JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2022MS003099

Key Points:

- We developed an Automated Machine Learning workflow to evaluate the utility of incorporating multiple trace gas columns in $PM_{2.5}$ estimates
- Tropospheric trace gas columns contain signatures of $PM_{2.5}$ precursors and improve $PM_{2.5}$ estimates
- We infer or link the possible relative importance of primary versus secondary sources of $PM_{2.5}$ using Automated Machine Learning and Spearman's rank correlation

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. Zheng,
zhonghua.zheng@outlook.com

Citation:

Zheng, Z., Fiore, A. M., Westervelt, D. M., Milly, G. P., Goldsmith, J., Karambelas, A., et al. (2023). 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*, *15*, e2022MS003099. https://doi.org/10.1029/2022MS003099

Received 18 MAR 2022
Accepted 3 FEB 2023

Author Contributions:

Conceptualization: Zhonghua Zheng, Arlene M. Fiore, Daniel M. Westervelt, George P. Milly
Data curation: Zhonghua Zheng, Alexandra Karambelas
Formal analysis: Zhonghua Zheng, Arlene M. Fiore, Daniel M. Westervelt, George P. Milly, Cynthia A. Randles, Antonio R. Paiva

Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed

Zhonghua Zheng[1,2,3], Arlene M. Fiore[3,4,5,6], Daniel M. Westervelt[3,5,7], George P. Milly[3], Jeff Goldsmith[5,8], Alexandra Karambelas[3], Gabriele Curci[9,10], Cynthia A. Randles[11,12], Antonio R. Paiva[11], Chi Wang[13], Qingyun Wu[14], and Sagnik Dey[15,16]

[1]Department of Earth and Environmental Sciences, The University of Manchester, Manchester, UK, [2]National Center for Atmospheric Research, Boulder, CO, USA, [3]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA, [4]Department of Earth and Environmental Sciences, Columbia University, New York, NY, USA, [5]Data Science Institute, Columbia University, New York, NY, USA, [6]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, [7]NASA Goddard Institute for Space Studies, New York, NY, USA, [8]Department of Biostatistics, Columbia University, New York, NY, USA, [9]Department of Physical and Chemical Sciences, University of L'Aquila, L'Aquila, Italy, [10]Center of Excellence in Telesensing of Environment and Model Prediction of Severe Events (CETEMPS), University of L'Aquila, L'Aquila, Italy, [11]ExxonMobil Technology and Engineering Company, Annandale, NJ, USA, [12]Now at International Methane Emissions Observatory, United Nations Environment Program, Paris, France, [13]Microsoft Corporation, Redmond, WA, USA, [14]College of Information Sciences and Technology, Pennsylvania State University, State College, PA, USA, [15]Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India, [16]Centre of Excellence for Research on Clean Air, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India

Abstract India is largely devoid of high-quality and reliable on-the-ground measurements of fine particulate matter ($PM_{2.5}$). Ground-level $PM_{2.5}$ concentrations are estimated from publicly available satellite Aerosol Optical Depth (AOD) products combined with other information. Prior research has largely overlooked the possibility of gaining additional accuracy and insights into the sources of PM using satellite retrievals of tropospheric trace gas columns. We evaluate the information content of tropospheric trace gas columns for $PM_{2.5}$ estimates over India within a modeling testbed using an Automated Machine Learning (AutoML) approach, which selects from a menu of different machine learning tools based on the data set. We then quantify the relative information content of tropospheric trace gas columns, AOD, meteorological fields, and emissions for estimating $PM_{2.5}$ over four Indian sub-regions on daily and monthly time scales. Our findings suggest that, regardless of the specific machine learning model assumptions, incorporating trace gas modeled columns improves $PM_{2.5}$ estimates. We use the ranking scores produced from the AutoML algorithm and Spearman's rank correlation to infer or link the possible relative importance of primary versus secondary sources of $PM_{2.5}$ as a first step toward estimating particle composition. Our comparison of AutoML-derived models to selected baseline machine learning models demonstrates that AutoML is at least as good as user-chosen models. The idealized pseudo-observations (chemical-transport model simulations) used in this work lay the groundwork for applying satellite retrievals of tropospheric trace gases to estimate fine particle concentrations in India and serve to illustrate the promise of AutoML applications in atmospheric and environmental research.

Plain Language Summary Ground-level fine particle ($PM_{2.5}$) concentrations are frequently estimated with freely available satellite Aerosol Optical Depth (AOD) products. We focus on India where sparse ground-based monitoring leaves gaps in our understanding of particle concentrations and the relative importance of different sources. We use an atmospheric chemistry model to test whether satellite retrievals of tropospheric trace gas columns can provide information on the origins of $PM_{2.5}$ and improve satellite-derived $PM_{2.5}$. We created an Automated Machine Learning workflow to evaluate the utility of incorporating multiple trace gas columns in $PM_{2.5}$ estimates, which represents nonlinear relationships between predictands and predictors while freeing users from selecting and tuning a specific machine learning model. On daily and monthly time scales, we quantify the relative information content of trace gas columns, AOD, meteorological fields, and emissions. We find that incorporating trace gas columns improves $PM_{2.5}$ estimates and may also enable inference of broad characteristics of particle composition.

Funding acquisition: Arlene M. Fiore, Daniel M. Westervelt, Jeff Goldsmith, Cynthia A. Randles
Investigation: Zhonghua Zheng
Methodology: Zhonghua Zheng, Arlene M. Fiore, Daniel M. Westervelt, George P. Milly, Cynthia A. Randles, Antonio R. Paiva
Project Administration: Arlene M. Fiore, Daniel M. Westervelt
Resources: Zhonghua Zheng, Arlene M. Fiore, Daniel M. Westervelt
Software: Zhonghua Zheng, Alexandra Karambelas, Gabriele Curci, Chi Wang, Qingyun Wu
Supervision: Arlene M. Fiore, Daniel M. Westervelt
Validation: Zhonghua Zheng
Visualization: Zhonghua Zheng
Writing – original draft: Zhonghua Zheng
Writing – review & editing: Zhonghua Zheng, Arlene M. Fiore, Daniel M. Westervelt, George P. Milly, Jeff Goldsmith, Alexandra Karambelas, Gabriele Curci, Cynthia A. Randles, Antonio R. Paiva, Chi Wang, Qingyun Wu, Sagnik Dey

1. Introduction

High levels of ambient fine particles (known as $PM_{2.5}$, particles 2.5 μm in diameter or smaller) pose a major environmental issue in India. As estimated by Chowdhury et al. (2019), nearly the entire population (99.9%) in India is exposed to annual $PM_{2.5}$ exceeding the previous World Health Organization (WHO) guideline of 10 $\mu g/m^3$. The latest WHO Global Air Quality Guidelines (AQG) announced on 22 September 2021, lowered the annual AQG level of $PM_{2.5}$ to 5 $\mu g/m^3$ (World Health Organization, 2021). To tackle the issue of air pollution, the Government of India launched the National Clean Air Program in January 2019, aimed at reducing particulate pollution by 20%–30% relative to 2017 levels by 2024. Monitoring air quality and understanding pollutant sources are critical to implementing effective air quality management plans, but India mostly lacks long-term, publicly accessible, reliable (i.e., quality controlled) measurements of particle composition that enable source attribution (Bali et al., 2021; Brauer et al., 2019). Although the Central Pollution Control Board (CPCB) has maintained a routine monitoring network for total $PM_{2.5}$ mass (composition unknown) and certain gas-phase species since 2008, the density of India's monitoring network (~0.14 monitors/million persons) is lower than other developing countries such as China (1.2 monitors/million persons) and developed countries such as USA (3.4 monitors/million persons), and leaves the majority of rural India entirely unmonitored (Bali et al., 2021; Brauer et al., 2019; Karambelas et al., 2018; Ravishankara et al., 2020).

Publicly available satellite products offer the opportunity to overcome limitations in spatiotemporal coverage and estimate $PM_{2.5}$ across India by combining satellite data with other information. Satellite aerosol optical depth (AOD) is often used to estimate $PM_{2.5}$ (Hoff & Christopher, 2009; van Donkelaar et al., 2006). Columnar AOD is combined with geophysical or statistical models that ingest additional meteorological data, emission inventories, chemical transport model simulations, and/or land use to estimate $PM_{2.5}$ and achieve better performance (Brauer et al., 2016; Xu et al., 2015). Typically, these approaches require high-quality ground-based measurements for model training and validation, which is particularly challenging in India due to the country's low monitor density relative to other world regions (e.g., U.S. and China).

Importantly, the possibility of gleaning additional insights into sources of PM from satellite retrievals of tropospheric trace gases has generally been overlooked. Trace gases including sulfur dioxide, nitrogen dioxide, and ammonia are precursors to fine particles that form via chemical reactions and thus should indicate the potential to form secondary PM. Other trace gases such as carbon monoxide (a product of incomplete combustion) and formaldehyde (produced during the oxidation of numerous organic gases) may correlate with emissions of aerosols or their precursor gases and may thus indicate primary (directly emitted) PM, as well as transported pollution of particles emitted or produced upwind. We evaluate here the potential for increased accuracy of ground-level $PM_{2.5}$ estimates in India by incorporating trace gas tropospheric columns in statistical approaches relating columnar AOD to surface $PM_{2.5}$. In this study, we use a chemical transport model as a testbed to assess the potential information content in tropospheric trace gas columns retrieved from satellite instruments.

