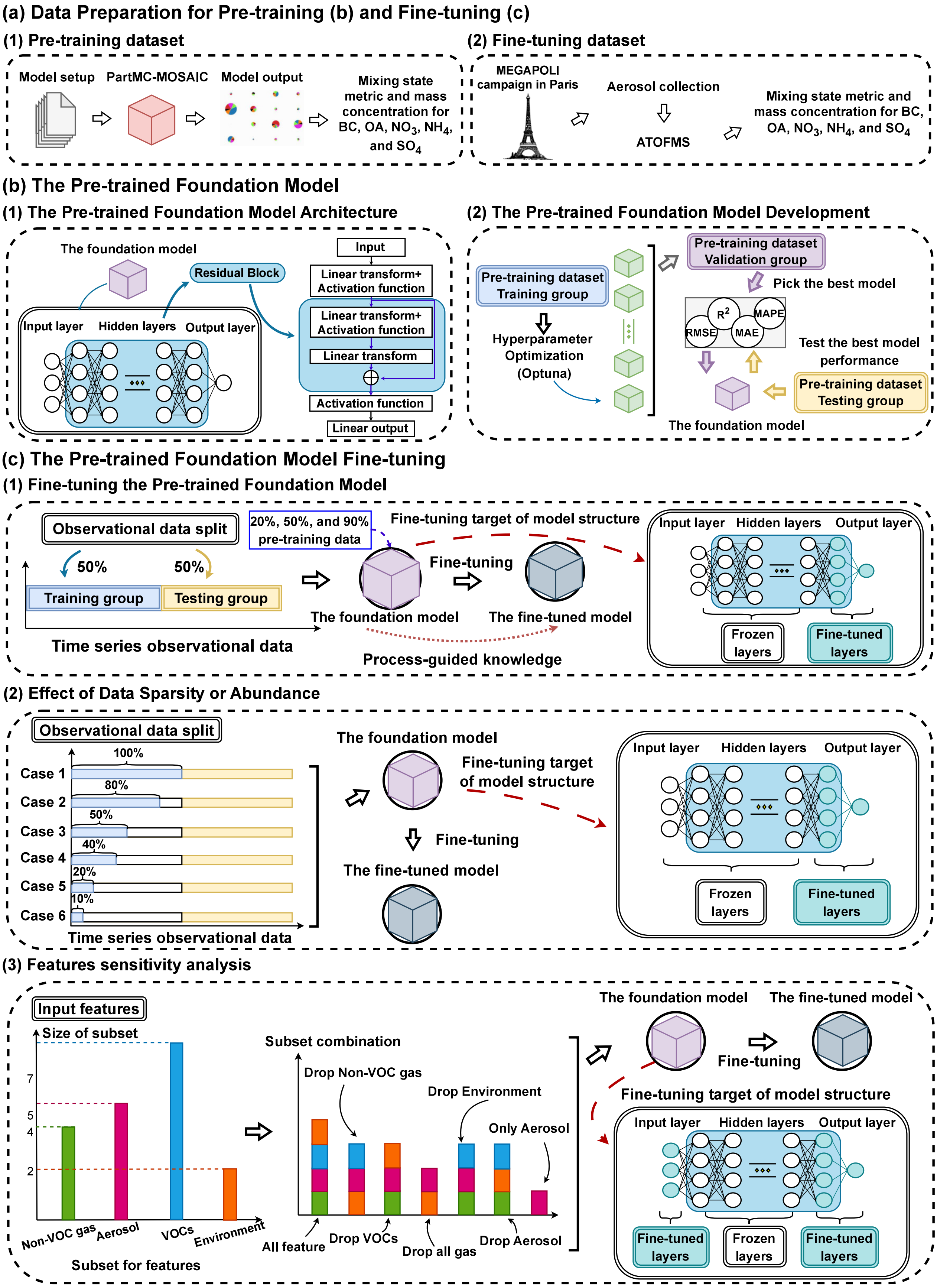


Figure 1. Overview of the workflow for the foundation and fine-tuned model development. (a) Data Preparation for Pretraining and Fine-tuning: (1) Generation of the pretraining data set for the foundation model using the PartMC-MOSAIC simulation. (2) Fine-tuning data set derived from the MEGAPOLI campaign observational data, source from Healy et al.[6] (b) The Pretrained Foundation Model: (1) Architecture of the foundation model,

Figure 1. continued

illustrating the overall structure (left) and the residual block structure (right). The "linear transformation" refers to a simple mathematical operation where each input is multiplied by a weight and added to a bias (a constant value), linearly altering the input values. (2) Model development through hyperparameter optimization, where the purple marker highlights the best foundation model selected during validation, and the yellow marker represents its performance on the pretraining testing set. (c) Fine-tuning the Pretrained Foundation Model: (1) Fine-tuning process, where observational data is chronologically split into a fine-tuning training set (first 50%) and a fine-tuning testing set (remaining data). The foundation model is pretrained using three different data volumes (20%, 50%, and 90% of the pretraining data) to investigate how the amount of pretraining data affects fine-tuning performance. (2) Effect of data sparsity or abundance, with different data subsets (case 1- 6) used for training and evaluated on the fine-tuning testing set. Fine-tuning targets include the output layer and adjacent layers. In the "Swapped Temporal Order" experiment, the roles of the data sets are reversed: the fine-tuning testing set is redivided (using the same proportional split) to serve as the training set, while the original fine-tuning training set is used for testing. (3) Feature sensitivity analysis, examining the model's response to different feature subsets, with fine-tuning targeting both input and output layers alongside adjacent layers.

(completely internal mixtures) to quantify the aerosol mixing state. This index uses diversity measures based on the Shannon entropy of chemical species distribution among particles.[26] The metric χ has been applied in field observations across various environments, including Paris during the MEGAPOLI (MEGAcities: Emissions, urban, regional and Global Atmospheric POLlution and climate effects, and Integrated tools for assessment and mitigation) campaign,[6] Northern California during CARES,[27] Pittsburgh, PA[10] and Beijing.[9] Due to the high cost of fieldwork and the complexity of required equipment, observations are often limited in temporal and spatial coverage, making it especially difficult to capture dynamic aerosol mixing state changes over long periods or large areas.[5] Therefore, simulations have become essential for studying aerosol mixing states, particularly when observational data is sparse or has gaps in spatial and temporal coverage.

The most direct way to simulate the aerosol mixing state and its index χ in modeling work is through a particle-resolved aerosol model. This approach tracks the composition of each particle, allowing χ to be calculated without assumptions about the aerosol mixing state.[28] However, substantial challenges remain in applying the particle-resolved aerosol model to study real-world aerosol mixing states. One of the primary gaps lies in the lack of necessary inputs, such as emission inventories and aerosol size distributions, that accurately reflect actual environmental conditions. Few studies have compared model outputs with observations to validate aerosol mixing state predictions. For example, Zhu et al.[29] validated the regional mixing-state-resolving SCRAM (size-composition resolved aerosol model) using MEGAPOLI data, and Zheng et al.[30] use the same observational data to validate a data-driven aerosol mixing state emulator trained by particle-solved aerosol simulations. Both studies rely on temporally aggregated data, smoothing out short-term variations like pollution spikes. This makes it difficult to ensure the reliability of the models in representing real-world aerosol mixing state changes.

In recent years, the rapid development of deep learning has driven the rise of foundation models.[31] Foundation models are pretrained on large-scale data sets, providing reasonable parameter initialization for downstream tasks.[31] The key is using transfer learning to apply "knowledge" from foundation models to related tasks or data sets, thereby facilitating the model development of the downstream tasks.[31,32] A notable example of a foundation model application is GPT-3.5, a large language foundation model pretrained on vast amounts of text and code.[33,34] ChatGPT, fine-tuned from GPT-3.5, is specifically optimized for conversational tasks.[31,35] In atmospheric sciences, foundation models are mainly applied in weather and climate modeling[36] (such as ClimateX,[37] FengWu,[38] PanGu,[39] FuXi,[40] GraphCast,[41] FourCastNet,[42] W-MAE,[43] and CliMedBert[44]), as well as in satellite remote sensing.[45−47] However, their use in atmospheric aerosol research remains limited, especially in the context of broader environmental applications.

This study proposes a foundation model-based approach that integrates pretraining on aerosol simulations with fine-tuning on limited observational data. The primary goal is to treat the emulation of aerosol simulation as a general task (focusing on predicting χ), with the prediction of χ in real-world environments as downstream tasks. Specifically, we developed a foundation model using particle-resolved model simulations to predict changes in χ, based on the information on aerosol species, gas species, and environmental factors. The foundation model incorporates process-guided knowledge, learned from process-based model results, providing reliable parameter initialization for downstream χ predictions. This method maximizes the utility of limited observational data while leveraging the advantages of simulations to capture the temporal evolution of aerosol mixing states under various conditions, reducing reliance on uncertain inputs like aerosol emission inventories. By fine-tuning the foundation model with a small amount of observational data, our approach demonstrates a clear improvement in predicting aerosol mixing states in real-world environments compared to traditional data-driven models. The scalability of the pretrained and fine-tuned models offers a viable solution for aerosol research and better quantifying aerosol environmental and health impacts, particularly in scenarios with limited observational data.

