Aerosol Science and Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uast20

Factors influencing ambient particulate matter in Delhi, India: Insights from machine learning

Kanan Patel, Sahil Bhandari, Shahzad Gani, Purushottam Kumar, Nisar Baig, Gazala Habib, Joshua Apte & Lea Hildebrandt Ruiz

To cite this article: Kanan Patel, Sahil Bhandari, Shahzad Gani, Purushottam Kumar, Nisar Baig, Gazala Habib, Joshua Apte & Lea Hildebrandt Ruiz (2023) Factors influencing ambient particulate matter in Delhi, India: Insights from machine learning, Aerosol Science and Technology, 57:6, 546-561, DOI: 10.1080/02786826.2023.2193237

To link to this article: https://doi.org/10.1080/02786826.2023.2193237

View supplementary material

Published online: 14 Apr 2023.

Submit your article to this journal

Article views: 402

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uast20

AEROSOL SCIENCE AND TECHNOLOGY
2023, VOL. 57, NO. 6, 546–561
https://doi.org/10.1080/02786826.2023.2193237

Factors influencing ambient particulate matter in Delhi, India: Insights from machine learning

Kanan Patel[a], Sahil Bhandari[b], Shahzad Gani[c,d], Purushottam Kumar[e], Nisar Baig[f], Gazala Habib[f], Joshua Apte[g,h], and Lea Hildebrandt Ruiz[a]

[a]McKetta Department of Chemical Engineering, The University of Texas at Austin, Austin, Texas, USA; [b]Lab for Environmental Assessment and Policy, University of British Columbia, Vancouver, British Columbia, Canada; [c]Institute for Atmospheric and Earth System Research/Physics, University of Helsinki, Helsinki, Finland; [d]Helsinki Institute of Sustainability Science, University of Helsinki, Helsinki, Finland; [e]Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA; [f]Department of Civil Engineering, Indian Institute of Technology, Delhi, New Delhi, India; [g]Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, California, USA; [h]School of Public Health, University of California, Berkeley, Berkeley, California, USA

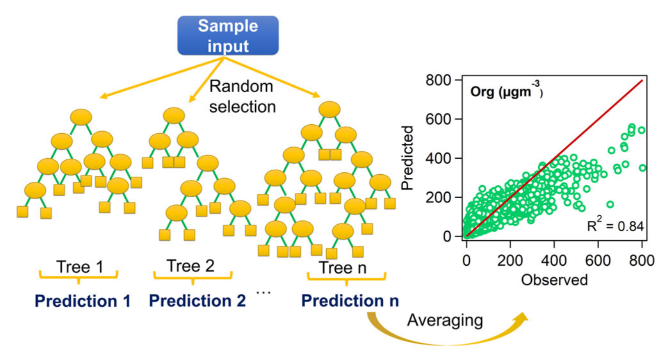
ABSTRACT

Concentrations of ambient particulate matter (PM) depend on various factors including emissions of primary pollutants, meteorology and chemical transformations. New Delhi, India is the most polluted megacity in the world and routinely experiences extreme pollution episodes. As part of the Delhi Aerosol Supersite study, we measured online continuous PM_1 (particulate matter of size less than 1 μm) concentrations and composition for over five years starting January 2017, using an Aerosol Chemical Speciation Monitor (ACSM). Here, we describe the development and application of machine learning models using random forest regression to estimate the concentrations, composition, sources and dynamics of PM in Delhi. These models estimate PM_1 species concentrations based on meteorological parameters including ambient temperature, relative humidity, planetary boundary layer height, wind speed, wind direction, precipitation, agricultural burning fire counts, solar radiation and cloud cover. We used hour of day, day of week and month of year as proxies for time-dependent emissions (e.g., emissions from traffic during rush hours). We demonstrate the applicability of these models to capture temporal variability of the PM_1 species, to understand the influence of individual factors via sensitivity analyses, and to separate impacts of the COVID-19 lockdowns and associated activity restrictions from impacts of other factors. Our models provide new insights into the factors influencing ambient PM_1 in New Delhi, India, demonstrating the power of machine learning models in atmospheric science applications.

ARTICLE HISTORY
Received 6 September 2022
Accepted 1 March 2023

EDITOR
Nicole Riemer

GRAPHICAL ABSTRACT



CONTACT Joshua S. Apte apte@berkeley.edu Department of Civil and Environmental Engineering, University of California Berkeley, 760 Davis Hall, Berkeley, CA 94720, USA; Lea Hildebrandt Ruiz lhr@che.utexas.edu McKetta Department of Chemical Engineering, The University of Texas at Austin, 200 E Dean Keeton St, Stop C0400 Austin, TX 78712-1139, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02786826.2023.2193237.

© 2023 American Association for Aerosol Research

1. Introduction

Atmospheric aerosols, or particulate matter (PM), are small solid or liquid like particles suspended in the atmosphere, which if inhaled can lead to adverse health effects and increase the risk of mortality (Schraufnagel et al. 2019). PM is known to increase the risk of various diseases including lung cancer, pulmonary infections, heart attacks, stroke, cataract, central nervous system disorders, chronic inflammatory diseases, age related disorders and cancer (Brook et al. 2010; Cohen et al. 2017; Kampa and Castanas 2008; Morakinyo et al. 2016). Air pollution (especially PM) leads to approximately 7 million human deaths every year (Campbell-Lendrum and Prüss-Ustün 2019; Health Effects Institute (HEI)) 2020; World Health Organization (WHO)) 2016). After high blood pressure, diet and tobacco smoking, ambient PM is the fourth leading factor contributing to global deaths annually (Campbell-Lendrum and Prüss-Ustün 2019). In addition to affecting human health, PM also affects climate (Rosenfeld et al. 2008; Samset et al. 2018). Aerosols have a direct radiative forcing because they scatter and absorb solar and infrared radiation in the atmosphere. They also cause an indirect radiative forcing by changing the formation and precipitation efficiency of clouds. Thus, reductions in PM, motivated by positive impacts on human health, will also have impacts on radiative forcing and therefore climate (Masson-Delmotte et al. 2022).