Artificial intelligence (AI) and data science methods, and machine learning (ML) methods in particular, have been developed and used in atmospheric and environmental studies over the last few years. This trend is likely to persist into the foreseeable future enabled by the rapid advances and tremendous needs in many areas, such as weather forecasting and predictions (Agrawal et al., 2019; Lagerquist et al., 2019; McGovern et al., 2017), Earth system modeling (Gentine et al., 2021; Irrgang et al., 2021; Reichstein et al., 2019), and climate analysis (Labe & Barnes, 2021; Toms et al., 2020). As an alternative to simple geophysical or statistical approaches, ML approaches such as Random Forest (RF) and Gradient Boosting have been applied to meld satellite estimates of AOD with weather and land use data to produce highly spatially and temporally resolved data sets of surface $PM_{2.5}$ concentration (Di et al., 2019; Geng et al., 2020; Rybarczyk & Zalakeviciute, 2018; Xiao et al., 2018). According to the "*No Free Lunch*" theorem (Wolpert, 1996), no one ML algorithm can be universally good for all data and problems. Instead, the nature of the problem, the data, and the purpose synergistically determine the appropriate learning algorithm for a problem. For example, a deep-learning-based model architecture trained to predict severe weather might not successfully predict an extreme air pollution episode. In some cases, given the sensitivity of the data-driven models, incorporating new predictors could shift the "ideal" learning algorithm from one to another (e.g., from linear to nonlinear). Even if the "best" learning algorithm is predefined (e.g., a neural network or a gradient boosting model), searching and tuning the hyperparameters (e.g., number of hidden layers in a neural network, or the learning rate of a gradient boosting model) usually depends on human knowledge and decisions. Furthermore, ML model development (training and selection) generally requires significant

computational resources. However, a model's performance is a reflection of how accurately it captures the importance of each feature and, accordingly, is able to extract information from them. It is thus essential to employ a tailored "best model" for each problem, since a model with better performance is more likely to provide a more accurate feature importance.

Concerns with machine learning computational efficiency have given rise to fast and economical software frameworks, known as Automated Machine Learning (AutoML) (Wang et al., 2021), in which "the user simply provides data, and the AutoML system automatically determines the approach that performs best for this particular application" (Hutter et al., 2019). AutoML frees domain scientists from selecting learners and hyperparameters and can potentially prevent suboptimal choices due to idiosyncrasies or ad-hocness. For example, Adams et al. (2020) have successfully employed AutoML (an R package "H_2O") for an optimal solution to correct low-cost air quality sensors.

In this study, we leverage the power of AutoML to evaluate the added benefit of including satellite retrievals of tropospheric trace gases in statistical models used to derive $PM_{2.5}$ estimates over India. We use a chemical transport model as a synthetic testbed for developing methods under spatially and temporally continuous ("perfect") data sets and use the AutoML as a tool to fit the regression of surface $PM_{2.5}$ given the meteorological fields, emission inventories, and satellite-like pseudo-data sets sampled from the model. Data-driven models can be derived from real world data or highly detailed geophysical (in our case, atmospheric chemistry) simulations. Here, we focus on data-driven models derived from a chemical transport model, which is driven by meteorology and emissions, without chemical data assimilation; consequently, the $PM_{2.5}$ fields over India simulated by the chemical transport model are considered "ground truth" (or "label" in Machine Learning) and "performance upper bound" in this paper. Data-driven models, like the ones we explore below, offer the chance to overcome computing limitations as a low-cost alternative to expensive high-resolution simulations with chemical transport models, and can provide new insights by revealing underlying relationships between the predictor variables and $PM_{2.5}$. Note the overarching goal of this study is not to provide regression models or $PM_{2.5}$ products. Instead, we aim to assess the improved accuracy that may be possible by incorporating satellite retrievals of tropospheric trace gases in combination with AOD and meteorological variables currently used to derive $PM_{2.5}$ from satellite products. The information obtained by blending together multiple data sets can provide guidance for developing future $PM_{2.5}$ products, especially over regions lacking widespread networks of particle mass and composition measurements.

2. Methods

2.1. GEOS-CHEM Simulations

We use simulations from the GEOS-Chem version 12.0.2 (The International GEOS-Chem User Community, 2018) chemical transport model as idealized pseudo-observations continuously available from ground-based and space-based platforms. The simulations were conducted for the year 2015 with a global 2° latitude × 2.5° longitude domain providing boundary conditions to a nested grid (0.25° latitude × 0.3125° longitude, ~25 × 30 km) and 47 non-uniform vertical layers over India (0–40°N and 60–100°E) as described in Karambelas et al. (2022). This nested grid configuration was loosely based on (Chaliyakunnel et al., 2019), which used the MERRA-2 reanalysis meteorology. Instead, we use GEOS-FP fields to achieve higher spatial resolution (Karambelas et al., 2022). We use the standard tropospheric and stratospheric chemistry (e.g., NO_X-O_X-HC-aerosol-Br with a simple secondary organic aerosol representation) and physics (Pai et al., 2020; Prashanth et al., 2021), and natural and biogenic emissions. Anthropogenic emissions are from the ECLIPSE anthropogenic emission inventory (Stohl et al., 2015) processed through the Harvard-NASA Emissions Component (Keller et al., 2014). Modeled surface $PM_{2.5}$ concentrations were previously evaluated against observations from India's Central Pollution Control Board (https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data) with findings presented in Karambelas et al. (2022). Briefly, GEOS-Chem underestimates annual average concentrations (Normalized Mean Bias (NMB) = −49%). Model performance improves during periods of higher concentrations such as pollution episodes (NMB = −33%) The model is able to reproduce spatial variations in concentrations across the country ($r^2 = 0.55$), again improving during periods of high pollution ($r^2 = 0.69$). More information on the simulations can be found in Karambelas et al. (2022).

2.2. Data Processing

The machine learning models can be expressed as

$$Y(lat, lon, t) = f(X(lat, lon, t)) \quad (1)$$

where Y (predictand) is the daily surface $PM_{2.5}$ concentration at location (*lat*, *lon*) at time *t*, and f denotes the function (machine learning model) for calculating Y. The variable X (predictor) could represent a single feature (e.g., "2-m air temperature") or a combination of features (e.g., "2-m air temperature" and "tropospheric vertical column of CO"). Note that "feature" and "field" are interchangeable when referring to "predictor." We construct surface $PM_{2.5}$ concentrations from the individual simulated chemical components (ammonium, nitrate, sulfate, black carbon, organic carbon, secondary organic aerosols, dust, and sea salt), and assume a relative humidity of 50%. Table 1 lists the features ("predictors") used in our analyses. We develop "pseudo-data sets" by sampling modeled fields at satellite overpass time. These data sets are "perfect" in the sense that no instrument noise or missed retrievals are introduced (e.g., due to clouds, etc.). Specifically, we use the Flexible Aerosol Optical Depth post-processing tool (Curci et al., 2015) to estimate AOD at 550 nm and dust AOD. These fields are sampled at 5:00 a.m. UTC to coincide with Terra's local 10:30 a.m. overpass. The tropospheric vertical columns (troposphere is defined as from the surface layer to model level 38 or about 48 hPa) of trace gases (CO, SO_2, NO_2, CH_2O, and NH_3) are sampled at 8:00 a.m. UTC to match satellite instruments (SO_2, NO_2, and CH_2O from Aura/OMI, and CO and NH_3 from Aqua/AIRS) with a local 1:30 p.m. overpass. Meteorological fields were averaged on a daily and a monthly basis for further analysis. The emission fields without daily variation were only used for monthly analyses. In other words, only the features with daily variance are utilized in daily machine learning models. Note that the "moderate PM2.5" months of April and August (Figure S1 in Supporting Information S1) were used as the hold-out samples (testing data) for validation purposes, and the remaining 10 months were used for the regionalization (see Section 2.3) and training in the AutoML workflow. The values on the ends of the $PM_{2.5}$ distribution, such as PM episodes that happened in December, are included in the training data to ensure that the training data is as representative of the full data distribution as feasible. To test model skill in estimating higher $PM_{2.5}$, we also examine the sensitivity to moving October data from training to testing, as higher $PM_{2.5}$ concentrations occur in October than in April or August (Figure S1 in Supporting Information S1). Our analysis is restricted to land grid cells (defined as land covering a fraction greater than 0.5 of any individual cell).

2.3. Delineating Geographical Regions

We perform regional analysis to facilitate comprehension of spatial patterns. Rather than define regions for our analysis based on prior studies, for example, based on climate regions (Hu et al., 2017) or $PM_{2.5}$ concentrations (Greenstone et al., 2015), we propose a simple data-driven unsupervised learning approach for regionalization (Figure 1). Our approach groups grid cells into a few regions (clusters) based on their spatiotemporal similarity. The regionalization consists of two steps: (a) Empirical Orthogonal Functions (EOFs) and Varimax Rotated EOFs (REOFs) analysis to reduce the dimensionality of the data set and capture the spatiotemporal patterns, and (b) *k*-means clustering to identify common regional patterns of variability across the EOFs.

2.3.1. EOF and REOF Analysis

Compared to supervised learning, where model performance is evaluated by a set of metrics (e.g., root-mean-square error) against validation data sets, unsupervised learning does not lend itself to quantitative evaluation. The principal component analysis (PCA) and its variant "varimax rotated PCA" have been widely applied in atmospheric and climate research, such as decomposing sea surface temperature (Lian & Chen, 2012) into REOFs to determine modes of variability. Motivated by a previous application of REOFs on the observed patterns of surface ozone (O_3) in the eastern United States (Fiore et al., 2003, 2022), we first applied PCA to derive the EOFs. The first four EOFs capture 55% of the variance in daily PM2.5 (Figure 1a). As the 5th EOF explains <5% additional variance, we select the first four EOFs and apply a varimax rotation which re-distributes the explained variance among the retained EOFs (Figure 1b). This approach should better identify regions where day-to-day variations are occurring coherently, as we might expect to occur on regional scales under the influence of common large-scale weather patterns.