■ MATERIALS AND METHODS

This study aims to develop a data-driven model that accurately predicts real-world aerosol mixing states by integrating process-based simulations with observations. The workflow overview for developing the foundation and fine-tuned models is illustrated in Figure 1. The figure outlines the data preparation for pretraining and fine-tuning data sets, the training of a pretrained foundation model, and its subsequent fine-tuning with observational data from the MEGAPOLI campaign.

Aerosol 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.[26] 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[6] and spatial[10] variability. Here, we focus on the mixing state of 100−700 nm aerosols, since the aerosol time-of-flight mass spectrometer (ATOFMS), used to determine χ observationally, 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

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

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

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 in Riemer and West.[26]

Data Preparation for Pretraining and Fine-Tuning.

Particle-Resolved Aerosol Model. PartMC-MOSAIC[28,48] is a stochastic particle-resolved aerosol model. The model details are described in Riemer et al.[28] and DeVille et al., DeVille et al.[49,50] for PartMC, and in Zaveri et al.[48] for MOSAIC.

In brief, the Lagrangian box model PartMC represents the evolution of aerosol particles in a fully mixed computational volume. We stochastically simulate the processes of emission, coagulation and dilution. 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,[51] 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.[52] The formation of secondary organic aerosol (SOA) is represented by the Secondary Organic Aerosol Model (SORGAM).[53]

Since the particle-resolved approach of PartMC-MOSAIC does not rely on any a priori assumptions about aerosol mixing states, it serves as a benchmark model for representing them. In this study, we used PartMC version 2.6.1 to generate the training, validation, and testing data sets for developing our foundation model. The training data set is used to train the model by learning the patterns in the data. The validation data set is applied to adjust model hyperparameters. After completing the training and validation, the testing data set is reserved for the final evaluation, providing an objective measure of the model's accuracy.

Pretraining Data Set: PartMC-MOSAIC Simulations. PartMC-MOSAIC is used to generate the data for pretraining the foundation model. This follows the procedure described in Zheng et al.[30,54] The PartMC-MOSAIC scenarios simulate the evolution of atmospheric aerosols using varying input parameters. By analyzing the hourly simulated output for each timestamp within each scenario and applying eq 1, we derive mixing state indices. The input parameters, listed in Table 1, including primary emissions of different aerosol types (e.g., carbonaceous aerosol, sea salt, and dust emissions, with contributions from Aitken mode, accumulation mode, and coarse mode size ranges), primary emissions of gas phase species (e.g., Sulfur dioxide(SO_2), Nitrogen dioxide(NO_2), Carbon monoxide(CO), and various volatile organic compounds), and meteorological parameters. Latin Hypercube Sampling (LHS) determined the parameter combination for each scenario. Each scenario involves 10,000 computational particles to model the aerosol population, starting at 6:00 a.m. local time, running for 24 h, and generating hourly outputs. Following postprocessing, each scenario provides 24 samples with the corresponding χ labels, one for each hourly snapshot during the scenario. The 90−5−5 framework has been widely used in deep learning applications across medicine,[55,56] chemistry,[57] and material.[58] We applied a shuffled 90−5−5 split at the scenario level for the 1,000 scenarios, assigning each entire scenario to either the training set (90%), validation set (5%), or test set (5%).

Fine-Tuning Data Set: MEGAPOLI Campaign Observations. Observational data from the MEGAPOLI winter campaign,[59,60] were collected at the Laboratoire d'Hygiéne de la Ville de Paris (LHVP), Paris, France from 15 January−11 February 2010, is used for transfer learning. This data set 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),[6] and the data set also provides measurements of many bulk aerosol and gas phase species, which are needed as features (predictive variables or inputs) for our foundation model. The specific steps taken to preprocess ATOFMS data can be found in SI Text S1. Table 2 details the specific aerosol species selected for χ calculation in both the PartMC simulation and MEGAPOLI observational data.

For the MEGAPOLI data, the air mass over the observation site originated from the ocean between January 15, 2010, at 8:00 PM local time and February 7, 2010, at midnight local time.[6] Outside this period, the air mass came from continental Europe.[6] Given the distinct emission sources of marine and continental air masses and the PartMC model's limitations in addressing all conditions simultaneously, our study concentrated on the marine period data, which constitutes over 80% (390 out of 466 timestamps) of the total data set.

The MEGAPOLI data set consists of time-series data, with sequential observations recorded over time. To avoid data leakage and ensure that the model's performance is evaluated on unseen data, the data set is split into training and testing sets based on chronological order. The first 50% of the data is used as the training set, while the remaining 50% is reserved for testing (Figure 1 c (1)). To differentiate these sets from those used in pretraining, we will refer to this specific training set as the "fine-

Table 1. List of Input Parameters and Their Sampling Ranges to Construct the Training and Testing Scenarios for the Foundation Model[a]

Parameters	Range
Environmental Variable	
Relative humidity (RH)	[0.4, 1)
Latitude	90°S, 90°N
Day of Year	[1, 365]
Temperature	Varies with day of the year and location assumptions for each scenario, remaining constant within each scenario
Gas Phase Emissions Scaling Factor	
SO_2, NO_2, NO, NH_3, CO, CH_3OH,	[0, 200%] of the reference scenario
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)	
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 h]

[a] The variables D_g, σ_g, and E_a refer to geometric mean diameter, geometric standard deviation, and number emission flux, respectively.

Table 2. Key Aerosol Species (Size Range: 100−700 nm) Used in Calculating the Mixing State Index χ for PartMC Simulations and MEGAPOLI Observational Data[a]

Data Source	Species 1	Species 2	Species 3	Species 4	Species 5
PartMC simulation	BC	SO_4	NO_3	NH_4	ARO1, ARO2, ALK1, OLE1, API1, API2, LIM1, LIM2, OC
MEGAPOLI observation	BC	SO_4	NO_3	NH_4	OA

[a] In PartMC, organic aerosols (OA) include aromatic hydrocarbons (ARO), alkanes (ALK), olefins (OLE), and other organic compounds (OC), corresponding to the OA category in the MEGAPOLI data set.

tuning training set" and the testing set as the "fine-tuning testing set."

The Pretrained Foundation Model. *The Pretrained Foundation Model Architecture.* The foundation model is trained using simulation data from PartMC-MOSAIC, employing the ResNet (Residual Neural Network)-like architecture for aerosol mixing state estimates. The model consists of an input layer, multiple residual blocks, and an output layer Figure 1 b (2)). The input layer is a linear layer that converts 18 input features to the desired hidden size (the number of neurons in the hidden layer of the foundation model), preparing the data for further processing. These input features (Table S2) include temperature (T), relative humidity (RH), carbon monoxide (CO), nitrogen oxides (NO_x), nitrous oxide (NO), ozone (O_3), xylene (XYL), ethene (ETH), acetone (AONE), toluene (TOL), paraffin carbon (PAR), internal olefin carbons (OLET), and acetaldehyde (ALD2), ammonium mass concentration (NH_4), sulfate mass concentration (SO_4), nitrate mass concentration (NO_3), black carbon mass concentration (BC), and organic aerosols mass concentration (OA). The output layer is a linear layer that reduces the dimensionality of the hidden features to the output feature (χ), making it suitable for regression tasks. ReLU activations follow each linear layer, introducing nonlinearity into the model, and enabling the model to learn complex patterns in the data.

The core of the model is composed of multiple residual blocks, as illustrated in Figure 1(c). Each block contains two linear layers interleaved with ReLU activation functions. A key feature of this module is the shortcut connection, which bypasses the two linear layers by directly adding the block's input to its output.[61] These residual connections help mitigate the vanishing gradient problem, enabling the network to learn more effectively, even as its depth increases.[62]

The Pretrained Foundation Model Development. Hyperparameters, set before training, play a crucial role in a model's ability to learn data patterns and make accurate predictions, alongside the parameters learned during model training (such as weights and biases). Selecting the optimal configuration of hyperparameters is critical for achieving high accuracy in the foundation model construction. Due to the massive number of potential models created by different hyperparameter combinations, We employ Optuna (version 3.6.1)[63] for hyperparameter optimization, leveraging Bayesian optimization techniques (Figure 1 b (1)). Table 3 outlines the hyperparameters and their respective ranges. The foundation model is implemented using PyTorch (version 2.3.0) within a Python 3.8.18 environment.