The US Environmental Protection Agency recognizes $PM_{2.5}$ (particles of size less than 2.5 micron) as a criteria pollutant and has set national ambient air quality standards (24-h standard set at 35 $\mu g/m^3$) to regulate its concentrations in the atmosphere (United States Environmental Protection 2022). Similarly, the World Health Organization has established guidelines for ambient air pollution levels to help policymakers across the world set standards for air quality management. The guidelines released in 2021 include several interim targets, designed to help countries with high pollution levels, with the final air quality guideline set at 15 $\mu g/m^3$ for 24-h average concentrations (World Health Organization (WHO)) 2021). Despite environmental regulations in high income countries and growing awareness in low and middle-income countries, reducing ambient PM remains a challenge. This is especially the case in low and middle-income countries, where approximately 90% of air pollution-related deaths occur (World Health Organization (WHO)) 2016). Megacities in developing nations are some of the biggest hotspots of air pollution exposure because of high population density and rapid urbanization and associated increases in industrial and road emissions (Molina 2021). Delhi, India is a rapidly growing urban center and is the second most populated city in the world, with a population of around 28 million (UN 2018). According to a recent estimate, Delhi is the world's most polluted megacity, on track to also become the world's most populated megacity by 2028 (United Nations: World urbanization prospects 2018; World Health Organization (WHO)) 2016). However, our understanding of the factors influencing air quality in Delhi is a work in progress (Baig et al. 2020; Bhandari et al. 2020; Gani et al. 2019; Gani et al. 2020; Guttikunda and Calori 2013; Guttikunda and Gurjar 2012; Jaiprakash et al. 2017; Pant et al. 2015; Pant, Guttikunda, and Peltier 2016; Pant and Harrison 2012; Patel et al. 2021a; Patel et al. 2021b).

PM can either be directly released into the atmosphere as primary emissions (referred to as primary aerosol) or formed through chemical reactions in the atmosphere and subsequent particle formation by partitioning to the aerosol phase (referred to as secondary aerosol). The eventual atmospheric fate of ambient aerosol depends on several factors including emissions, meteorology, atmospheric chemistry as well as sinks (condensation, coagulation, deposition, etc.). For example, ambient temperature affects thermodynamics and gas-particle partitioning of PM, wind speed and planetary boundary layer height affect ventilation. Further, winters in Delhi are often associated with temperature inversions (where temperature close to earth's surface is lower than the layer above), which traps the pollutants close to earth's surface, leading to high surface concentrations and haze-like conditions. Traditionally, deterministic models such as atmospheric chemical transport models have been used to understand the influence of changing source emissions and meteorology on ambient pollutant concentrations. However, they require prior knowledge of emission profiles and reaction pathways (for secondary formation) which are not yet satisfactorily established for Delhi.

Recently, some studies have shown the capability of statistical models such as predictive machine learning (ML) models to capture the temporal variability of ambient air pollutants, given sufficient data is available to train these models (Christopoulos et al. 2018; Feng et al. 2019; Grange et al. 2021; Lovrić et al. 2021; Nair et al. 2021; Pande et al. 2022; Qin et al. 2022; Rubal and Kumar 2018; Stirnberg et al. 2021; Wang et al. 2020; Yang et al. 2021; Wang et al. 2022; Yu et al. 2016; Zhang et al. 2022). The ML models offer several advantages – they have higher computational efficiency and offer the flexibility of leveraging and capturing measured data. Furthermore, recent advances in technology (especially aerosol mass spectrometry) have made it possible to measure aerosol composition and concentrations at a high mass and temporal resolution (Baltensperger et al. 2010; Pratt and Prather 2012; Zhang et al. 2007). As part of the Delhi Aerosol Supersite (DAS) study, we collected long-term PM_1 concentrations and composition in Delhi at high temporal resolution (Arub et al. 2020; Bhandari et al. 2020; Gani et al. 2019; Gani et al. 2020; Patel et al. 2021a; Patel et al. 2021b). The objective of the DAS study was to understand the factors influencing ambient PM_1 in Delhi.

For the analysis presented in this paper we utilized over three years of data collected as part of the DAS study to build machine learning models using random forest regression. The objective of this study was to investigate the influence of meteorology and emission proxies on the PM_1 variability in Delhi by using random forest regression. We used several meteorological parameters and emission proxies as predictor variables in these models (see Section 2 for details), which can capture the non-linear response of PM_1 to changing emissions, meteorology and atmospheric oxidizing capacity as discussed in Section 3. We also performed sensitivity analysis, where we varied one feature at a time by keeping the other parameters to a constant value to understand the influence of each of these parameters in the variability of PM_1 and its constituents.

2. Materials and methods

Data: Non-refractory PM_1 (NR-PM_1; particulate matter of size less than 1 micron that flash vaporizes at the vaporizer temperature of 600 °C) composition and concentration was measured at the Delhi Aerosol Supersite (DAS) by using a Quadrupole Aerosol Chemical Speciation Monitor (Q-ACSM, Aerodyne Research, Billerica, MA, USA) (Ng et al. 2011). DAS is located at the Indian Institute of Technology, Delhi campus in New Delhi. Details on the instrument set up, operation, calibration and data processing are presented in our previous publications (Bhandari et al. 2020; Gani et al. 2019; Gani et al. 2020; Patel et al. 2021a; Patel et al. 2021b). Data was collected every ~1-min and was post-averaged to 1 h for the analysis presented here. Data from Jan 2017 to Feb 2020 was used to develop the models.