2.3.2. *k*-Means Clustering

The *k*-means clustering is an unsupervised learning approach and has been applied for ecoregion delineation (Kumar et al., 2011), environmental risk zoning (Shi & Zeng, 2014), and aerosol mixing state regionalization (Zheng et al., 2020). Qualitatively, we gauge successful implementations of clustering by the emergence of spatially contiguous regions without the direct guidance of spatial information (e.g., providing the algorithm

Table 1
Features (Fields) Definitions

Type	Feature (fields)	Description	Temporal resolution
Meteorological	T2M	2-m air temperature	Daily and monthly
	RH	2-m relative humidity	
	PBLH	Planetary boundary layer height	
	U10M	10-m eastward wind	
	V10M	10-m northward wind	
	PRECTOT	Total precipitation	
Satellite (aerosol)	AOT_C	Aerosol optical thickness (or AOD) at 550 nm	
	AOT_DUST_C	Aerosol optical thickness (or AOD) of dust at 550 nm	
Satellite (trace gases)	CO_trop	tropospheric vertical column of CO	
	SO_2_trop	tropospheric vertical column of SO_2	
	NO_2_trop	tropospheric vertical column of NO_2	
	CH_2O_trop	tropospheric vertical column of CH_2O	
	NH_3_trop	tropospheric vertical column of NH_3	
Emission	EmisDST_Natural	Dust emissions from natural sources (EmisDST1_Natural + EmisDST2_Natural + EmisDST3_Natural + EmisDST4_Natural), number indicates GEOS-Chem size bin	
	EmisNO_Fert	NO emissions from fertilizer	
	EmisNO_Lightning	NO emissions from lightning	
	EmisNO_Ship	NO emissions from ships	
	EmisNO_Soil	NO emissions from soil	
	EmisBC_Anthro	Black carbon aerosol emissions from anthropogenic sources (EmisBCPI_Anthro + EmisBCPO_Anthro), "PI" refers to "hydrophilic" and "PO" refers to "hydrophobic"	Monthly
	EmisBC_BioBurn	Black carbon aerosol emissions from biomass burning (EmisBCPI_BioBurn + EmisBCPO_BioBurn)	
	EmisOC_Anthro	Organic carbon aerosol emissions from anthropogenic sources (EmisOCPI_Anthro + EmisOCPO_Anthro)	
	EmisOC_BioBurn	Black carbon aerosol emissions from biomass burning (EmisOCPI_BioBurn + EmisOCPO_BioBurn)	
	EmisCH_2O_Anthro	Formaldehyde (CH_2O) emissions from anthropogenic sources	
	EmisCH_2O_BioBurn	CH_2O emissions from biomass burning	
	EmisCO_Anthro	CO emissions from anthropogenic sources	
	EmisCO_BioBurn	CO emissions from biomass burning	
	EmisCO_Ship	CO emissions from ships	
	EmisNH_3_Anthro	NH_3 emissions from anthropogenic sources	
	EmisNH_3_BioBurn	NH_3 emissions from biomass burning	
	EmisNH_3_Natural	NH_3 emissions from natural sources	
	EmisNO_Aircraft	NO emissions from aircraft	
	EmisNO_Anthro	NO emissions from anthropogenic sources	
	EmisNO_BioBurn	NO emissions from biomass burning	
	EmisSO_2_Aircraft	SO_2 emissions from aircraft	
	EmisSO_2_Anthro	SO_2 emissions from anthropogenic sources	
	EmisSO_2_BioBurn	SO_2 emissions from biomass burning	
	EmisSO_4_Anthro	SO_4 emissions from anthropogenic sources	

Note. All fields are taken from the GEOS-Chem simulation; meteorological data and emissions fields are input to the chemical transport model simulations whereas the column concentrations are simulated output.

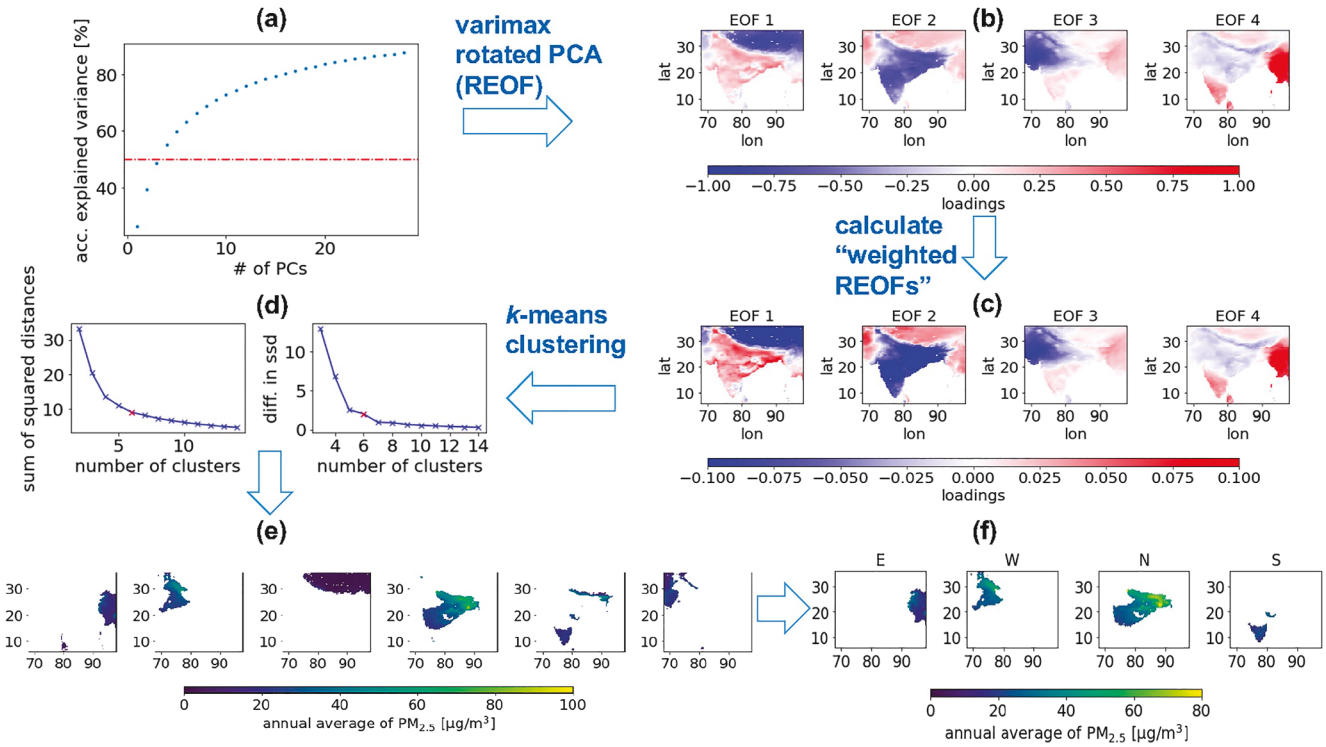

Figure 1. The workflow of delineating geographical regions. (a) Principal component analysis to derive the Empirical Orthogonal Functions (EOFs) that capture over 50% of the variance (first four EOFs); (b) varimax-rotated loadings for the selected EOFs; (c) weighted averaged loadings for the selected REOFs; (d) "Elbow method" to determine the number of regions (clusters); (e) regions based on $k = 6$ from k-means clustering; (f) four regions that intersect with India's land pixels.

with latitude and longitude). We then multiply the EOF loadings by the corresponding explained variance to produce "weighted REOFs" (Figure 1c) as the input for the k-means clustering so that Euclidean distances among them correctly capture the relationships with respect to the original feature space. We use the "elbow method" (Figure 1d) to identify an optimal trade-off point and select six clusters (Hastie et al., 2009). Then we select four regions that intersect with India's land pixels as our study areas. Note that the four regions (Figure 1e) contain not only India but also nearby countries, such as Bangladesh, Nepal, and Myanmar. Additionally, Region C (the union of four regions) and Region A (India and its neighbors, including all land grid cells within the nested grid of the simulations) are considered in this study to examine patterns at various spatial scales (see Section 2.4).

2.4. Automated Machine Learning (AutoML)

Rather than using a specific machine learning approach (e.g., RF) to build regression models and quantify the importance of various features (fields), here we use a lightweight Python library "FLAML" (a Fast and Lightweight AutoML library) (Wang et al., 2021) as the tool for the AutoML task. This library chooses a search order optimized for both computational cost and model error, and selects the learner, hyperparameters, sample size, and resampling strategy iteratively. When tested on a large open-source AutoML benchmark, FLAML has superior performance compared to the top-ranked AutoML libraries, but with much smaller computational and time budgets (Wang et al., 2021). Given our modeling formulation, we configured the AutoML for a regression task with "auto" for the estimator list, optimizing the R^2 metric, and assigned a time budget of "5,400 s" (1.5 hr) for each AutoML experiment. The "auto" scheme of ML estimator models consists in this library of tree-based approaches, namely, LightGBM (Light Gradient Boosting Machine, Ke et al., 2017), XGBoost (eXtreme Gradient Boosting, Chen & Guestrin, 2016), CatBoost (categorical boosting, Prokhorenkova et al., 2018), RF (Breiman, 2001), and Extra-Trees (Extremely randomized trees, Geurts et al., 2006). We then compare the best estimator (the specific learning algorithm/model with optimized hyperparameters) from AutoML with two baseline models: the default configurations of XGBoost (xgboost.XGBRegressor) and RF (sklearn.ensemble.RandomForestRegressor).

Table 2
Experimental Design and Core Questions

	Feature	Time scale	Region
Core questions	Do tropospheric trace gases improve PM2.5 estimates?	How does the ranking of features vary at different time scales?	How does the ranking of features vary in different regions?
Experiments	Data from the collection of all the grid cells falling into a certain region:	- daily - Monthly	- E/S/W/N: individual region from Figure 1f
	- with trace gas columns		- C: the union of four regions (E + S + W + N)
	- without trace gas columns but with AODs		
	- without trace gas columns and without AODs		- A: India and the neighboring countries (all land grid cells from the simulations, including Region C)

2.5. Experimental Design

We conduct a series of comparisons and answer three core questions by using the best estimators trained from AutoML (Table 2). First, the maximum benefit of tropospheric trace gas columns (using modeled proxies as described in Section 2.2) for surface $PM_{2.5}$ estimates can be determined by assessing the improvement in estimator performance when trace gas columns are used as features, in addition to meteorological variables, emissions, and AOD. We also test the model performance in the absence of AOD (removing total and dust AOD) when trace gases, meteorological variables, and emissions are available. Second, the same feature combination but different data (monthly vs. daily) can be used to estimate the maximum information content possible from tropospheric trace gas columns and other input variables (using "perfect" model data sets) at different time scales. Monthly estimates, in comparison to daily estimates, attempt to capture spatial patterns and seasonal cycles but are unable to incorporate daily weather data. Third, the best estimators trained on data at different-sized regions ingested on different temporal averaging periods (monthly vs. daily) provide insights on whether any benefit from including tropospheric trace gas columns is spatially equivalent. Table 3 provides an overview of the training sample size for each experiment.