Fine-Tuning the Pretrained Foundation Model. Transfer learning leverages knowledge from pretrained models to enhance model performance on related tasks or different data sets.[64] The most common strategy for knowledge transfer is fine-tuning a pretrained foundation model on a new data set.[65] Methods include fine-tuning all neural network parameters,[66] only the parameters of the last few layers,[67] or using the pretrained model as a fixed feature extractor with a classifier.[68]

In this study, we fine-tuned the foundation model using the fine-tuning training set. The primary strategy involved updating the final layers, including the output layer, with the number of

Table 3. Hyperparameters and Range Values for Optimization

Hyperparameter	Range
batch_size	[16, 32, 64, 128]
hidden_size	[128, 256, 512, 1024]
learning_rate[a]	$10^{-6} - 10^{-4}$
num_blocks[b]	10 − 20 [step = 1]

[a]The learning rate is sampled between 1e-5 and 1e-3, and the sampling is done in a logarithmic space. [b]The number of blocks is sampled between 10 and 20, inclusive. The step = 1 parameter ensures that Optuna will consider every integer within this range.

Table 4. Hyperparameters and Range Values for Fine-Tuned Model Optimization

Hyperparameter	Range	Sampling Method
batch_size	1	NA
L2 regularization coefficient (λ)[a]	10^{-6}–10^{-3}	Log-uniform sampling
learning_rate	10^{-7}–10^{-4}	Log-uniform sampling
Number of fine-tuned layers (Final layers)[b]	1−5	Integer sampling
Number of fine-tuned layers (Initial layers)[c]	1	NA

[a]L2 regularization coefficient (λ): a small λ allows the model more flexibility but increases the risk of overfitting, while a larger λ improves regularization at the cost of potential underfitting. [b]Number of fine-tuned layers (Final layers): The final few layers of the model, usually including the output layer and preceding layers. [c]Number of fine-tuned layers (Initial layers): The initial layers of the model, responsible for extracting fundamental features; the fine-tuning input layer is only applied in feature sensitivity analysis to account for changes in the number of input features.

trainable layers as an experimental variable (Figure 1 c). The minimal fine-tuning case updated only the output layer, while the maximal case adjusted the last five layers (Table 4). For feature sensitivity experiments, the same fine-tuning strategy was applied, but since the number of input features varied, the input layer was also fine-tuned to accommodate these changes (Figure 1 c (3) and Table 4).

The fine-tuning data set was normalized using the same standard as the pretraining data set. Due to the limited size of the fine-tuning training set (maximum 195 samples), we set the batch size to 1 to ensure each sample contributed fully to model updates. We applied Ridge Regression (L2 regularization), which penalizes large weights by adding an L2 penalty term to the loss function.[69] This constraint reduces model variance and improves generalization, particularly when training data is limited.[70] Hyperparameters, including the L2 regularization coefficient, learning rate, and the number of fine-tuned layers, were optimized using Optuna (details in Table 4). We designed a series of experiments to evaluate the effectiveness of the fine-tuned model on observational data. The fine-tuned model retains the structure of the pretrained model while adjusting only the last two layers. Additional details are presented in Text S3 in the SI.

Impact of Pretraining Data Volume on Fine-Tuning Performance. In the first experiment, we investigated the relationship between the volume of pretraining data and the performance of fine-tuned models. Specifically, pretrained models were constructed using 20%, 50%, and 90% of the pretraining data, while the fine-tuning data set was kept constant for performance evaluation.

We compared the performance of these fine-tuned models with that of baseline models trained solely on the fine-tuning training set. The baseline models included a linear regression (LR) model and the "best tree-based model" selected through Automated Machine Learning (AutoML) using the Fast Lightweight Automated Machine Learning (FLAML) framework.[71−73] FLAML optimized model selection, hyperparameters, sample sizes, and resampling strategies, evaluating models such as Light Gradient Boosting Machine (LightGBM),[74] eXtreme Gradient BoostingXGBoost (XGBoost),[75] Random Forest (RF)[76] and Extremely randomized trees (Extra-Trees).[77] For consistency, all models were evaluated on the corresponding fine-tuning testing sets.

Effect of Data Sparsity or Abundance. The second experiment examined the impact of training data sample size on fine-tuned model performance (Figure 1 c (2)). The fine-tuning training set was divided into six chronological subsets containing 10%, 20%, 40%, 50%, 80%, and 100% of the data, and these subsets were used to fine-tune the pretrained model, with performance evaluated on the fine-tuning testing set. Baseline models (AutoML and LR) were trained with the corresponding subsets for comparison. Additionally, to explore whether the temporal partitioning of fine-tuning data affects the results, we swapped the roles of the training and testing set by dividing the fine-tuning testing set into six similar chronological subsets; these subsets were then used to fine-tune the pretrained model, with performance evaluated on the fine-tuning training set, while the corresponding baseline models were trained with the same subsets for comparison.

Features Sensitivity Analysis. The third experiment tested the model's sensitivity to different input features using two approaches: feature subset exclusion (Figure 1 c (3)) and permutation feature importance (PFI; see Model Evaluation and Interpretation for details). We created four feature subsets (Table S2) and combined them into six different subsets: 1) excluding non-VOC gases, 2) excluding VOCs, 3) excluding both gases and VOCs, 4) excluding environmental parameters, 5) excluding aerosols, and 6) using aerosol data only. Fine-tuning focused on the input and output layers of the neural network, as well as adjacent layers, to accommodate changes in input feature size. Additionally, PFI was applied to assess the impact of each feature on model performance. These approaches allowed us to identify the most critical features for accurate predictions.

Model Evaluation and Interpretation. Model performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2. Note that MAPE is always expressed as a percentage, while RMSE and MAE retain the units of the measured quantity. In this study, since χ is dimensionless and ranges from 0 to 1, RMSE and MAE values are float points, whereas MAPE values represent percentage errors.

PFI is used to evaluate the importance of features across the entire data set.[78] Feature importance is determined by measuring the increase in the model's prediction error after shuffling the feature's values.[79] This process disrupts the relationship between the feature and the target variable, and the resulting increase in the model error reflects the extent to which the model relies on that feature for its predictions. A substantial increase in error indicates that the feature is crucial for the model's predictions, while no change suggests the feature

Table 5. Performance Comparisons of Fine-Tuning Models with Different Pretraining Data Volumes and Baseline Models (LR and AutoML)

Training data	Model Type	R^2/RMSE
Fine-tuning training set	LR[a] (Baseline model)	0.0689/0.0635
Fine-tuning training set	AutoML[b] (Baseline model)	0.1915/0.0592
20% of pretraining data	Pretrained Model	Before Fine-tuning: −1.6689/0.1076 After Fine-tuning: 0.2651/0.0565
50% of pretraining data	Pretrained Model	Before Fine-tuning: −0.4882/0.0803 After Fine-tuning: 0.4199/0.0502
90% of pretraining data	Pretrained Model	Before Fine-tuning: −0.2278/0.0780 After Fine-tuning: 0.6373/0.0397

[a] LR represents linear regression model; AutoML. [b] Refers to the best-performing tree-based model selected by the automated machine learning process, which includes XGBoost, LightGBM, Random Forest (RF), and Extra-Trees.