Table 1. Predictor (input) variables used in the models.

S.No	Predictor variable	Shortform
1	Ambient Temperature	T
2	Relative Humidity	RH
3	Planetary Boundary Layer Height	H
4	Wind Speed	WS
5	Wind Direction	WD
6	Precipitation	Precip.
7	Solar Radiation	SR
8	Cloud Cover	CC
9	Fire Counts	Fire
10	Hour of the Day	HOD
11	Day of the Week	DOW
12	Month of the year	MOY

Table 1 summarizes the predictor (input) variables used in the models. They include meteorological parameters – ambient temperature (T, measured 10 m from the ground), relative humidity (RH), planetary boundary layer height (H), wind speed (WS), wind direction (WD), precipitation (P), solar radiation (SR), cloud cover (CC). Agricultural burning fire counts in the northwest states are used as proxy for burning emissions and were obtained from the NASA fire information for resource management system, FIRMS, which uses the moderate resolution imaging spectroradiometer, MODIS, collection 6 (Fire Information for Resource Management System (FIRMS)) 2022; Giglio, Schroeder, and Justice 2016; Justice et al. 1998) dataset. Furthermore, hour of day (HOD), day of week (DOW) and month of year (MOY) were used as categorical variables to account for emissions specific to certain times (e.g., vehicular emissions during peak traffic hours, the differences in vehicular emissions on weekday versus weekend due to different traffic conditions, and biomass burning during the winter months). Hourly T, H, WS, WD data were obtained from NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) (Bosilovich, Lucchesi, and Suarez 2016; Durre, Vose, and Wuertz 2006; Gelaro et al. 2017; Mccarty et al. 2016; MERRA-2 website 2022). We found wind direction to be a reasonable proxy for wind trajectory at this location (see Section S2 in the online supplementary information (SI) for details). Solar radiation and precipitation data were obtained from the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis 5th Generation (ERA5) (Muñoz-Sabater et al. 2021). RH and cloud cover were obtained from the Indira Gandhi International Airport (IGIA; 8 km from our site), with the data retrieved from the Iowa Environmental Mesonet (IEM) archive. Cloud cover data were obtained in the METAR code format (Automated Surface Observing System (ASOS)) 1998). We compared these cloud cover data to those obtained from MODIS aboard the Aqua and Terra satellites (using the algorithm from Christiansen, Carlton, and Henderson 2020) as part of our quality check and found that they compared relatively well (the data matched well ~80% of the times; see

Section S1 in the SI for details on cloud data comparison). We used the airport cloud data instead of the MODIS cloud data for our analysis because (1) airport data are available at a higher (∼1 h) resolution while MODIS overpasses occur once a day and (2) airport data are in "oktas" (Oktas 2022), so they include details on the extent of cloud cover while MODIS data are extracted as "clear" or "cloudy" day flags.

The factors were selected to account for the variability induced due to meteorology as well as certain emission sources (as proxies). The models do not directly account for secondary formation, although parameters such as solar radiation serve as proxies for photochemistry which leads to the formation of secondary pollutants. While deterministic models such as chemical transport modeling can account for the formation of secondary pollutants, they are parameter rich and computationally intensive. Machine learning models provide the simplicity of fewer parameters and better computational efficiency relative to such models (see Section 1). Despite fewer parameters, they can be a powerful means to understand and estimate air quality if trained with enough data (Yang et al. 2021).

Model: Random-forest regression was used to predict NR-PM_1 (hereafter referred to as PM_1) and its constituents by using the predictor variables described above. The python package, scikit-learn (Pedregosa et al. 2011) was used for the analysis. Random forest regression is a decision tree-based modeling approach where multiple decision trees are generated in parallel by using boot-strapped training data and splits in each decision tree are made along the 'best' of a subset of randomly chosen parameters (Breiman 2001). The splits are performed one at a time and they divide the predictor feature space into multiple regions. The predicted value of the estimator variable (PM_1 and its constituents in our case) in a region is its average value in that region. The 'best' split is the one that minimizes the residual sum of squares (RSS, Equation (1))

$$\mathrm{RSS} = \sum_{j=1}^{n} \sum\nolimits_{i:x_i \epsilon R_j} \left(y_i - \hat{y}_{R_j}\right)^2 \tag{1}$$

Here y_i is the true value of the estimator variable for predictor variables (x_i) belonging to region j and $\hat{y}_{R_j}$ is its average value in region j. The final prediction is the average of all predictions made by the multiple decision trees.

The advantages of random forest regression include that it generates de-correlated trees (Breiman 2001) since the subset features for splitting are chosen at random. Further, the use of multiple trees ensures that it can handle a high number of predictor variables without the risk of overfitting. This is also one of the reasons that it is often referred to as a "parameter rich" model, although technically it is non-parametric (i.e., makes no assumptions about the type of mapping function). Furthermore, unlike other parameter rich machine learning models (such as artificial neural networks and deep learning), it offers the opportunity to measure feature importance and ranking (Breiman 2001; Rosina et al. 2020). We used recursive feature elimination (RFE) and 5-fold cross validation to determine feature ranking by using R^2 scores. As the name suggests, the goal of RFE is to select features (or factors) by recursively considering smaller sets of features. The importance of each feature is determined by the variance explained by them. The least important features are "pruned" from the current set of features until the desired number of features are reached (Pedregosa et al. 2011). Further, since bootstrapped training data are used in the models, the unused/unseen data can be used to simultaneously test the models. These unseen data are called the out-of-bag (OOB) data, the scores associated with them are called the OOB scores and the predictions associated with them are called the OOB predictions. We used the OOB scores to test the models and to tune the hyper parameters, namely the depth of trees and the number of trees. The 'depth' of a node is the number of edges from that node to the tree's root node. The depth of the tree is the depth of its deepest leaf node. The number of trees correspond to the trees used for estimating the averaged prediction. Increasing the number of trees may increase the variance explained but it also increases the computational cost, and after a certain number of trees the increase in variance is negligible. Increasing the depth of trees also increases model variance but the increase is insignificant beyond a certain point. We tuned the number and depth of trees such that further increase in any of these parameters yielded negligible further increase in the OOB score (see Figures S6–11 in the SI).