2.6. Feature Importance Attribution

In Data Science, "feature importance" refers to a score that represents how useful the feature is at predicting the target variable. However, the type of feature importance score differs for different learning algorithms and results in values with varying orders of magnitude. For example, the feature importance of Extra-Trees is based on "impurity" (the normalized total reduction of the mean squared error brought by that feature), LightBGM's feature importance is by default based on "split" (the numbers of times the feature is used in a tree node in the model), and XGBoost usually uses "gain" (the average Gini impurity/information gain across all splits the feature is used in). However, the interpretability of machine learning models is still a challenging issue, and there is no consensus over which technique for determining feature importance is superior. For instance, although Shapley Additive exPlanations (Lundberg & Lee, 2017) is a unified approach to interpreting model predictions, the assumption that an ML prediction can be represented by a sum of contributions from each feature may not hold for highly nonlinear models (Gosiewska & Biecek, 2019). Therefore, we employ the default "feature importance" attribute of the approach selected by AutoML, as it is most frequently employed by domain-specific researchers.

Then, we derive a "ranking score" metric to unify the comparison of feature importance from different learning algorithms (e.g., Extra-Trees, LightGBM, and XGB). For each estimator, we rank the feature importance values from lowest to highest and assign a "ranking score" to each feature based on the rank order of the corresponding

Table 3
Training Sample Size (10 Months in the Year 2015)

Time scale	Region E	Region S	Region W	Region N	Region C	Region A
Daily	230,128	113,392	274,816	574,256	1,192,592	2,174,512
Monthly	7,570	3,730	9,040	18,890	39,230	71,530

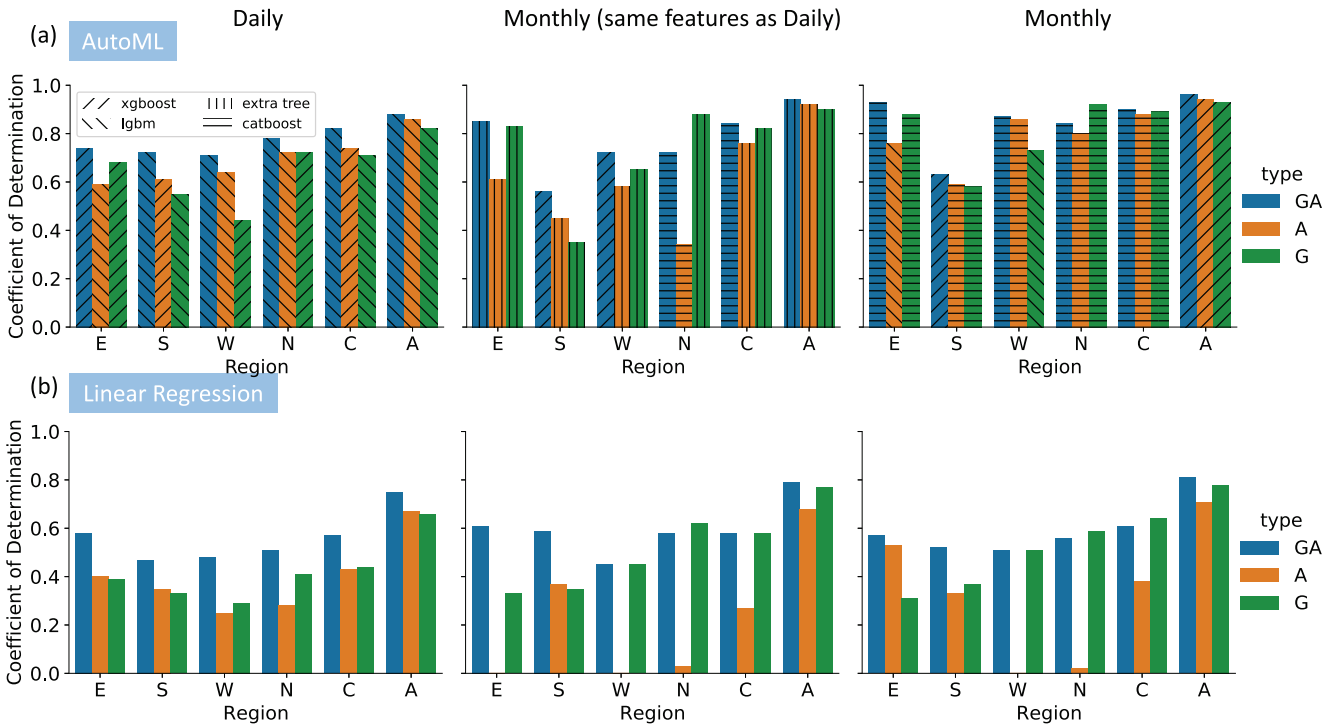

Figure 2. Improved predictive capability of $PM_{2.5}$ estimates when adding trace gas columns to other feature types in Table 1. Estimators are trained based on (a) Automated Machine Learning and (b) linear regression. The left and middle panels differ in time scale (daily vs. monthly), and the middle and right panels differ in feature numbers (see Section 2.2 and Table 1). GA: both trace gas columns and Aerosol Optical Depths (AODs) are available; A: AODs are available but trace gas columns are not available; G: trace gas columns are available but AODs are not available. Note the bars with the coefficient of determination lower than 0 are not shown, and results are based on testing data.

feature importance value. That is, the least important feature has a score of 1, the second to least important feature has a score of 2, and so on. As a result, ranking scores are bounded between 1 (least important) and the number of features (most important), which converts the feature importance of different estimators to the same scale for comparison. We also group features within the same type (Table 1) and compute the mean ranking score and the standard deviation within each type.

3. Results and Discussion

3.1. Including Multiple Trace Gas Modeled Columns Generally Improves $PM_{2.5}$ Estimates

We first evaluate whether $PM_{2.5}$ estimates improve in accuracy when we add trace gas tropospheric columns simulated by the GEOS-Chem model to the simulated meteorological variables, AOD, and emissions. We apply AutoML-derived nonlinear models and linear regression (LR). The coefficient of determination (R^2, based on an ordinary least-squares regression) between $PM_{2.5}$ simulated by GEOS-Chem versus that predicted with machine learning approaches is used as a metric for accuracy. These regressions provide average estimates, across the study area and time scales, of the association between trace gas columns and $PM_{2.5}$. As a comparison, we also apply the same approach to AODs.

While the "best estimator" from AutoML varies in space and time, we observe an increase in R^2 when simulated columnar trace gases are included in nonlinear and LR models (Figure 2), implying that trace gases contain signatures useful for $PM_{2.5}$ estimation. However, including AODs as features in the presence of trace gas columns does not guarantee improved performance, and sometimes impairs the model performance (e.g., the difference between "GA" (both trace gas columns and AODs are available) and "G" (trace gas columns are available but AODs are not available) in monthly Region N). Given that the models with nonlinear relationships exhibit higher R^2 compared to the linear model, here we focus on the results from AutoML. The comparisons between "GA" and "A" with different ML model assumptions are discussed in Section 3.4. Some emission inventories in the

model are only available at the monthly resolution, while others vary day-by-day. By comparing the results of monthly $PM_{2.5}$ estimates using only the emissions available at daily time scales ("Monthly (same features as Daily)") versus all emission inventories ("Monthly"), we find that using "all emission inventories" yields higher R^2 (e.g., 0.02 to 0.15 improvement in R^2 for "GA") at monthly time scales, implying that more accurate emission inventories (captured by the monthly emission features) will improve $PM_{2.5}$ estimates. In the following sections, we will keep the "Monthly" results that utilized emission data available at both monthly and daily time scales for analysis, since they yielded the best overall estimates of $PM_{2.5}$ in this study.

The improvements in R^2 vary spatially and temporally at the regional scale. In Region E, adding trace gas columns alongside AOD (GA) boosted daily R^2 from 0.59 to 0.74, and monthly from 0.76 to 0.93 compared to AOD alone (A). But in Region W, the increases in R^2 are moderate (+0.07 for daily and +0.01 for monthly). Given that other features already account for 86% of the variance in Region W on the monthly scale, adding trace gases only results in marginal increases in R^2 (note the improvement is not a linear addition, it reconstructs the interactions among all features, not only the interactions between other features and trace gases). Marginal increases in R^2 also occur for Region N, where the daily and monthly increases are 0.06 and 0.04, respectively. The increases in R^2 are not proportional to the baseline (without trace gases; A). For example, although the baseline monthly R^2 in Region S is relatively low, its increase is similar to other regions. The lower R^2 values for monthly $PM_{2.5}$ estimates in Region S may be due to insufficient samples, as this region's sample size is approximately one-fifth to half that of the other regions.

When training on data from the union of our four individual regions (Region C) or all the land grid cells as a whole (Region A), the inclusion of trace gases always contributes to a higher R^2. Especially, trace gases in Region C increased R^2 from 0.74 to 0.82 at the daily scale. At a larger geospatial scale (Region A), although the baseline R^2 values on the daily scale (0.86) and monthly scale (0.94) are well explained by meteorological fields, emission inventory, and AODs, the presence of trace gas columns can further explain variance (0.02 for both) in $PM_{2.5}$ and improve the estimates. However, a cost may be associated with the minor improvement, depending on the effort required to acquire additional data. As such, it is necessary to weigh the trade-offs.

3.2. The Relative Importance of Trace Gas Columns to Accurate $PM_{2.5}$ Estimates Varies Spatially and Temporally

We compare ranking scores among the types (Figure 3) and features (Figures 4 and 5), defined in Table 1, to study the relative importance of trace gas columns for improving $PM_{2.5}$ estimates over India, where a type or feature with a higher ranking score indicates its higher importance compared to other types or features in the regression. The ranking scores shown in Figures 3–5 suggest that the important features and feature interactions differ in space and time, which may be explained by regional and temporal differences in the dominant sources and the interactions with meteorological conditions.