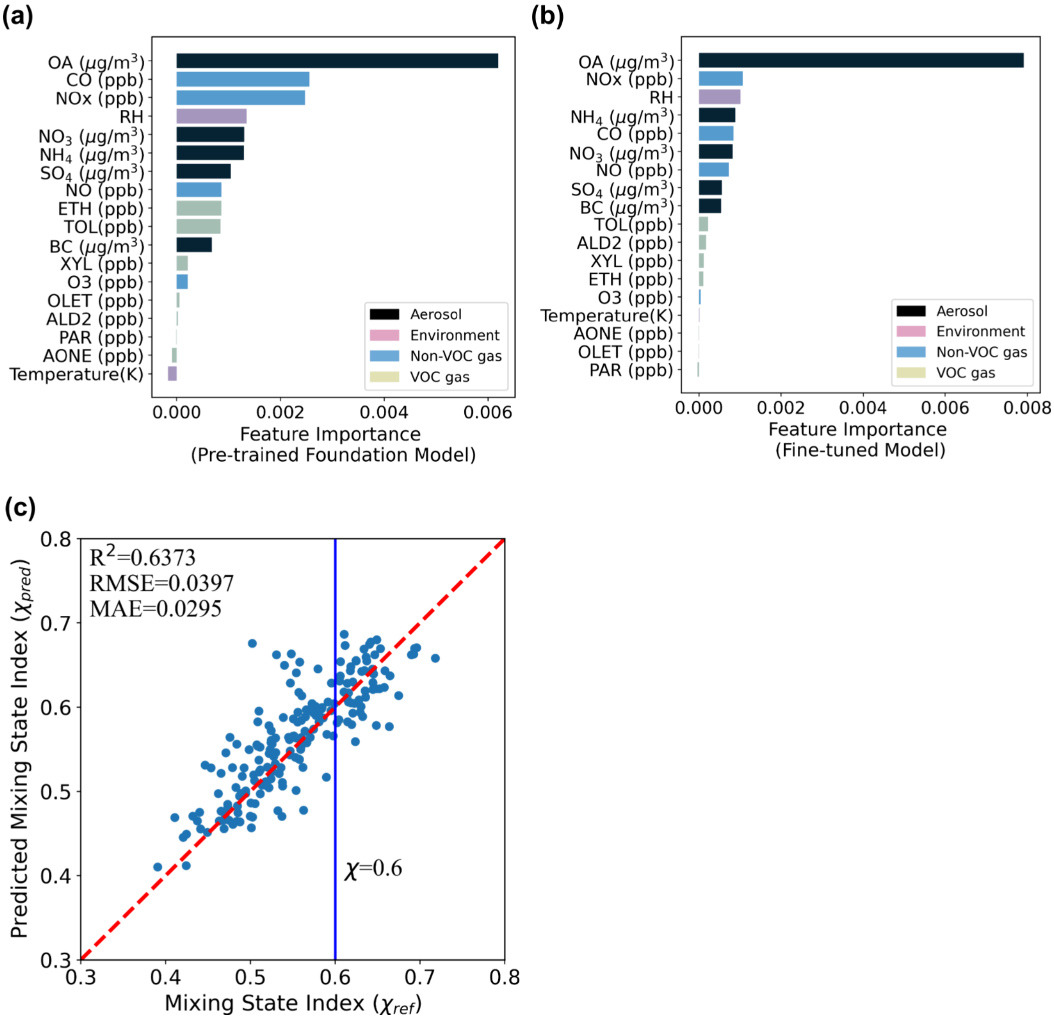


Figure 2. (a) Feature importance of the pretrained foundation model, evaluated using PFI on the fine-tuning test set. (b) The feature importance of the fine-tuned model (trained by 100% of fine-tuning training data) was evaluated using PFI on the fine-tuning test set. (c) Performance of the optimal fine-tuned model trained by 100% of fine-tuning training data, including R^2, RMSE and MAE on the fine-tuning testing set.

is unimportant. In this study, PFI was calculated multiple times using various permutations of different features, with the number of calculations set at 100.

■ RESULTS AND DISCUSSION

Impact of Pretraining Data Volume on Fine-Tuning Performance. The optimized foundation model, determined through hyperparameter tuning, comprises 15 residual blocks, each with 512 hidden units. Additional details are presented in Text S2 and Figure S3. Before fine-tuning, all pretrained models exhibited negative R^2 (Table 5), indicating that the predictions of the model are worse than simply using the mean of the observed data as a prediction. This poor performance cloud be attributed to two main factors. First, although the pretraining data was designed to capture a broad range of feature variations and dependencies related to χ, they do not reflect the actual feature distribution in real-world conditions (Figure S2 in the SI). Second, the calculation of χ depends on aerosol mass

Table 6. Performance Comparisons among the Fine-Tuned Models, AutoML, and Linear Regression on Different Fine-Tuning Training Sets[c]

Training data group (Data size)	Model Type	R^2	RMSE	MAE	MAPE (%)
10% of training data (20)	LR[a]	−118.27	0.7192	0.6025	111.16
10% of training data (20)	AutoML[b]	−1.2036	0.0978	0.0810	15.75
10% of training data (20)	Fine-tuning	0.1758	0.0598	0.0412	7.63
20% of training data (39)	LR	−1.9922	0.1139	0.0864	15.80
20% of training data (39)	AutoML	0.2016	0.0588	0.0476	9.04
20% of training data (39)	Fine-tuning	0.3634	0.0525	0.0390	7.15
40% of training data (78)	LR	0.2207	0.0581	0.0465	8.43
40% of training data (78)	AutoML	0.2012	0.0589	0.0475	8.99
40% of training data (78)	Fine-tuning	0.5133	0.0459	0.0353	6.46
50% of training data (97)	LR	0.3516	0.0530	0.0409	7.42
50% of training data (97)	AutoML	0.3318	0.0538	0.0435	8.05
50% of training data (97)	Fine-tuning	0.5577	0.0438	0.0334	6.15
80% of training data (156)	LR	0.1022	0.0624	0.0513	9.80
80% of training data (156)	AutoML	0.1408	0.0610	0.0501	9.44
80% of training data (156)	Fine-tuning	0.6084	0.0412	0.0316	5.87
100% of training data (195)	LR	0.0689	0.0635	0.0527	10.14
100% of training data (195)	AutoML	0.1915	0.0592	0.0491	9.43
100% of training data (195)	Fine-tuning	0.6373	0.0397	0.0295	5.41

[a]LR represents the linear regression model; AutoML. [b]Refers to the best-performing tree-based model selected by the automated machine learning process, which includes XGBoost, LightGBM, Random Forest (RF), and Extra-Trees. [c]The model performance was evaluated on the fine-tuning testing set.

concentration and compositional differences between individual aerosol particles. Without the contextual support of real-world data, the input features in the pretrained model are insufficient to capture these compositional differences. This limitation is evident in the PFI results for the pretrained model (Figure 2(a)), which show a relatively even distribution of feature importance, reflecting the model's generalization over simulated scenarios.

By comparison, all fine-tuned model outperforms the baseline models (Table 5). the fine-tuned model demonstrates a more focused feature importance distribution (Figure 2(b)). Key variables, such as OA, gain prominence, while the importance of certain VOC-related features (e.g., ETH and TOL) diminishes. This shift suggests that fine-tuning with observational data provides the important context to align the model with real-world feature distributions, thereby enhancing its ability to capture critical aerosol processes.

Table 5 clearly shows that larger pretraining data sets lead to better fine-tuned model performance. Substantially reducing the pretraining data volume (e.g., using only 20% instead of 90%) undermines the quality of the learned representations, thereby limiting the effectiveness of the fine-tuning stage. Given that the fine-tuning data is limited, its improvement is heavily dependent on the richness and quality of the information extracted during pretraining. In other words, while fine-tuning with observational data is essential for achieving high performance, it cannot function independently but should be built upon a robust and well-informed pretrained model. Adequate simulation data during pretraining plays a critical role in constructing a solid initial model.

Effects of Data Sparsity or Abundance on Fine-Tuned Model Performance. Table 6 compares the performance of the fine-tuned model with baseline models (AutoML and LR) across different training data sizes. The results show that fine-tuned models consistently outperformed the baseline models. As the fraction of training data increased from 10% to 50%, the performance of all models improved to varying degrees, with smaller data sets resulting in poorer results. Notably, all models exhibited poor performance when the training data was limited to 20% or less of the total, corresponding to fewer than 40 samples. Figure S5 illustrates the changes in feature importance as the amount of fine-tuning data increases. When the training data proportion does not exceed 50%, the magnitude and ranking of feature importance are adjusted compared to the base model, yet the overall distribution remains relatively uniform. For example, the feature importance of OA even decreases, while some VOCs such as ETH and TOL still retain a certain degree of importance. This suggests that although a limited amount of fine-tuning data can provide some information, it is insufficient to fully align the model with the true feature distribution, leading to relatively poorer overall performance.

AutoML and Linear Regression (LR) are regarded as "empirical models" because they rely solely on patterns and correlations found in the data, without using any knowledge or process-based models to explain how things work. When the fraction of the fine-tuning training set exceeds 50%, the performance of empirical models (AutoML and LR) began to degrade as the data quantity increased. Figure S4(a) shows that the additional data included in the 80% and 100% fine-tuning training sets mainly consists of χ values between 0.6 and 0.7, whereas only about 30% of the fine-tuning testing data falls within this range. This discrepancy in distribution reduces the consistency between the fine-tuning training and testing sets, introducing noise. This inconsistency, combined with the small testing set of 195 samples, increased sensitivity to such discrepancies, leading to poorer performance of empirical models. In contrast, the performance of the fine-tuned model steadily improves as its feature importance progressively aligns with key variables, OA becomes increasingly dominant, while the importance of certain VOC-related features (ETH and TOL) declines (Figure S5).