The COVID-19 pandemic led to large variations in urban air quality in different parts of the world (Chauhan and Singh 2020; Gautam 2020; Kumar et al. 2020; Kumar 2020; Mahato, Pal, and Ghosh 2020; Manchanda et al. 2021; Sharma, Jain, and Lamba 2020) In Delhi, the COVID-19 lockdown restrictions in 2020 were implemented in 4 phases, starting Mar 25th and continuing until the end of May (see Figures S12 and 13 in the SI for details lockdown restrictions). We used the model predictions during the lockdown period as the expected concentrations under business-as-usual conditions (if the lockdown did not happen)

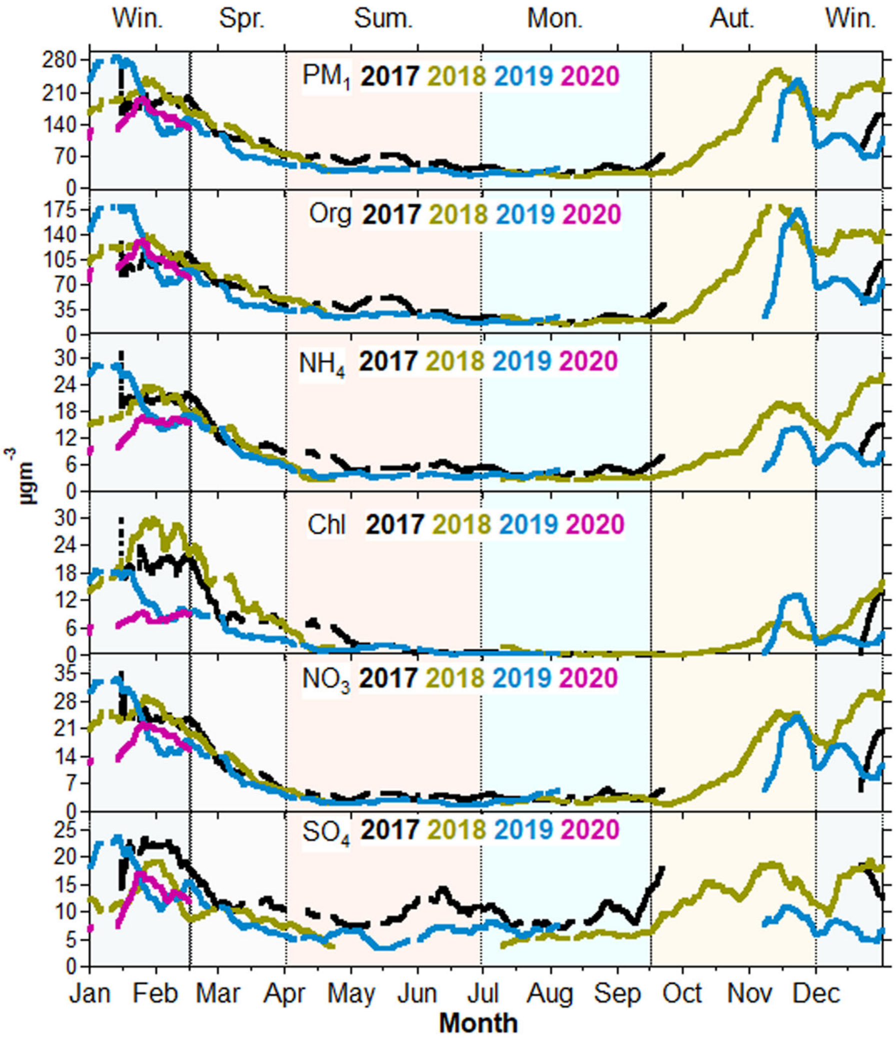


Figure 1. Bi-weekly moving averages of PM$_1$ and its constituents from Jan 2017 to Feb 2020. The seasons are categorized as winter ("Win.", December to mid-February), spring ("Spr.", mid-February to March), summer ("Sum.", April to June), monsoon ("Mon", July to mid-September) and autumn ("Aut.", mid-September to November).

because they are trained to estimate concentrations under typical conditions. Next, we used the "% difference" of observed concentrations relative to predicted concentrations (during the lockdown period) to quantify the influence of the lockdown restrictions. Finally, we compared this to our previous methodology (Patel et al. 2021b) of using "% difference" of observed concentrations relative to historical concentrations (during the same period as the lockdown) to quantify the influence of the lockdowns (see Section 3.4).