On a daily scale, trace gas columns from GEOS-Chem (NO_2, SO_2, and CH_2O) are the most important factors that boost the performance of $PM_{2.5}$ estimates in Region E (Figure 4). The order of other types remains similar (Figure 3) when trace gas columns or AODs are not included as features in this region. Region S, however, shows that the inclusion of the trace gas columns rearranges the order of feature importance among types for daily $PM_{2.5}$ estimates (Figure 3). Without the use of trace gas columns to estimate daily $PM_{2.5}$ levels, the most significant feature type is AOD, followed by meteorological fields and emissions. When trace gas columns are considered, the relative importance of AOD decreases, and meteorological fields (e.g., V10M and planetary boundary layer) take precedence, implying that AOD and trace gas columns may contain redundant information over Region S. Meteorological fields and AODs are the most important factors for $PM_{2.5}$ estimates in Region W and Region N when trace gas columns are not available. But the trace gas columns (SO_2, NH_3, NO_2) are as important as the meteorological fields (V10M, U10M, PBLH) when they are included, implying that both secondary aerosol production via chemical reactions (e.g., formation of ammonium sulfate and nitrate) and physical processes (e.g., transport and dispersion) determine $PM_{2.5}$ distributions within these regions. The model trained from Region C (four regions as a whole) shows that AODs can explain a large fraction of the variance of $PM_{2.5}$ when trace gas columns are missing. However, with the presence of trace gas columns, meteorological fields are the most important factors that modulate $PM_{2.5}$ estimates. This discrepancy may also indicate redundant information in AODs and trace gas columns. In a larger area (Region A), regardless of the presence of trace gas columns, meteorological fields are the dominant features. The lower overall ranking score for either feature type when both AOD and trace gas columns are included may be explained by the high correlation between AODs and all trace gas columns

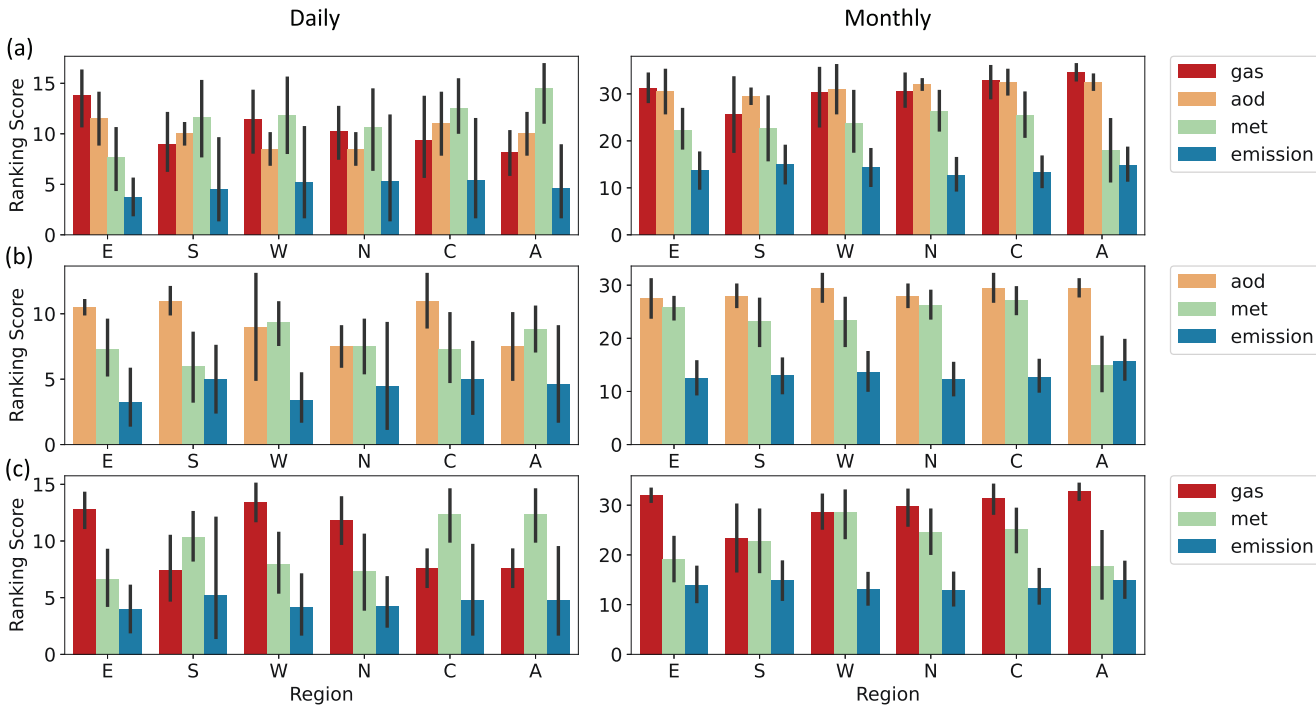

Figure 3. Ranking scores of Aerosol Optical Depth (AODs), meteorological fields, emission inventory, and trace gas column simulated by the GEOS-Chem model in estimating modeled $PM_{2.5}$. (a) Both trace gas columns and AODs are available; (b) AODs are available but trace gas columns are not available; (c) trace gas columns are available but AODs are not available. Means and standard deviations of the ranking scores are derived from Automated Machine Learning-trained "best estimators" within the same type.

(Figure S2 in Supporting Information S1), which leads to features to be selected from either type as they contain similar information. Notably, while daily estimates indicate that emissions are the least important features in all cases, this could be because much of the predictive information can be inferred from the other features, due to a scarcity of varying emission data, or because the source at daily time scales does not change spatially as much in the model.

For monthly $PM_{2.5}$ estimates, although the patterns of the relative importance of different feature types remain similar among the four regions and Region C, the specific ranking order of features varies in different regions.

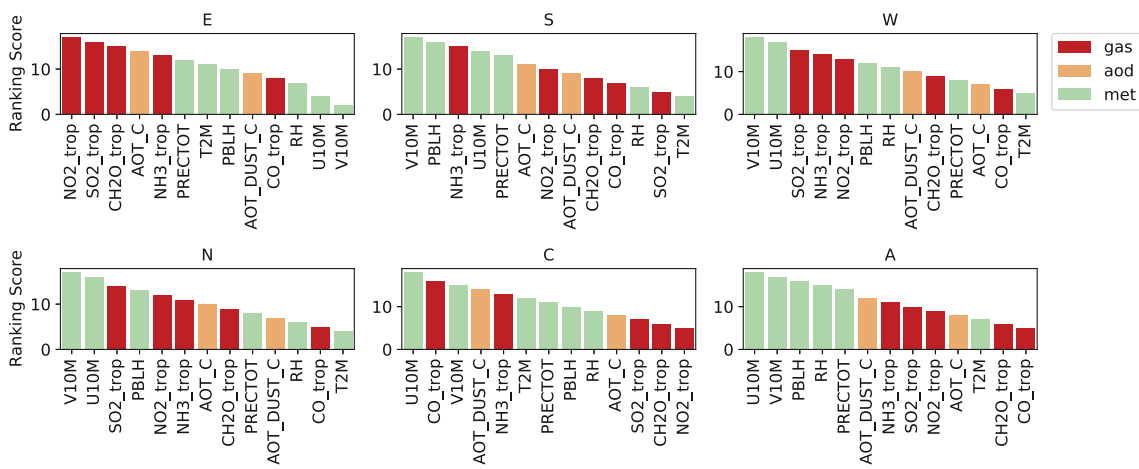

Figure 4. Ranking scores of features from Aerosol Optical Depths (AODs), meteorological fields, and trace gas columns in estimating daily $PM_{2.5}$. The ranking scores are derived from Automated Machine Learning-trained "best estimators" that incorporate both trace gas columns and AOD, as well as meteorological fields and an emission inventory. Ranking scores of emissions are not presented, but not all emissions have low ranking scores. For example, "EmisNO_Lightning" and "EmisNO_Soil" lie between "RH" and "U10M" in Region E, and "EmisNO_Fert" is between "U10M" and "SO_2_trop" in Region N.

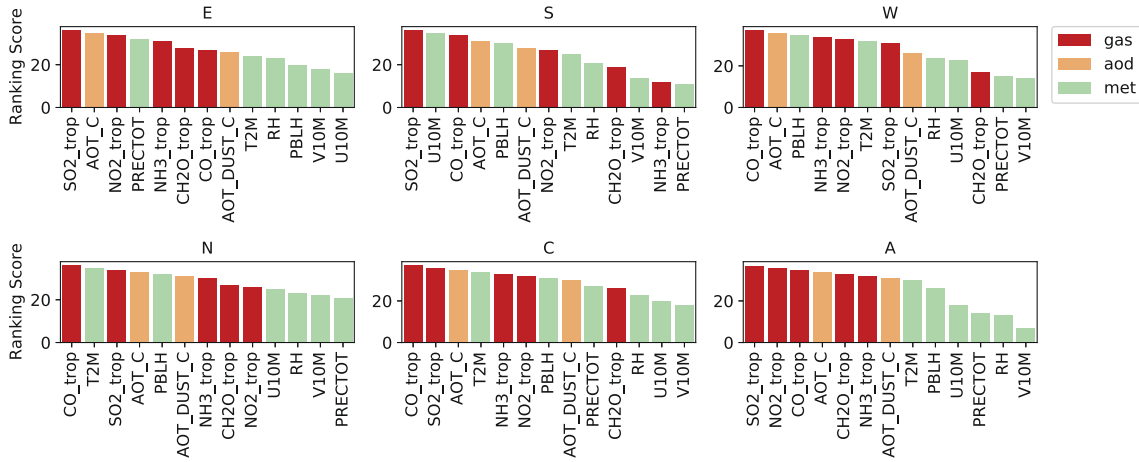

Figure 5. Ranking scores of features from Aerosol Optical Depths (AODs), meteorological fields, and trace gas columns in estimating monthly $PM_{2.5}$. The ranking scores are derived from Automated Machine Learning-trained "best estimators" that incorporate both trace gas columns and AOD, as well as meteorological fields and an emission inventory. Ranking scores of emissions are not presented, but not all emissions have low ranking scores.