Figure 3 shows the temporal variation of the mixing state index (χ) over time, with the black sold line representing the reference mixing state index (χ_{ref}), and the colored dashed lines

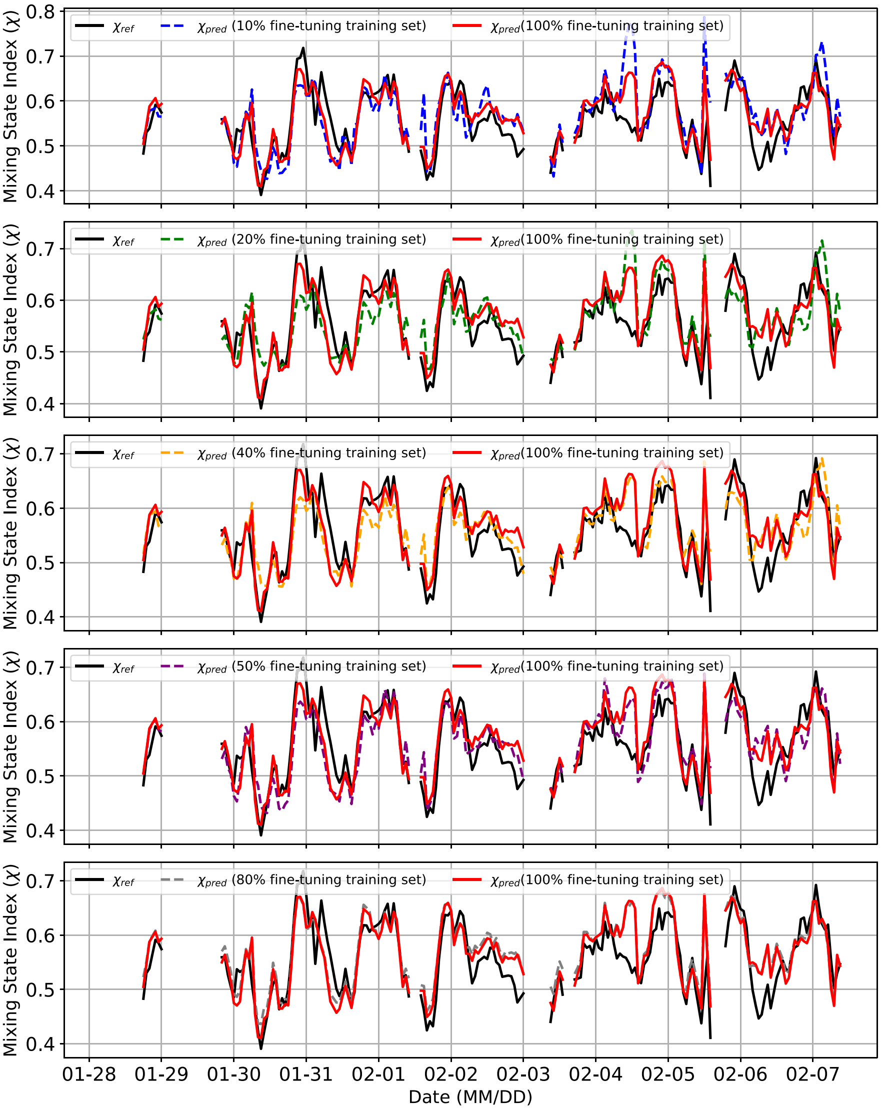


Figure 3. Comparison of model results obtained using different fractions of the fine-tuning training data. Each case is individually compared against both the reference data and the best-performing model (fine-tuned with 100% of the training set).

indicating predictions based on varying scales of fine-tuning training sets (10%, 20%, 40%, 50%, 80%, 100%). As the training set size increases, predictions align more closely with the reference (χ_{ref}). However, between February 4 and 5, the fine-tuned model consistently exhibits varying degrees of overestimation. This is mainly because the model learned from the fine-tuning training data that higher concentrations of gaseous components generally correspond to higher χ values (Figure S6); yet, around February 4, the observed χ values dropped sharply, failing to reach the high levels expected from the increased gaseous concentrations. Moreover, when the non-VOC gaseous features were removed, the fine-tuned model no longer showed overestimation during February 4−5, which further substantiates this point (Figure 4). This discrepancy indicates that the current set of input features used to describe the composition of individual particles is limited.

The results of the "Swapped Temporal Order" experiment (Table S3 in the SI) indicate that traditional models (LR and AutoML) exhibit significant fluctuations, suggesting that the data may experience temporal distribution drift. Traditional models are susceptible to local changes in data distribution and struggle to capture common patterns across different time segments, making them more susceptible to short-term or localized temporal variations. In contrast, whether using the original temporal order (with the first half of the data for training and the second half for testing) (Table 6) or the swapped order (with the latter half for training and the first half for testing) (Table S3), the performance of the fine-tuned model steadily improves as the training data increases, consistently outperforming the baseline models. Notably, with a comparable amount of data in the swapped temporal order experiment, the R^2 variation of the fine-tuned model is only about 6%. This is attributed to the integration of pretrained knowledge, which enables the model to capture common features across different time periods rather than merely fitting short-term fluctuations, thereby enhancing its robustness against temporal distribution drift. It is important to note that our current experiments are based on data spanning only about two months, so the impact of long-term temporal drift remains to be evaluated.

Integrated Feature Importance and Sensitivity Analysis for Aerosol Mixing State Prediction Models. This study integrates Permutation Feature Importance (PFI) and

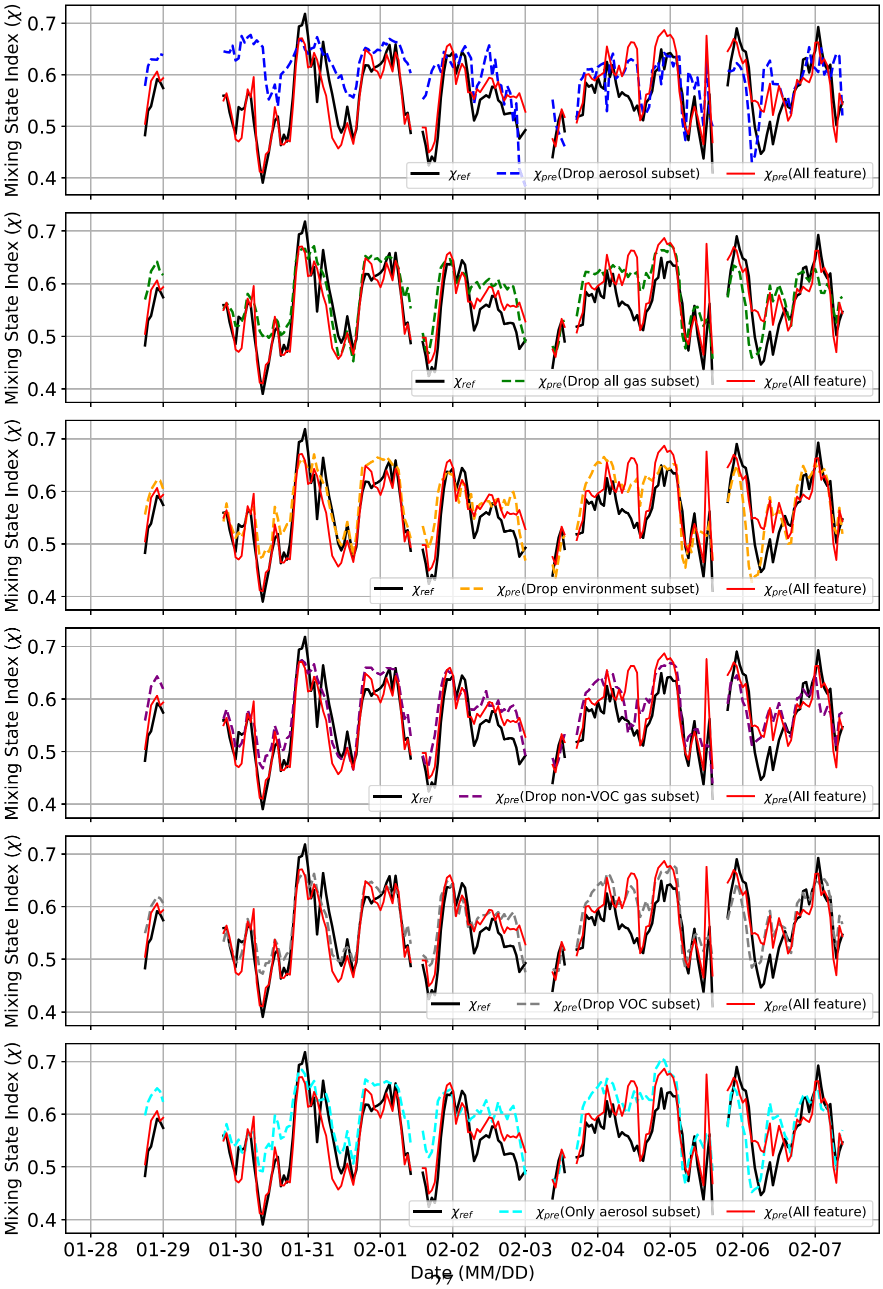


Figure 4. Comparison of model results obtained using different feature subset combinations. Each case is individually compared against both the reference data and the best-performing model (fine-tuned with all input features).

sensitivity analysis to assess the impact of different features on predicting aerosol mixing state with the best fine-tuned model from the previous section. The model, trained on the 100% of the fine-tuning training set, achieves an R^2 value of 0.6373 on the fine-tuning testing set, indicating good agreement with observational aerosol mixing state data (Figure 2(c)). However, it tends to overestimate mixing state index values (χ) below 0.6 due to discrepancies between the fine-tuning training and testing data sets.