3. Results and discussion

3.1. Temporal trends

As shown in Figure 1, the temporal trends of the concentrations of PM$_1$ and its constituents have been consistent over the last few years – with higher concentrations during autumn, winter and lower concentrations during spring, summer and monsoon. Chloride (Chl) and sulfate (SO_4) concentrations are more variable than others, especially in Jan/Feb (winter), which could partly be due to a more diverse source mix, such as industrial emissions, trash burning, etc., at that time of the year. Figure S14 in the SI shows the trends of PM$_{2.5}$ concentrations recorded at the nearest monitoring station that is operated by the Delhi Pollution Control Committee (DPCC), R.K. Puram (3 km away). The trends in PM$_{2.5}$ are comparable to those observed in PM$_1$ recorded at our site – there is an increase in the concentrations during the colder months (autumn and winter), and a decrease in concentrations during the warmer months. Further, there is no consistent decrease in concentrations over the years, similar to the PM$_1$ trends. Figure S15 shows the trends of select meteorological variables including temperature, planetary boundary layer height and relative humidity over the last few years. Their trends have also been

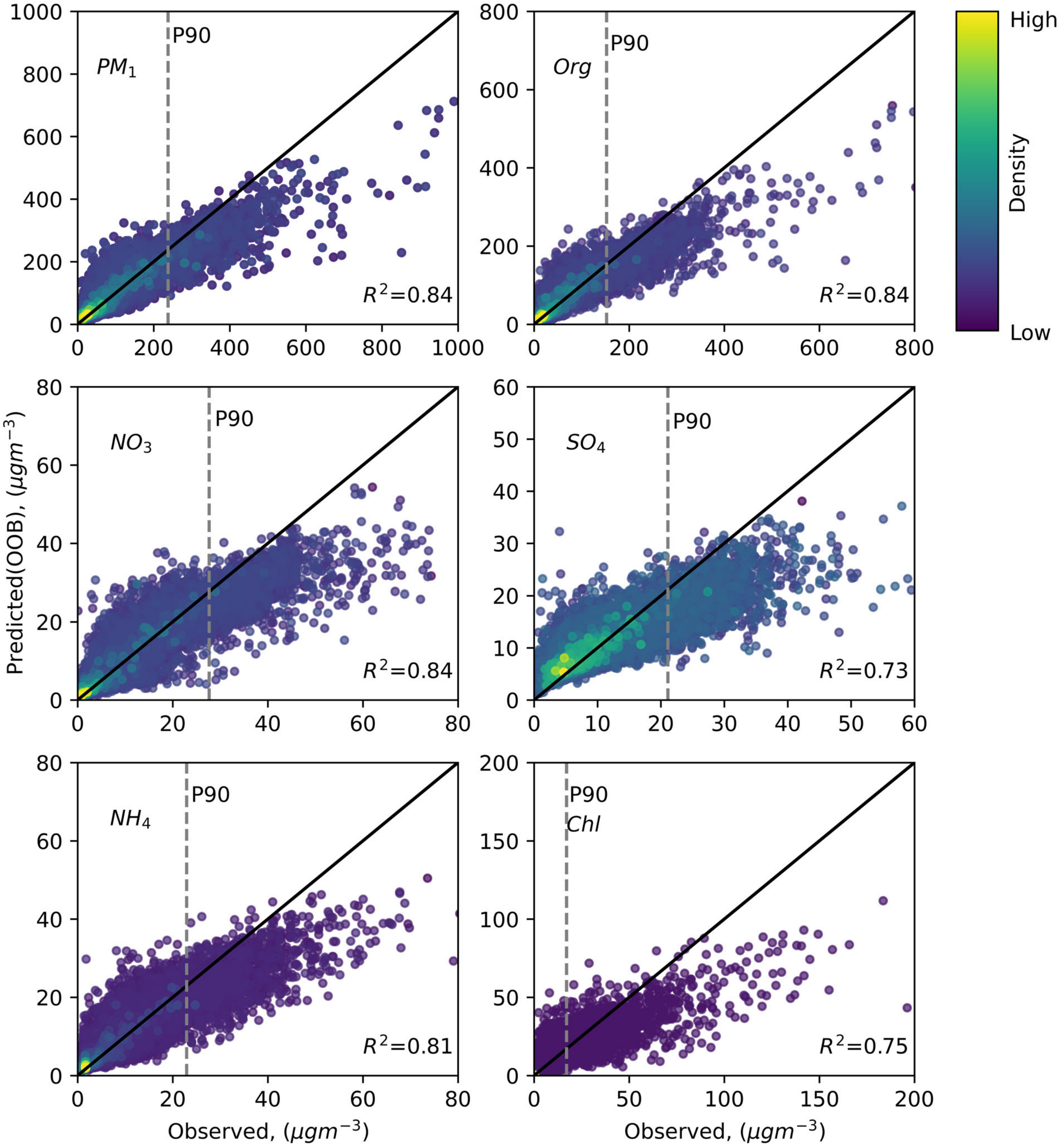


Figure 2. Predicted (OOB) vs observed concentrations of PM_1 and its constituents, colored based on density of points based on binned data. P90 line shows the 90% percentile points based on the observed data.

consistent across these years. Meteorology affects PM_1 concentrations and composition; for example, temperature influences PM_1 concentrations through gas-particle partitioning and thermodynamics, relative humidity affects concentrations by governing available absorbing mass and aqueous phase chemistry, and planetary boundary layer height affects PM_1 concentrations through vertical mixing and dilution. Thus, by using machine learning modeling, in Section 3.2, we investigate whether meteorology and select emission proxies can explain the variability of PM_1 and its constituents at the Delhi Aerosol Supersite.