The results can be partially related to the dominant sources. For example, a slightly lower mass fraction of sulfate in Region W (~13% of total $PM_{2.5}$) compared to other regions (14%–17%) corresponds to a lower ranking score of "SO2_trop." When building a model for a larger geospatial extent (Region A), the ranking score of meteorological fields declines (Figure 5). A possible reason is that the ML model for Region A assumes the same interactions among features for all regions and gives lower importance to meteorological variables compared to other types. But in reality, meteorology varies across the country and different meteorological factors may play different roles in different regions. On the other hand, monthly averaged precipitation and relative humidity in Region A are less correlated with $PM_{2.5}$ (Figure S3 in Supporting Information S1) than in other regions on monthly or daily scales, implying that information about processes (e.g., wet deposition) is diluted.

3.3. Implications for $PM_{2.5}$ Speciation

Along with the feature importance generated by AutoML, we use Spearman's rank correlation coefficient to infer the chemical composition of $PM_{2.5}$. Note that feature importance (or ranking score) and Spearman's rank correlation coefficient address different aspects of our analysis. Feature importance concerns the interactions among features and features' contribution to the predictive capability, but it does not reveal the individual relationship between target variable and each feature. On the other hand, Spearman's rank measures the monotonic relationship (whether linear or not) between two variables but does not necessarily inform on its significance in a predictive model with other features. Here we show that, in the regions where trace gas columns are associated with higher ranking scores, the correlation between trace gas columns and $PM_{2.5}$ are also high. We find that tropospheric columns of SO_2 and NO_2 contribute most to the variation of daily and monthly $PM_{2.5}$ estimates in Region E based on ranking scores, along with the highest Spearman's rank correlation coefficients. The agreement between AutoML and Spearman's rank correlation suggests secondary inorganic PM (sulfate and nitrate) are critical species modulating the $PM_{2.5}$ concentrations in Region E. The results are consistent with the identification of the source of $PM_{2.5}$ in North-Eastern India, where multivariate analysis of the collected samples indicates that sulfate and nitrate contribute the most to the variations (Khare & Baruah, 2010). Another study utilizing the MERRA-2 database indicates that SO_4 contributed to 45% of AOD over the northeastern region of India (Pathak & Bhuyan, 2022). Thus, the agreement between the ranking scores and the Spearman's correlations provide evidence for chemical speciation of $PM_{2.5}$ and thus potential to infer source attribution in the subregions. Figure 6 also suggests that SO_2 (Region S) and CO (Region W and Region N) have higher correlation coefficients with $PM_{2.5}$ compared to other trace gas columns, consistent with the monthly ranking scores (Figure 5). The findings suggest that sulfate may modulate monthly $PM_{2.5}$ variability in Region S, whereas the co-variation of the primary pollutant CO with monthly $PM_{2.5}$ may indicate co-emission or co-production with $PM_{2.5}$ in Regions W and N.

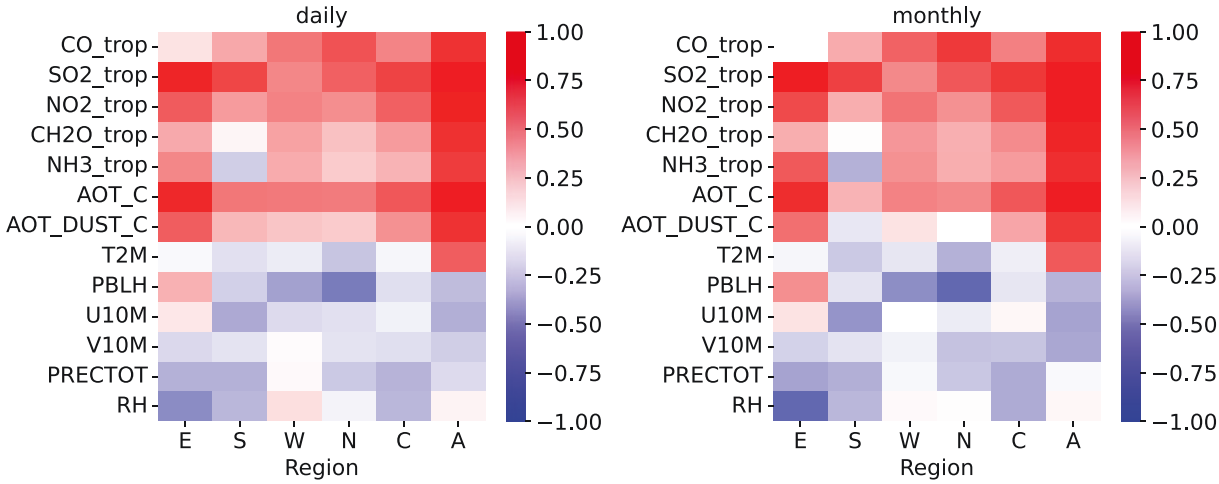

Figure 6. Spearman's rank correlation coefficient between $PM_{2.5}$ and AODs, meteorological fields, and trace gas columns.

3.4. AutoML Consistently Outperforms Baseline Machine Learning Models

We use the independent testing data to evaluate the advantage of using AutoML by comparing the "best estimator" from AutoML with LR and two commonly used nonlinear ML models: RF and XGBoost (XGB). As expected, the performance of ML models varies case by case (Figure 7a). For example, although the performance of RF is generally worse than XGBoost, it outperforms XGBoost when estimating monthly $PM_{2.5}$ in Region N. Therefore, it is not the best practice to implement the same machine learning algorithm for every region. Instead, the best estimators trained by AutoML outperform RF, XGBoost, and LR, especially at the regional scale. As shown in Figure 2, XGBoost (with improved hyperparameter configurations) is selected as the best estimator for Region E on a daily scale, while LightGBM estimators are selected for other regions. On the monthly scale, the best estimators for the same region differ. For instance, Catboost estimators, which are not chosen on a daily scale, are selected as the best estimators for Regions E, W, N, and C. Because AutoML consists of a set of different learning algorithms and explores several possible hyperparameter combinations of each algorithm when training the estimators, it is at least no worse than user-chosen models. We also show that an increase in data volume (e.g., daily compared to monthly) is likely to narrow the gap in R^2, highlighting the importance of attaining a large quantity of high-quality data for data-driven model development.

To assess whether trace gas columns can improve the $PM_{2.5}$ estimates, we repeat the regressions by implementing the above learning algorithms on a daily and monthly scale without trace gas columns, and calculate the difference in R^2 (Figure 7b). We show that no matter the choice of learning algorithms, including trace gas columns consistently results in a higher R^2. Although the most obvious improvements come from LR, the best estimators from LR are in general worse than the nonlinear models, confirming that the assumption of a nonlinear relationship between features and $PM_{2.5}$ is more appropriate. In accordance with the present results, previous studies (e.g., Porter & Heald, 2019; Tai et al., 2010) have demonstrated that the correlations between PM and meteorological conditions are complex.

3.5. Is "Big Data" Always Better?

The above analysis has shown that a larger volume of data at the same time scale (e.g., Region A) results in a better predictive capability compared to separate models for each of its subregions. However, such comparisons are potentially misleading, because the testing data sets are different and model performance is "averaged" across regions. Two questions are raised here: (a) How well does a "generalized model" trained on the larger region perform when applied to the sub-regions it encompasses ("spatial mismatch")? (b) What role do trace gas products play in the application of a "spatially mismatched" model? Both questions are in line with the emphasis of Data-Centric AI, as a generalized model attempts to make use of the additional available samples to more robustly infer the predictive relationships, but it is possible that only a part of the data is indeed useful. On the other hand, while the first principles (e.g., chemical reactions) should be universal, due to incomplete features such as human

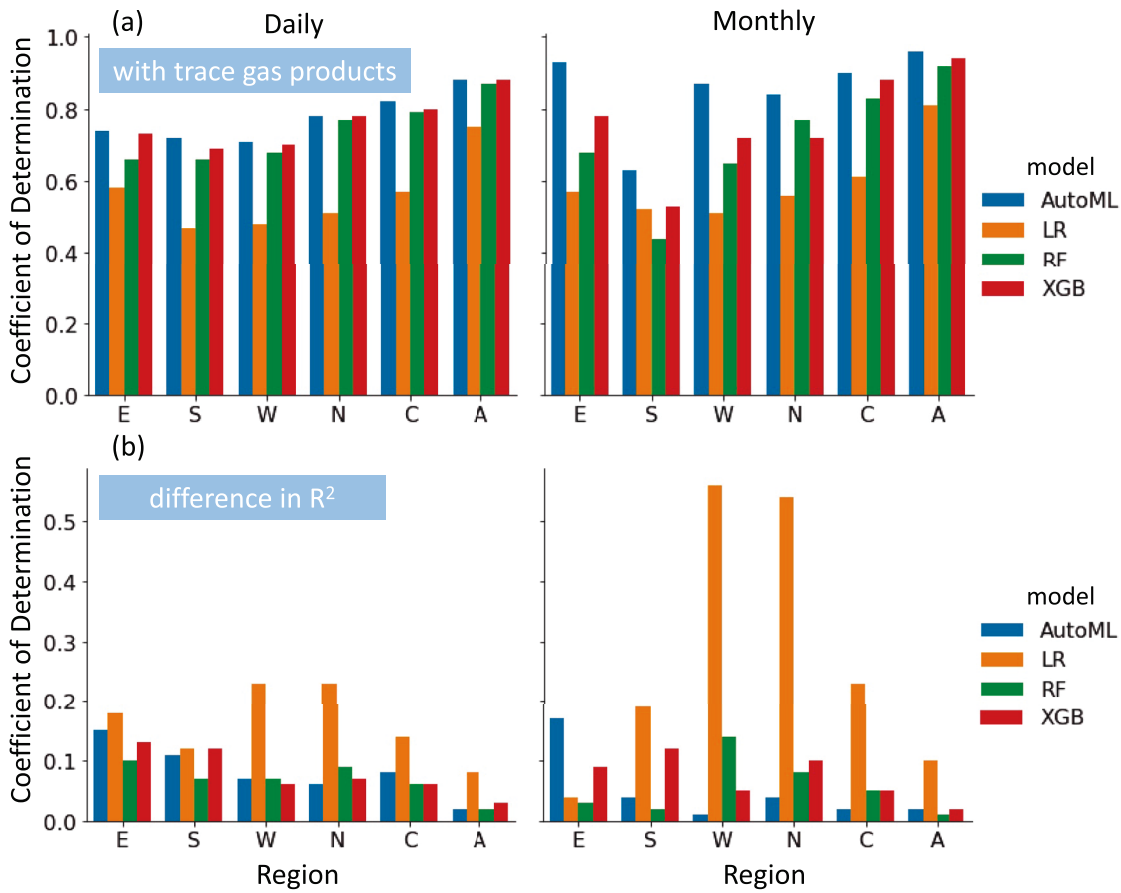

Figure 7. Performance of different estimators in estimating daily and monthly $PM_{2.5}$. (a) Estimators trained from the data with trace gas columns; (b) Increases in R^2 from (a) compared to the estimators without considering trace gas columns as features. Estimators are trained from Automated Machine Learning (AutoML), linear regression (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGB).

activities, the underlying physical and chemical processes embedded in different places are distinct. For example, the models are unaware of prior knowledge such as the number of vehicles and power plants, and the complex regional meteorological processes, which would suggest using a model tailored to each region.