Figure 2(b) indicates that aerosol mass concentrations impact their mixing state, with OA showing the highest feature importance due to its greater mass concentration compared to other aerosol species (BC, NH_4, SO_4 and NO_3) (Figure S4(b) and Figure S4(c)). This highlights OA's key role in predicting the mixing state index in the MEGAPOLI project. Sensitivity analysis confirms this, as shown in Table 7 and Figure 4, when aerosol mass concentration data is removed, it fails to capture the variability in χ. This reduction is expected since aerosol concentrations are central to calculating the mixing state index, reflecting aerosol population diversity. While models using only aerosol mass concentration data outperform those without aerosol data, they still lack detailed information on individual particle composition, which leads to suboptimal model performance.

As illustrated in Table 7, removing environmental data (T and RH) has a similar impact on model performance as removing all gas data (including VOC and non-VOC gases), with R^2 dropping to 0.4827 and 0.4686, and RMSE increasing to 0.0480 and 0.0431, respectively. The combination of environmental and gas data yields better model performance than using

Table 7. Fine-Tuning Performance with Different Feature Combinations

Feature sets	Input size	R^2	RMSE	MAE	MAPE (%)
All feature included	18	0.6373	0.0397	0.0295	5.41
Drop VOC gas	11	0.5902	0.0422	0.0350	6.53
Drop non-VOC gas	14	0.5719	0.0431	0.0362	6.74
Drop all gas	7	0.4686	0.0480	0.0398	7.45
Drop environmental data	16	0.4827	0.0474	0.0386	7.20
Drop aerosol	13	−0.3972	0.0778	0.0614	11.92
Only aerosol	5	0.1931	0.0592	0.0492	9.28

either type alone, highlighting their complementary roles in predicting aerosol composition. This finding is further supported by PFI results, which show that the importance of certain non-VOC gases (such as NO_x, NO, CO) and RH is comparable to that of certain aerosol mass concentrations, indicating their critical role in particle composition description for model prediction. Additionally, when non-VOC gas data are removed, although the overall R^2 performance is slightly reduced, certain errors, such as the widespread overestimation observed around February 4, are effectively mitigated (Figure 4). This suggests that while these features can improve model accuracy, they may also introduce some errors.

While VOC data shows the lowest sensitivity, reflected in its PFI value, this does not imply that VOC data is unimportant. The reduced sensitivity may be attributed to the winter setting of the MEGAPOLI campaign, characterized by limited photochemical activity and minimal secondary organic aerosol (SOA) formation.[80−82] In contrast, a similar study conducted in summer could yield different results, as VOC aging and subsequent SOA production may affect the particle mixing state.[81,82] Time-series analysis of χ (Figure 4) reveals that removing VOC data causes deviations, particularly on January 30 and February 1, underscoring its role in capturing the dynamics of χ. Further investigation is necessary to understand this observation fully. In contrast, the lower importance of temperature may be due to the limited variability of the observed data, which covers only one month. Finally, the type of features used has a more pronounced impact on model performance than the number of input features. Models using only aerosol features outperform those with more features that exclude aerosol data.

For future field campaigns, defining the "minimum data requirements" for effective model training in aerosol mixing state studies could be essential. Key elements include the mass concentration of aerosol species (OA, BC, NH_4, SO_4, and NO_3) and environmental parameters (temperature and relative humidity), as indicated by the impact analysis of this study. Combining aerosol and relevant gas data better captures aerosol composition diversity, thereby enhancing model accuracy. A representative sample size is also crucial for robust model performance across varying environments. In this case, where aerosols are primarily from local sources with limited variability,[6,59,60] 3—4 days of hourly data can achieve reliable model performance (Table 6). However, for complex aerosol sources or long-term, cross-seasonal predictions, more extensive data are needed to capture changes in mixing state indexes.

Implications. In this study, fine-tuned models incorporating process-based knowledge demonstrate enhanced robustness and stability under data quality uncertainties, consistently outperforming empirical models when applied to MEGAPOLI winter data. Our foundation and fine-tuned models are scalable and capable of continuous learning, allowing them to adapt to different scenarios more accurately. Instead of retraining the model from scratch, additional pertinent data sets can be incrementally added to update the model's knowledge, saving time and enhancing accuracy. This approach also reduces the dependence of aerosol models on detailed input data, such as emission inventories, by using the foundation model to capture relationships between aerosol mixing states and related features across various scenarios. This enables the effective use of simulation data, conserving computational resources. Our research offers a practical method for accurately studying aerosol mixing states by integrating simulations with observational data, enhancing the reliability of research findings, and providing a viable strategy for future global aerosol studies. Additionally, this framework can be applied to other fields, improving the reliability and accuracy of predictions and supporting global research initiatives.

Several strategies could be considered in future research to enhance the robustness and applicability of the model. First, the emulator in this study is fine-tuned based on the MEGAPOLI winter data set. Due to the spatial and temporal limitations of the data, it is currently applicable for estimating variations in χ during the MEGAPOLI winter period. Given the major current challenge of limited observational data, establishing a detailed mixing state database and defining standardized criteria—such as the selection of species for mixing state calculations—would be crucial. In future research, broader data sets covering more extensive spatial and temporal ranges can be incorporated, enabling the emulator to estimate χ variations in other regions, and even globally, as well as its seasonal variations. Second, integrating more advanced architectures, such as Transformers,[83] along with incorporating process-based constraints into the loss function, could improve the accuracy of process-guided fine-tuning. When larger data sets become available, it will also be essential to account for spatiotemporal patterns in deep learning models to effectively manage the vastness of the data sets and address their uneven global distribution. Approaches such as Graph Neural Networks (GNNs)[84] and data assimilation techniques could be explored to address these challenges. Insights can be drawn from AI-enabled air quality models, which face similar challenges. Third, to further enhance model versatility, integrating additional static data, such as land-use information could be considered. This integration would refine the model's predictions by providing additional context regarding ambient environmental conditions. Land-use data, for example, could help in distinguishing urban, rural, and industrial areas, improving the emulator's ability to capture aerosol variability across different settings. These improvements would strengthen the model framework, providing a solid foundation for aerosol research and extending its applicability to environmental and health studies.

■ ASSOCIATED CONTENT

Data Availability Statement

Code and simulation data to reproduce the foundation model and fine-tuned model are available at https://github.com/envdes/code_MEGAPOLI_Foundation_Model. MEGAPOLI observational data will be made available on request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.4c00329.

Detailed descriptions of the PartMC-MOSAIC and MEGAPOLI data sets, the hyperparameter results of both the pretrained model and the fine-tuned model, and the results of the "Swapped Temporal Order" experiment (PDF)

AUTHOR INFORMATION

Corresponding Authors

Zhonghua Zheng − *Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, U.K.;* orcid.org/0000-0002-0642-650X; Email: zhonghua.zheng@manchester.ac.uk

David Topping − *Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, U.K.;* orcid.org/0000-0001-8247-9649; Email: david.topping@manchester.ac.uk

Nicole Riemer − *Department of Climate, Meteorology and Atmospheric Sciences, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States;* Email: nriemer@illinois.edu

Authors

Fei Jiang − *Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, U.K.*

Hugh Coe − *Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, U.K.; National Centre for Atmospheric Sciences, The University of Manchester, Manchester M13 9PL, U.K.*

Robert M. Healy − *Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment, Conservation and Parks, Toronto M9P 3V6, Canada;* orcid.org/0000-0002-1920-9846

Laurent Poulain − *Atmospheric Chemistry Department (ACD), Leibniz Institute for Tropospheric Research (TROPOS), Leipzig 04318, Germany*

Valérie Gros − *Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, IPSL, Université Paris-Saclay, Gif-sur-Yvette 91191, France*

Hao Zhang − *Department of Earth and Environmental Sciences, The University of Manchester, Manchester M13 9PL, U.K.*

Weijun Li − *Department of Atmospheric Sciences, School of Earth Sciences, Zhejiang University, Hangzhou 310027, China;* orcid.org/0000-0003-4887-4260

Dantong Liu − *Department of Atmospheric Sciences, School of Earth Sciences, Zhejiang University, Hangzhou 310027, China;* orcid.org/0000-0003-3768-1770

Matthew West − *Department of Mechanical Science and Engineering, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.4c00329

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester, and York. The authors acknowledge the assistance given by Research IT and Computational Shared Facility 3 (CSF3) at The University of Manchester. Z.Z. appreciates the support provided by the academic start-up funds from the Department of Earth and Environmental Sciences at The University of Manchester and the Aerosol Science Career Development Grant from The Aerosol Society. We thank Reetahan Mukhopadhyay for his contributions to a related project.

REFERENCES

(1) Raes, F.; Dingenen, R. V.; Vignati, E.; Wilson, J.; Putaud, J.-P.; Seinfeld, J. H.; Adams, P. Formation and cycling of aerosols in the global troposphere. *Atmos. Environ.* **2000**, *34*, 4215−4240.