3.2. Machine learning modeling

The model performance for each of the variables is shown in Figure 2 (scatter plot) and Figure S16 (Root Mean Squared Error, RMSE, in the SI). Overall, the random forest models are able to capture over 70% of the variance in all species ($R^2 > 0.7$) indicating that

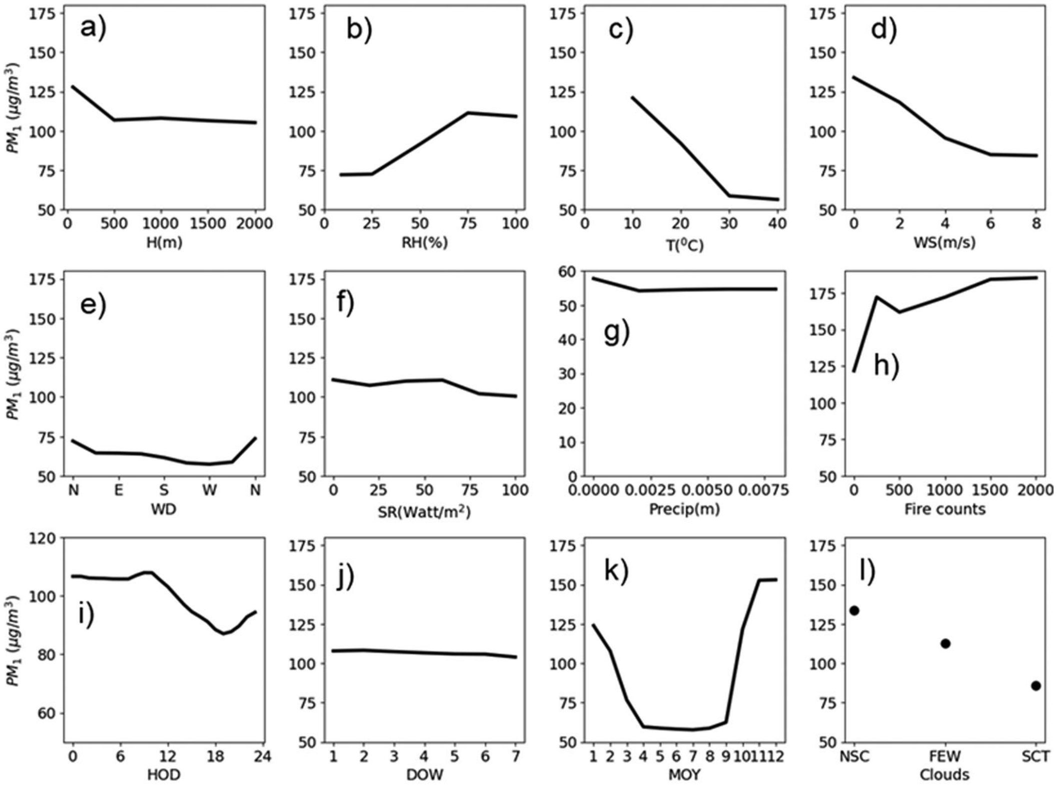


Figure 3. Sensitivity analysis for PM_1. MOY = 2, HOD = 10 and DOW =1 for all the cases except h) Fire Counts where MOY was set to 10 (Oct), when significant fires are experienced. Clouds in l) are labeled as NSC: No Significant clouds, FEW: FEW clouds (< =25% Coverage) and SCT Scattered clouds (< =50% Coverage). See details on cloud coverage in section S1 of the SI.

meteorology and emission proxies are sufficient to capture most of their variability. The explained variance is highest for PM_1 (0.84), organics (Org; 0.84) and nitrate (NO_3; 0.84), followed by ammonium (NH_4; 0.81), chloride (Chl; 0.75) and sulfate (SO_4; 0.73), suggesting that the concentrations of chloride and sulfate are influenced by a greater number of factors and are thus more variable than the other species (Figure 3), consistent with their temporal trends in Figure 1.

The concentrations of all the species are underestimated at higher concentrations (at values usually greater than P90; Figure 2). Previous studies that have reported an underestimation at higher concentrations have noted that random forest-based models tend to have higher biases in predicting larger values due to fewer number of data samples available at higher concentrations, which is also seen through less dense data points at higher concentrations in Figure 2 (Arlot and Lerasle 2014; Bengio and Grandvalet 2004; Yang et al. 2021). Underprediction at higher concentrations could also be due to emissions during the episodic events not being fully captured by the models. For example, festivals such as Diwali and Lohri which are usually associated with high air pollution are not accounted for by the model. Further, emissions from the unregulated brick kiln industry can be sporadic and thus cannot be fully accounted for by the temporal proxies (Misra et al. 2020). Although the RMSE values are higher for autumn and winter (Figure S16 in the SI), the normalized RMSE values (Figure S17) are within 0.2–0.4 for most species (other than Chl) across all seasons. The values are higher during the summer for NH_4, SO_4 and NO_3, indicating that the unexplained variance is somewhat higher relative to mean concentrations in summer for these species. This points to the role of factors such as increased photochemistry and other reaction pathways (e.g., the formation of organonitrates and organosulfates) which are not directly accounted for by the models. The high NRMSE values of Chl may be due to low mean concentrations and highly variable data, consistent with the density scatter plot (Figure 2). Nevertheless, these models are able to capture the majority of the temporal variability of PM_1 and its species with the chosen features. In the next section, we use the models to perform a sensitivity analysis.

3.3. Influence of different parameters – sensitivity analysis

A sensitivity analysis was performed to understand the influence of different parameters on the variability of PM_1 and its constituents. This was achieved by varying one parameter at a time while keeping the other parameters constant, usually set as their overall average, unless otherwise specified.

Figure 3 shows the sensitivity analysis for PM_1, Figure S18 in the SI shows the results obtained from recursive feature elimination (described in Section 2) and Table S2 shows the feature ranking and importance for the different models.