Here we assess the applicability of the "best estimators" trained from larger areas (Region A and Region C) by applying them back to the sub-regions (E, S, W, N). The results (Figure 8a), however, suggest that the generalized model (which includes every grid cell in the sub-regions) is not universally the best solution. For example, incorporating data from other regions (Region C) can improve the monthly $PM_{2.5}$ estimates in Region S, but the estimates suffer as a result of gaining more irrelevant data (Region A), implying that fundamentally different governing processes control $PM_{2.5}$ variability in that region. Such noise has the potential to mislead machine learning algorithms. On the contrary, the monthly $PM_{2.5}$ estimates in Region N benefit from more information. As a result, models that perform well at larger geographic scales may ensure generally good performance overall, but fail to capture the specificity in pollutant sources and meteorological processes for each of their subregions, because tailored models may be necessary if pollutant sources and meteorological processes vary from region to region. Even if we "mistakenly" apply the model across spatial scales, we find that the presence of trace gas columns benefits models (Figure 8b). In our cases, including trace gas columns as features does not impair predictive capability.

4. Conclusions

We use an Automated Machine Learning (AutoML) approach on a modeling testbed to evaluate the information content of tropospheric trace gas columns for fine particle estimates in India. We quantify the relative information

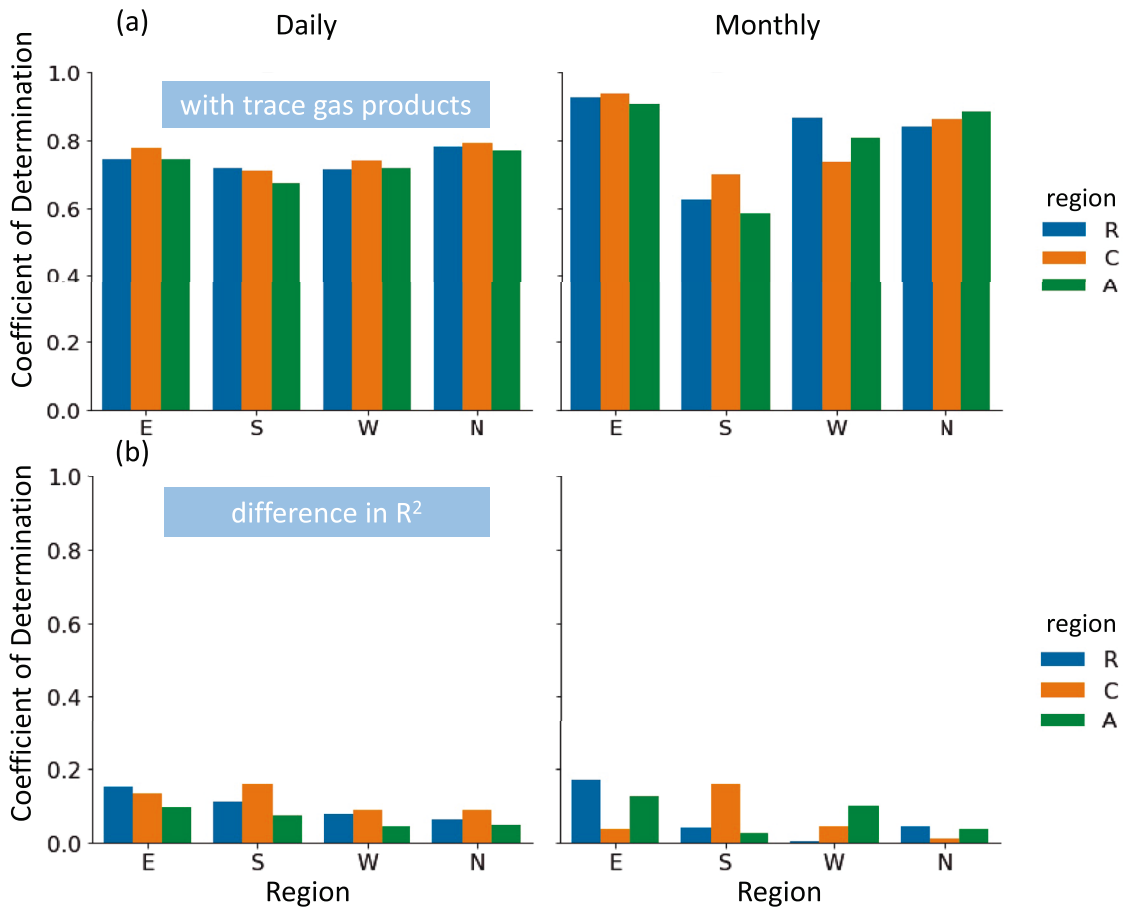

Figure 8. Performance of estimators in estimating daily and monthly $PM_{2.5}$ across spatial scales. (a) Estimators trained from the data with trace gas columns; (b) Increases in R^2 from (a) compared to the estimators without considering trace gas columns as features. Estimators are trained from the region (R), the group of regions (C), and all land grid cells (A) and applied back to each region. The differences between the predictions with and without trace gas columns as features are significant, based on Wilcoxon signed-rank test.

content of trace gas columns, AODs, meteorological fields, and emissions for four sub-regions within India, and on daily versus monthly time scales. As a byproduct, an unsupervised-learning-based regionalization strategy is developed to delineate geographical regions with similar daily patterns of variability for analysis.

Our results suggest that incorporating trace gas modeled columns enhances $PM_{2.5}$ estimates in general, regardless of model assumptions. The enhancements in predictive capability differ in both space and time. Using the ranking scores and Spearman's rank correlation, we can infer the possible particle composition, and thus a broad characterization of $PM_{2.5}$ sources. For example, we infer that $PM_{2.5}$ variability in Region E (see Figure 6) is modulated by secondary PM (sulfate and nitrate). However, in Region W and Region N, primary pollutants—as indicated by a strong correlation with CO—may modulate monthly $PM_{2.5}$ variability, whereas meteorological processes influence the daily $PM_{2.5}$ variability.

Our comparison of AutoML-derived models against selected baseline ML models demonstrates that AutoML is at least as good as user-chosen models. We ask the question "Is Big Data always better?" and find a nuanced answer that is regionally dependent. Even in these "spatially mismatched" models, however, we show that incorporating trace gas products can still improve $PM_{2.5}$ estimates across India.

Additional analysis using AOD data from the local 1:30 p.m. overpass and the mean of the 10:30 a.m. and 1:30 p.m. overpasses demonstrates that our results are consistent regardless of the local overpass time of AOD employed. Figures S4–S9 and Table S1 in Supporting Information S1 illustrate for all of these combinations that estimator performance improves when trace gas columns are utilized as features and AutoML is implemented,

for all three AOD overpass temporal sampling approaches. When a month (October) is shifted from training to testing data, our conclusions remain unchanged (Figures S10–S11 in Supporting Information S1). The modeling testbed, like any other modeling-based study, is necessarily hampered by simulation accuracy. Considering the biases in chemical transport models, there may be a discrepancy when applying the models developed from our pseudo-data sets (model simulations) to actual satellite retrievals. The variable importance identified in this study is limited by the physical and chemical representations and their uncertainties in the chemical transport model. Real satellite observations and ground measurements are needed to evaluate and improve such "simulation data-driven" models, as well as to compare our $PM_{2.5}$ estimations with those of other $PM_{2.5}$ products. However, the idealized pseudo-observations used in this work demonstrate potential for satellite retrievals of tropospheric trace gases to improve fine particle estimates in India and may contain information to interpret their origin. Our analysis also highlights the promising application of AutoML in atmospheric and environmental research. Future $PM_{2.5}$ estimates, for example, may benefit from the trace gas columns acquired by high-resolution geostationary satellites.

Data Availability Statement

Scripts and data to reproduce the results and figures are preserved at https://doi.org/10.5281/zenodo.6363824 (Zheng, 2022) or https://github.com/zzheng93/code_DSI_India_AutoML. The raw data from GEOS-Chem simulations used for Automated Machine Learning and analysis in this study are available at https://doi.org/10.7916/nwx1-jt94 (Zheng et al., 2022).

Acknowledgments
We acknowledge ExxonMobil Technology and Engineering Company and Columbia Data Science Institute Seed Funds for supporting this work. We are grateful for helpful discussions with Dr. Ruth S. DeFries and Dr. Marianthi-Anna Kioumourtzoglou. We would like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1755088. ZZ acknowledges support from NCAR Advanced Study Program Postdoctoral Fellowship. SD acknowledges support from IIT Delhi for Chair Professor Fellowship. We appreciate the careful reading of our manuscript and the many insightful comments and suggestions from three anonymous reviewers.