(2) Pöschl, U. Atmospheric Aerosols: Composition, Transformation, Climate and Health Effects. *Angew. Chem., Int. Ed.* **2005**, *44*, 7520−7540.

(3) Després, V. R.; Huffman, J. A.; Burrows, S. M.; Hoose, C.; Safatov, A. S.; Buryak, G.; Fröhlich-Nowoisky, J.; Elbert, W.; Andreae, M. O.; Pöschl, U.; Jaenicke, R. Primary biological aerosol particles in the atmosphere: a review. *Tellus B: Chemical and Physical Meteorology* **2022**, *64*, 15598.

(4) Winkler, P. The growth of atmospheric aerosol particles as a function of the relative humidity—II. An improved concept of mixed nuclei. *J. Aerosol Sci.* **1973**, *4*, 373−387.

(5) Riemer, N.; Ault, A. P.; West, M.; Craig, R. L.; Curtis, J. H. Aerosol Mixing State: Measurements, Modeling, and Impacts. *Reviews of Geophysics* **2019**, *57*, 187−249.

(6) Healy, R. M.; Riemer, N.; Wenger, J. C.; Murphy, M.; West, M.; Poulain, L.; Wiedensohler, A.; O'Connor, I. P.; McGillicuddy, E.; Sodeau, J. R.; Evans, G. J. Single particle diversity and mixing state measurements. *Atmospheric Chemistry and Physics* **2014**, *14*, 6289−6299.

(7) Bondy, A. L.; Bonanno, D.; Moffet, R. C.; Wang, B.; Laskin, A.; Ault, A. P. The diverse chemical mixing state of aerosol particles in the southeastern United States. *Atmospheric Chemistry and Physics* **2018**, *18*, 12595−12612.

(8) Lee, A. K.; Rivellini, L.-H.; Chen, C.-L.; Liu, J.; Price, D. J.; Betha, R.; Russell, L. M.; Zhang, X.; Cappa, C. D. Influences of Primary Emission and Secondary Coating Formation on the Particle Diversity and Mixing State of Black Carbon Particles. *Environ. Sci. Technol.* **2019**, *53*, 9429−9438.

(9) Yu, C.; Liu, D.; Broda, K.; Joshi, R.; Olfert, J.; Sun, Y.; Fu, P.; Coe, H.; Allan, J. D. Characterising mass-resolved mixing state of black carbon in Beijing using a morphology-independent measurement method. *Atmospheric Chemistry and Physics* **2020**, *20*, 3645−3661.

(10) Ye, Q.; Gu, P.; Li, H. Z.; Robinson, E. S.; Lipsky, E.; Kaltsonoudis, C.; Lee, A. K.; Apte, J. S.; Robinson, A. L.; Sullivan, R. C.; Presto, A. A.; Donahue, N. M. Spatial Variability of Sources and Mixing State of Atmospheric Particles in a Metropolitan Area. *Environ. Sci. Technol.* **2018**, *52*, 6807−6815.

(11) Rissler, J.; Nordin, E. Z.; Eriksson, A. C.; Nilsson, P. T.; Frosch, M.; Sporre, M. K.; Wierzbicka, A.; Svenningsson, B.; Löndahl, J.; Messing, M. E.; Sjogren, S.; Hemmingsen, J. G.; Loft, S.; Pagels, J. H.; Swietlicki, E. Effective Density and Mixing State of Aerosol Particles in a Near-Traffic Urban Environment. *Environ. Sci. Technol.* **2014**, *48*, 6300−6308.

(12) Schutgens, N. A. J.; Stier, P. A pathway analysis of global aerosol processes. *Atmospheric Chemistry and Physics* **2014**, *14*, 11657−11686.

(13) Jacobson, M. Z. Strong radiative heating due to the mixing state of black carbon in atmospheric aerosols. *Nature* **2001**, *409*, 695−697.

(14) Chung, S. H.; Seinfeld, J. H. Global distribution and climate forcing of carbonaceous aerosols. *Journal of Geophysical Research: Atmospheres* **2002**, *107*, AAC 14-1.

(15) Fierce, L.; Bond, T. C.; Bauer, S. E.; Mena, F.; Riemer, N. Black carbon absorption at the global scale is affected by particle-scale diversity in composition. *Nature Communication* **2016**, *7*, 12361.

(16) Fierce, L.; Riemer, N.; Bond, T. C. Toward Reduced Representation of Mixing State for Simulating Aerosol Effects on

Climate. *Bulletin of the American Meteorological Society* **2017**, *98*, 971−980.

(17) Liu, D.; et al. Black-carbon absorption enhancement in the atmosphere determined by particle mixing state. *Nature Geoscience* **2017**, *10*, 184−188.

(18) Yao, Y.; Curtis, J. H.; Ching, J.; Zheng, Z.; Riemer, N. Quantifying the effects of mixing state on aerosol optical properties. *Atmospheric Chemistry and Physics* **2022**, *22*, 9265−9282.

(19) Wang, J.; Cubison, M. J.; Aiken, A. C.; Jimenez, J. L.; Collins, D. R. The importance of aerosol mixing state and size-resolved composition on CCN concentration and the variation of the importance with atmospheric aging of aerosols. *Atmospheric Chemistry and Physics* **2010**, *10*, 7267−7283.

(20) Ching, J.; Riemer, N.; West, M. Black carbon mixing state impacts on cloud microphysical properties: Effects of aerosol plume and environmental conditions. *Journal of Geophysical Research: Atmospheres* **2016**, *121*, 5990−6013.

(21) Ching, J.; Fast, J.; West, M.; Riemer, N. Metrics to quantify the importance of mixing state for CCN activity. *Atmospheric Chemistry and Physics* **2017**, *17*, 7445−7458.

(22) Shen, W.; Wang, M.; Riemer, N.; Zheng, Z.; Liu, Y.; Dong, X. Improving BC Mixing State and CCN Activity Representation With Machine Learning in the Community Atmosphere Model Version 6 (CAM6). *Journal of Advances in Modeling Earth Systems* **2024**, *16*, No. e2023MS003889.

(23) Ching, J.; Riemer, N.; West, M. Impacts of black carbon mixing state on black carbon nucleation scavenging: Insights from a particle-resolved model. *Journal of Geophysical Research: Atmospheres* **2012**, *117*, D23209.

(24) Ching, J.; Kajino, M. Aerosol mixing state matters for particles deposition in human respiratory system. *Sci. Rep.* **2018**, *8*, 8864.

(25) Ching, J.; Kajino, M.; Matsui, H. Resolving aerosol mixing state increases accuracy of black carbon respiratory deposition estimates. *One Earth* **2020**, *3*, 763−776.

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

(27) O'Brien, R. E.; Wang, B.; Laskin, A.; Riemer, N.; West, M.; Zhang, Q.; Sun, Y.; Yu, X.; Alpert, P.; Knopf, D. A.; Gilles, M. K.; Moffet, R. C. Chemical imaging of ambient aerosol particles: Observational constraints on mixing state parameterization. *Journal of Geophysical Research: Atmospheres* **2015**, *120*, 9591−9605.

(28) Riemer, N.; West, M.; Zaveri, R. A.; Easter, R. C. Simulating the evolution of soot mixing state with a particle-resolved aerosol model. *Journal of Geophysical Research: Atmospheres* **2009**, *114*, D09202.

(29) Zhu, S.; Sartelet, K. N.; Healy, R. M.; Wenger, J. C. Simulation of particle diversity and mixing state over Greater Paris: a model−measurement inter-comparison. *Faraday Discuss.* **2016**, *189*, 547−566.

(30) Zheng, Z.; Curtis, J. H.; Yao, Y.; Gasparik, J. T.; Anantharaj, V. G.; Zhao, L.; West, M.; Riemer, N. Estimating Submicron Aerosol Mixing State at the Global Scale With Machine Learning and Earth System Modeling. *Earth and Space Science* **2021**, *8*, 9479−9496.

(31) Bommasani, R.; Hudson, D. A.; Adeli, E.; et al. On the Opportunities and Risks of Foundation Models. *ArXiv preprint* **2022−07−12**, DOI: 10.48550/arXiv.2108.07258. (Accessed: 2025−02−09)

(32) Olson, M. H.; Hergenhahn, B. R. *An introduction to theories of learning*, ninth ed. ed.; Psychology Press: New York, 2016; OCLC: 914472558.