The top five features for PM_1 (Table S2) are T, H, fire counts, WD and MOY. These features can explain around 80% of the variance in PM_1 (Figure S18). Recent studies in Los Angeles, U.S. (Yang et al. 2021) and Beijing, China (Su et al. 2020) also found H, RH, T and WD to be amongst the top factors for random forest regression models predicting $PM_{2.5}$, indicating that meteorology is consistently important in governing PM variability across different cities around the world. Figure 3 is generally consistent with feature ranking shown in Table S2 (which includes features listed in the order of their importance in the model; details shown in Section 2). For example, there is noticeable variability with T, H, MOY, and Fires, which are amongst its top ranked features. However, it does not encompass the total variability; for example, Figure 3e shows the variability with WD but the pattern is valid for MOY = 2, HOD = 10 and DOW =1 only. Thus, it is essentially a 2-dimensional visualization of a complex multi-dimensional variance problem. Nevertheless, it provides interesting insights – the decrease in PM_1 with wind speed (Figure 3d) indicates wind speed plays an important role in providing ventilation. While we had previously demonstrated some effect of ventilation provided by wind (Bhandari et al. 2020), this analysis presented here allows us to separate out the effect of other parameters. Solar radiation, which is expected to promote formation of secondary PM, was not found to be very important for PM_1 variability (Figure 3f and Table S2). This could in part be due to covariance with other features such as temperature. For example, higher radiation may be associated with higher temperatures which can lead to some evaporation of PM. In some other studies, ozone was included as a predictor variable as a proxy for secondary PM formation in the model. For example, Yang et al. 2021 found that including ozone increased model R^2 from 0.53 to 0.65. We did not include ozone in our models because it was not measured at our site.

Although there is some variability with HOD (Figure 3i), HOD ranks low (11th) in feature ranking, indicating that after accounting for the influence of meteorological parameters varying through the day, PM_1 does not vary significantly with HOD. This could be because PM_1 includes primary (PA) and secondary aerosol (SA) where one of these species complements the other in the diurnal cycle – e.g., primary emissions usually occur during the morning and evening hours (traffic, biomass burning, etc.), while secondary formation is during the daytime, when gas phase precursors are oxidized in the presence of sunlight to form low volatility products which partition to the particle phase. If primary emissions were more important than secondary formation, we would see a larger peak during traffic hours and a higher feature importance for HOD. This suggests that the fraction of secondary aerosol is comparable to primary aerosol in Delhi, which is consistent with our previous analyses: we have shown that even though both PA and SA reduced during the COVID-19 lockdowns in 2020, the ratio of PA/SA did not decrease, and SA continued to dominate (Patel et al. 2021b). We also found that while primary aerosol dominated during high pollution episodes, on average, secondary aerosol was the dominant contributor to PM_1 in Delhi (Bhandari et al. 2020; Patel et al. 2021a). Thus, our current analysis provides additional evidence for the importance of secondary as well as primary aerosol in Delhi.

DOW does not impact PM_1 variability (Figure 3j), indicating that weekdays versus weekends do not significantly impact PM_1 concentrations. The variability with MOY (Figure 3k and rank 5 in Table S2 in the SI) suggests that even if all the other factors including meteorology were constant across the year, there would still be increased concentrations during the winter months (e.g., December), pointing to increased emissions such as domestic biomass burning during those months. Reducing these season specific emissions will help reduce the PM_1 concentrations by $\sim$ 60% ($\sim$150 $\mu g/m^3$ in winter months versus $\sim$60 $\mu g/m^3$ in summer months, when normalized for meteorology). This analysis demonstrates that while meteorology plays an important role in contributing to high concentrations in Delhi during winter, increased sources also contribute significantly to pollution events.

PM_1 and cloud cover (Figure 3l) are anti-correlated to each other. This could potentially be due to aerosol-cloud interactions where studies have shown that biomass burning aerosol might have a negative influence on the cloud cover (Feingold, Jiang, and Harrington 2005; NASA SC11 2011). This has been hypothesized to be due to the semi-direct aerosol effect on clouds where certain aerosols absorb sunlight and warm the atmosphere relative to ground/surface temperature. This "heating" reduces the cloud fraction by accelerating the process of evaporation of existing clouds and by reducing the upward movement of moisture needed to form clouds (Hansen, Sato, and Ruedy 1997). Such suppression of clouds from biomass burning aerosol has also been observed in the biomass burning regions of Brazil (Koren et al. 2004). Consistently high combustion and biomass burning aerosol in Delhi (Bhandari et al. 2022; Patel et al. 2021a; Patel et al. 2021b) could partly explain this observation. Another explanation for this observation could be the covariance of cloud cover and temperature because we observed that certain cloudy periods were associated with warmer temperatures (Figure S19 in the SI). This is consistent with the radiation effects of clouds, if larger cloud cover is observed during nighttime (International Satellite Cloud Climatology Project (ISCCP) 2010; UIUC: The Weather World Project 2010). Indeed, we observed that nighttime temperatures during SCT periods (corresponding to relatively high cloud fraction) were higher in many seasons (Figure S20).

To our knowledge, this is the first analysis to study the anti-correlation between cloud cover and PM_1 concentrations recorded through ground-based measurements in a polluted city in South Asia. Our results are different from previous studies performed in regions with much lower concentrations. For example, Christopher and Gupta 2010 found that $PM_{2.5}$ concentrations in the continental U.S. were not correlated to cloud cover (i.e., they found no significant differences in $PM_{2.5}$ concentrations on clear versus cloudy days). More recently, Christiansen, Carlton, and Henderson 2020 found that while the $PM_{2.5}$ concentrations in the continental U.S. during clear sky conditions were statistically higher than cloudy sky conditions, the trends of $PM_{2.5}$ constituents were different – nitrate concentrations were consistently higher during cloudy periods and sulfate concentrations were higher during cloudy periods in winter months. The differences in the observations between U.S. and a polluted region in South Asia are likely driven by differences in PM composition, loadings, and the underlying meteorology. We recognize that the anti-correlation of fraction of cloud cover and PM_1 may be due to reasons other than aerosol-cloud interactions. In this study, our analysis was limited by PM_1 data collected at a single site and cloud cover data collected at the Indira Gandhi International Airport (IGIA; 8 km from our site). We encourage future studies to investigate this in detail to further understand the impact of cloud cover on air pollution in the region.