References

Adams, M. D., Massey, F., Chastko, K., & Cupini, C. (2020). Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction. *Atmospheric Environment*, *230*, 117479. https://doi.org/10.1016/j.atmosenv.2020.117479

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., & Hickey, J. (2019). Machine learning for precipitation nowcasting from radar images. *ArXiv:1912.12132 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1912.12132

Bali, K., Dey, S., & Ganguly, D. (2021). Diurnal patterns in ambient PM2.5 exposure over India using MERRA-2 reanalysis data. *Atmospheric Environment*, *248*, 118180. https://doi.org/10.1016/j.atmosenv.2020.118180

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V., Dentener, F., et al. (2016). Ambient air pollution exposure estimation for the global burden of disease 2013. *Environmental Science & Technology*, *50*(1), 79–88. https://doi.org/10.1021/acs.est.5b03709

Brauer, M., Guttikunda, S. K., Dey, S., Tripathi, S. N., Weagle, C., & Martin, R. V. (2019). Examination of monitoring approaches for ambient air pollution: A case study for India. *Atmospheric Environment*, *216*, 116940. https://doi.org/10.1016/j.atmosenv.2019.116940

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chaliyakunnel, S., Millet, D. B., & Chen, X. (2019). Constraining emissions of volatile organic compounds over the Indian subcontinent using space-based formaldehyde measurements. *Journal of Geophysical Research: Atmospheres*, *124*(19), 10525–10545. https://doi.org/10.1029/2019JD031262

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM Press. https://doi.org/10.1145/2939672.2939785

Chowdhury, S., Dey, S., Guttikunda, S., Pillarisetti, A., Smith, K. R., & Di Girolamo, L. (2019). Indian annual ambient air quality standard is achievable by completely mitigating emissions from household sources. *Proceedings of the National Academy of Sciences of the Unites States of America*, *116*(22), 10711–10716. https://doi.org/10.1073/pnas.1900888116

Curci, G., Hogrefe, C., Bianconi, R., Im, U., Balzarini, A., Baró, R., et al. (2015). Uncertainties of simulated aerosol optical properties induced by assumptions on aerosol physical and chemical properties: An AQMEII-2 perspective. *Atmospheric Environment*, *115*, 541–552. https://doi.org/10.1016/j.atmosenv.2014.09.009

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., et al. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, *130*, 104909. https://doi.org/10.1016/j.envint.2019.104909

Fiore, A. M., Jacob, D. J., Mathur, R., & Martin, R. V. (2003). Application of empirical orthogonal functions to evaluate ozone simulations with regional and global models. *Journal of Geophysical Research*, *108*(D14), 4431. https://doi.org/10.1029/2002JD003151

Fiore, A. M., Milly, G. P., Hancock, S. E., Quiñones, L., Bowden, J. H., Helstrom, E., et al. (2022). Characterizing changes in eastern U.S. Pollution events in a warming world. *Journal of Geophysical Research: Atmospheres*, *127*(9), e2021JD035985. https://doi.org/10.1029/2021JD035985

Geng, G., Meng, X., He, K., & Liu, Y. (2020). Random forest models for PM2.5 speciation concentrations using {MISR} fractional AODs. *Environmental Research Letters*, *15*(3), 034056. https://doi.org/10.1088/1748-9326/ab76df

Gentine, P., Eyring, V., & Beucler, T. (2021). Deep learning for the parametrization of subgrid processes in climate models. In G. Camps-Valls, D. Tuia, X. X. Zhu, & M. Reichstein (Eds.), *Deep learning for the Earth sciences* (1st ed., pp. 307–314). Wiley. https://doi.org/10.1002/9781119646181.ch21

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gosiewska, A., & Biecek, P. (2019). Do not trust additive explanations. *arXiv preprint arXiv:1903.11420*.

Greenstone, M., Nilekani, J., Pande, R., Ryan, N., Sudarshan, A., & Sugathan, A. (2015). Lower pollution, longer lives: Life expectancy gains if India reduced particulate matter pollution. *Economic and Political Weekly*, 40–46.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hoff, R. M., & Christopher, S. A. (2009). Remote sensing of particulate pollution from space: Have we reached the promised land? *Journal of the Air & Waste Management Association*, *59*(6), 645–675. https://doi.org/10.3155/1047-3289.59.6.645

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. *Environmental Science & Technology*, *51*(12), 6936–6944. https://doi.org/10.1021/acs.est.7b01210

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.) (2019). *Automated machine learning: Methods, systems, challenges*. Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Karambelas, A., Fiore, A. M., Westervelt, D. M., McNeill, V. F., Randles, C. A., Venkataraman, C., et al. (2022). Investigating drivers of particulate matter pollution over India and the implications for radiative forcing with GEOS-Chem-TOMAS15. *Journal of Geophysical Research: Atmospheres*, *127*(24), e2021JD036195. https://doi.org/10.1029/2021JD036195

Karambelas, A., Holloway, T., Kinney, P. L., Fiore, A. M., DeFries, R., Kiesewetter, G., & Heyes, C. (2018). Urban versus rural health impacts attributable to $PM_{2.5}$ and O_3 in northern India. *Environmental Research Letters*, *13*(6), 064010. https://doi.org/10.1088/1748-9326/aac24d

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Keller, C. A., Long, M. S., Yantosca, R. M., Da Silva, A. M., Pawson, S., & Jacob, D. J. (2014). HEMCO v1.0: A versatile, ESMF-compliant component for calculating emissions in atmospheric models. *Geoscientific Model Development*, *7*(4), 1409–1417. https://doi.org/10.5194/gmd-7-1409-2014

Khare, P., & Baruah, B. P. (2010). Elemental characterization and source identification of $PM_{2.5}$ using multivariate analysis at the suburban site of North-East India. *Atmospheric Research*, *98*(1), 148–162. https://doi.org/10.1016/j.atmosres.2010.07.001

Kumar, J., Mills, R. T., Hoffman, F. M., & Hargrove, W. W. (2011). Parallel k-means clustering for quantitative ecoregion delineation using large data sets. *Procedia Computer Science*, *4*, 1602–1611. https://doi.org/10.1016/j.procs.2011.04.173

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. https://doi.org/10.1029/2021MS002464

Lagerquist, R., McGovern, A., & Gagne, D. J., II. (2019). Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, *34*(4), 1137–1160. https://doi.org/10.1175/WAF-D-18-0183.1

Lian, T., & Chen, D. (2012). An evaluation of rotated EOF analysis and its application to tropical Pacific SST variability. *Journal of Climate*, *25*(15), 5361–5373. https://doi.org/10.1175/JCLI-D-11-00663.1

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1

Pai, S. J., Heald, C. L., Pierce, J. R., Farina, S. C., Marais, E. A., Jimenez, J. L., et al. (2020). An evaluation of global organic aerosol schemes using airborne observations. *Atmospheric Chemistry and Physics*, *20*(5), 2637–2665. https://doi.org/10.5194/acp-20-2637-2020

Pathak, B., & Bhuyan, P. K. (2022). Characteristics of atmospheric pollutants over the northeastern region of India. In *Asian atmospheric pollution* (pp. 367–392). Elsevier. https://doi.org/10.1016/B978-0-12-816693-2.00016-0

Porter, W. C., & Heald, C. L. (2019). The mechanisms and meteorological drivers of the summertime ozone–temperature relationship. *Atmospheric Chemistry and Physics*, *19*(21), 13367–13381. https://doi.org/10.5194/acp-19-13367-2019

Prashanth, P., Speth, R. L., Eastham, S. D., Sabnis, J. S., & Barrett, S. R. H. (2021). Post-combustion emissions control in aero-gas turbine engines. *Energy & Environmental Science*, *14*(2), 916–930. https://doi.org/10.1039/D0EE02362K

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

Ravishankara, A. R., David, L. M., Pierce, J. R., & Venkataraman, C. (2020). Outdoor air pollution in India is not only an urban problem. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(46), 28640–28644. https://doi.org/10.1073/pnas.2007236117

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12), 2570. https://doi.org/10.3390/app8122570

Shi, W., & Zeng, W. (2014). Application of k-means clustering to environmental risk zoning of the chemical industrial area. *Frontiers of Environmental Science & Engineering*, *8*(1), 117–127. https://doi.org/10.1007/s11783-013-0581-5

Stohl, A., Aamaas, B., Amann, M., Baker, L. H., Bellouin, N., Berntsen, T. K., et al. (2015). Evaluating the climate and air quality impacts of short-lived pollutants. *Atmospheric Chemistry and Physics*, *15*(18), 10529–10566. https://doi.org/10.5194/acp-15-10529-2015

Tai, A. P. K., Mickley, L. J., & Jacob, D. J. (2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment*, *44*(32), 3976–3984. https://doi.org/10.1016/j.atmosenv.2010.06.060

The International GEOS-Chem User. (2018). *Geoschem/Geos-Chem: Geos-Chem 12.0.2 release*. Zenodo. https://doi.org/10.5281/ZENODO.1455215

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002

van Donkelaar, A., Martin, R. V., & Park, R. J. (2006). Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing. *Journal of Geophysical Research*, *111*(D21), D21201. https://doi.org/10.1029/2005JD006996

Wang, C., Wu, Q., Weimer, M., & Zhu, E. (2021). FLAML: A fast and lightweight AutoML library. In *MLSys*.

Wolpert, D. H. (1996). The lack of A priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: Executive summary*. World Health Organization.

Xiao, Q., Chang, H. H., Geng, G., & Liu, Y. (2018). An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environmental Science & Technology*, *52*(22), 13260–13269. https://doi.org/10.1021/acs.est.8b02917

Xu, J.-W., Martin, R. V., van Donkelaar, A., Kim, J., Choi, M., Zhang, Q., et al. (2015). Estimating ground-level $PM_{2.5}$ in eastern China using aerosol optical depth determined from the GOCI satellite instrument. *Atmospheric Chemistry and Physics*, *15*(22), 13133–13144. https://doi.org/10.5194/acp-15-13133-2015

Zheng, Z. (2022). zzheng93/code_DSI_India_AutoML: First release (Version v0.0.0) [Software]. Zenodo. https://doi.org/10.5281/ZENODO.6363824

Zheng, Z., Ching, J., Curtis, J. H., Yao, Y., Xu, P., West, M., & Riemer, N. (2020). Unsupervised regionalization of particle-resolved aerosol mixing state indices on the global scale. *ArXiv:2012.03365 [Physics]*. Retrieved from http://arxiv.org/abs/2012.03365

Zheng, Z., Fiore, A. M., Westervelt, D. M., Milly, G. P., Goldsmith, J., Karambelas, A., et al. (2022). Automated machine learning to evaluate the information content of tropospheric trace gas columns for fine particle estimates over India: A modeling testbed [Dataset]. Columbia University. https://doi.org/10.7916/NWX1-JT94