(33) Chen, M.; Tworek, J.; Jun, H.; et al. Evaluating Large Language Models Trained on Code. *ArXiv preprint* **2021−07−14**, DOI: 10.48550/arXiv.2107.03374. (Accessed: 2025−02−09)

(34) Neelakantan, A.; Xu, T.; Puri, R.; et al. Text and Code Embeddings by Contrastive Pre-Training. *ArXiv preprint* **2022−01−24**, DOI: 10.48550/arXiv.2201.10005. (Accessed: 2025−02−09)

(35) Zhou, C.; Li, Q.; Li, C.; et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv preprint* **2023−05−01**, DOI: 10.48550/arXiv.2302.09419. (Accessed: 2025−02−09)

(36) Chen, S.; Long, G.; Jiang, J.; et al. Foundation Models for Weather and Climate Data Understanding: A Comprehensive Survey. *ArXiv preprint* **2023−12−05**, DOI: 10.48550/arXiv.2312.03014. (Accessed: 2025−02−09)

(37) Nguyen, T.; Brandstetter, J.; Kapoor, A.; et al. ClimaX: A foundation model for weather and climate. *ArXiv preprint* **2023−12−18**, DOI: 10.48550/arXiv.2301.10343. (Accessed: 2025−02−09)

(38) Chen, K.; Han, T.; Gong, J.; et al. FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead. *ArXiv preprint* **2023−04−06**, DOI: 10.48550/arXiv.2304.02948. (Accessed: 2025−02−09)

(39) Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533−538.

(40) Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; Li, H. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science* **2023**, *6*, 190.

(41) Lam, R.; Sanchez-Gonzalez, A.; Willson, M. et al.; GraphCast: Learning skillful medium-range global weather forecasting. *ArXiv preprint*, 2023-08-04; DOI: 10.48550/arXiv.2212.12794. (Accessed: 2025−02−09).

(42) Pathak, J.; Subramanian, S.; Harrington, P.; et al. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *ArXiv preprint* **2022−02−22**, DOI: 10.48550/arXiv.2202.11214. (Accessed: 2025−02−09)

(43) Man, X.; Zhang, C.; Feng, J.; et al. W-MAE: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *ArXiv preprint* **2023−12−15**, DOI: 10.48550/arXiv.2304.08754. (Accessed: 2025−02−09)

(44) Webersinke, N.; Kraus, M.; Bingler, J. A.; et al. ClimateBert: A Pretrained Language Model for Climate-Related Text. *ArXiv preprint* **2022−12−17**, DOI: 10.48550/arXiv.2110.12010. (Accessed: 2025−02−09)

(45) Yuan, Y.; Lin, L. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 474−487.

(46) Cong, Y.; Khanna, S.; Meng, C.; et al. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. *ArXiv preprint* **2023−01−15**, DOI: 10.48550/arXiv.2207.08051. (Accessed: 2025−02−09)

(47) Reed, C. J.; Gupta, R.; Li, S.; et al. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. *ArXiv preprint* **2023−09−22**, DOI: 10.48550/arXiv.2212.14532. (Accessed: 2025−02−09)

(48) Zaveri, R. A.; Easter, R. C.; Fast, J. D.; Peters, L. K. Model for Simulating Aerosol Interactions and Chemistry (MOSAIC). *Journal of Geophysical Research: Atmospheres* **2008**, *113*, D13204.

(49) DeVille, R. E. L.; Riemer, N.; West, M. Weighted Flow Algorithms (WFA) for stochastic particle coagulation. *Reviews of Geophysics* **2011**, *230*, 8427−8451.

(50) DeVille, L.; Riemer, N.; West, M. Convergence of a generalized Weighted Flow Algorithm for stochastic particle coagulation. *Journal of Computational Dynamics* **2019**, *6*, 69.

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

(52) Zaveri, R. A.; Easter, R. C.; Peters, L. K. A computationally efficient Multicomponent Equilibrium Solver for Aerosols (MESA). *Journal of Geophysical Research: Atmospheres* **2005**, *110*, D24203.

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

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

(55) Levakov, G.; Rosenthal, G.; Shelef, I.; Raviv, T. R.; Avidan, G. From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. *Human Brain Mapping* **2020**, *41*, 3235−3252.

(56) Zhang, F.; Li, Z.; Zhang, B.; Du, H.; Wang, B.; Zhang, X. Multimodal deep learning model for auxiliary diagnosis of Alzheimer's disease. *Neurocomputing* **2019**, *361*, 185−195.

(57) Xu, H.; Lin, J.; Zhang, D.; Mo, F. Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. *Nat. Commun.* **2023**, *14*, 3095.

(58) Wan, J.; Jiang, J.-W.; Park, H. S. Machine learning-based design of porous graphene with low thermal conductivity. *Carbon* **2020**, *157*, 262−269.

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

(60) Healy, R. M.; Sciare, J.; Poulain, L.; Crippa, M.; Wiedensohler, A.; Prévôt, A. S. H.; Baltensperger, U.; Sarda-Estève, R.; McGuire, M. L.; Jeong, C.-H.; McGillicuddy, E.; O'Connor, I. P.; Sodeau, J. R.; Evans, G. J.; Wenger, J. C. Quantitative determination of carbonaceous particle mixing state in Paris using single-particle mass spectrometer and aerosol mass spectrometer measurements. *Atmospheric Chemistry and Physics* **2013**, *13*, 9479−9496.

(61) He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. *ArXiv preprint* **2025−12−10**, DOI: 10.48550/arXiv.1512.03385. (Accessed: 2025−02−09)

(62) He, K.; Zhang, X.; Ren, S.; et al. Identity Mappings in Deep Residual Networks. *ArXiv preprint* **2016−07−25**, DOI: 10.48550/arXiv.1603.05027. (Accessed: 2025−02−09)

(63) Akiba, T.; Sano, S.; Yanase, T.; et al. Optuna: A Next-generation Hyperparameter Optimization Framework. *ArXiv preprint* **2019−07−25**, DOI: 10.48550/arXiv.1907.10902. (Accessed: 2025−02−09)

(64) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *Journal of Big Data* **2016**, *3*, 9.

(65) Guo, Y.; Shi, H.; Kumar, A.; et al. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. *ArXiv preprint* **2018−11−21**, DOI: 10.48550/arXiv.1811.08737. (Accessed: 2025−02−09)

(66) Girshick, R. B.; Donahue, J.; Darrell, T.; et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *ArXiv preprint* **2014−10−22**, DOI: 10.48550/arXiv.1311.2524. (Accessed: 2025−02−09)

(67) Long, M.; Cao, Y.; Wang, J.; et al. Learning Transferable Features with Deep Adaptation Networks. *ArXiv preprint* **2015−05−27**, DOI: 10.48550/arXiv.1502.02791. (Accessed: 2025−02−09)

(68) Razavian, A. S.; Azizpour, H.; Sullivan, J.; et al. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *ArXiv preprint* **2014−05−12**, DOI: 10.48550/arXiv.1403.6382. (Accessed: 2025−02−09)

(69) Cortes, C.; Mohri, M.; Rostamizadeh, A. L2 Regularization for Learning Kernels. *ArXiv preprint* **2012−05−09**, DOI: 10.48550/arXiv.1205.2653. (Accessed: 2025−02−09)

(70) van Laarhoven, T. L2 Regularization versus Batch and Weight Normalization. *ArXiv preprint* **2017−06−16**, DOI: 10.48550/arXiv.1706.05350. (Accessed: 2025−02−09)

(71) Wang, C.; Wu, Q.; Weimer, M.; Zhu, E. FLAML: A Fast and Lightweight AutoML Library. *MLSys 2021*.

(72) Zhang, C.; Hu, Q.; Su, W.; Xing, C.; Liu, C. Satellite spectroscopy reveals the atmospheric consequences of the 2022 Russia-Ukraine war. *Science of The Total Environment* **2023**, *869*, 161759.

(73) Zheng, Z.; Fiore, A. M.; Westervelt, D. M.; Milly, G. P.; Goldsmith, J.; Karambelas, A.; Curci, G.; Randles, C. A.; Paiva, A. R.; Wang, C.; Wu, Q.; Dey, S. Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed. *Journal of Advances in Modeling Earth Systems* **2023**, *15*, No. e2022MS003099.

(74) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

(75) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *ArXiv preprint* **2016−06−10**, DOI: 10.48550/arXiv.1603.02754. (Accessed: 2025−02−09)

(76) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(77) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Machine Learning* **2006**, *63*, 3−42.

(78) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340−1347.

(79) Fisher, A.; Rudin, C.; Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* **2019**, *20*, 1−81.

(80) Hallquist, M.; et al. The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmospheric Chemistry and Physics* **2009**, *9*, 5155−5236.

(81) Jimenez, J. L.; et al. Evolution of Organic Aerosols in the Atmosphere. *Science* **2009**, *326*, 1525−1529.

(82) Xu, W.; Han, T.; Du, W.; Wang, Q.; Chen, C.; Zhao, J.; Zhang, Y.; Li, J.; Fu, P.; Wang, Z.; Worsnop, D. R.; Sun, Y. Effects of Aqueous-Phase and Photochemical Processing on Secondary Organic Aerosol Formation and Evolution in Beijing, China. *Environ. Sci. Technol.* **2017**, *51*, 762−770.

(83) Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is All you Need. *ArXiv preprint* **2017−12−06**, DOI: 10.48550/arXiv.1706.03762. (Accessed: 2025−02−09)

(84) Lam, R.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416−1421.