The sensitivity analyses for the constituents of PM_1 are presented in Figures S21–25 in the SI. The top five features of organics (T, H, WD, Fires, MOY) explain around 80% of its variance (Figure S18 and Table S2). Similar to PM_1, Org does not vary significantly with HOD (Figure S21), due to the reasons mentioned above. We wanted to understand how the variability of Org with T compared with other studies that have looked at this effect, which is assumed to be mainly due to the semi-volatile nature of Org. In Figure S26, we normalized the mean predicted organic aerosol (OA) concentrations by the highest mean concentration (obtained for lowest temperature, $\sim 65\ \mu g/m^3$ Figure S21) and used the curve as a proxy for mass fraction remaining of OA as a function of temperature (otherwise obtained from thermodenuder measurements) (An et al. 2007; Faulhaber et al. 2009; Louvaris et al. 2017). We compared this with the results of Grieshop et al. 2009, who developed these curves for fresh and aged biomass burning plumes using thermodenuder measurements. The OA loadings used in their study ($40–90\ \mu g/m^3$) were comparable to those measured here. The resulting curve lies in between the fresh and aged OA curves, consistent with the influence of primary (proxy for fresh) and secondary (proxy for aged) OA on OA in Delhi. This analysis shows a proof-of-concept method which may be used to gather insights about OA volatility where thermodenuder measurements are not available.

T, WD, Fires are amongst the top features for chloride (Figure S22 and Table S2 in the SI). The importance of T and the drop to near zero concentrations at higher temperatures is consistent with the volatile nature of ammonium chloride (Salcedo et al. 2006). The importance of the NW wind direction (Table S2 and Figure S22) is consistent with our previous hypothesis, based on a smaller dataset, of the influence of industrial emissions (e.g., hydrochloric acid released from steel pickling in the NW region) on chloride in Delhi (Gani et al. 2019). The importance of Fires is consistent with our previous hypothesis of the influence of regional agricultural fires on chloride in Delhi (Bhandari et al. 2020; Gani et al. 2019; Gani et al. 2020; Patel et al. 2021a; Patel et al. 2021b). The morning peak in the variability with HOD (Figure S22) suggests the influence of emissions released during that period (6–8 AM). One potential source could be trash burning because it usually occurs during the morning (colder) hours, especially during the colder months (Bhandari et al. 2020).

Similar to chloride, nitrate is also most influenced by temperature, consistent with the volatile nature of ammonium nitrate (Table S2 and Figure S24). Further, MOY is ranked third in terms of feature importance for nitrate, highest amongst all species (Table S2). This suggests increased sources and processes during certain months of the year, such as biomass burning, contributing to high nitrate concentrations. Another reason could be that factors such as WD are more important for chloride (because of industrial emissions from NW as mentioned above) and not as important for nitrate, thus pushing WD below MOY for nitrate.

The top four features for sulfate are T, WD, RH and WS respectively (Table S2 and Figure S23 in the SI). The importance of wind is likely because of the influence of power plants located outside the city (Jain and Sharma 2020). This is consistent with the longer atmospheric lifetime and long-range transport of sulfate relative to the other species (Kallos et al. 2007; Seinfeld and Pandis 2006). The importance of RH is consistent with aqueous phase chemistry influencing ammonium sulfate formation at higher RH (Bhandari et al. 2020; Jaiprakash et al. 2017; Wang et al. 2016). Like sulfate, ammonium also has RH in the top 5 features (Table S2 and Figure S25). The importance of temperature for sulfate is surprising (Table S2), because sulfuric acid (which is neutralized by ammonia to form particulate ammonium sulfate), has low volatility and remains in aerosol/particle phase even at higher temperatures (Patel et al. 2020; Seinfeld and Pandis 2006). Some explanations for the importance of T for sulfate could be (a) correlation between temperature and solar radiation resulting in higher oxidizing capacity at higher temperature, which would lead to increased oxidation of sulfur dioxide (SO_2) to sulfuric acid or (b) the increased demand (and therefore generation) for electricity at higher temperatures which would lead to higher SO_2 emissions. The latter hypothesis is supported by our observations during the COVID-19 lockdowns in 2020 where we noted reduced sulfate concentrations, which were correlated to reduced electricity generation during the period (see Section 3.4. and Andrew 2020; Patel et al. 2021b).

The sensitivity analysis with RF predictions provides interesting insights into the influence of predictor variables on PM_1 concentrations. However, because the models are trained on real world data where the input variables are codependent, this approach cannot fully capture the impact of individual predictor variables on the predictions. We encourage future studies to use newly emerging techniques such as “explainable machine learning” (Hou et al. 2022) which may be able to better account for codependency among variables and provide more robust insights into the impact of individual variables on the predictions for such data sets.

3.4. Inferring the influence of the COVID-19 lockdown from the model

The COVID-19 lockdowns in India were implemented in four phases – Phase 1 (25 March to 14 April), 2 (15 April to 3 May), 3 (4–17 May), and 4 (18–31 May), with restrictions easing with time (summary of restrictions shown in Figures S12 and S13 in the SI) (Patel et al. 2021b). Figure 4 compares the influence of COVID-19 lockdown on the air quality in Delhi quantified using two approaches – (1) using the model predictions with the input features from the lockdown period as expected concentrations (under business-as-usual conditions if the lockdown did not happen) and computing the “% difference” between observed and expected concentrations as the influence of the lockdown restrictions and (2) using the average historical (2017–2019) concentrations as proxy for expected concentrations (under business-as-usual conditions without the lockdown) and computing the “% difference” between observed and historical concentrations (Patel et al. 2021b). While the models may not be able to completely capture the trends as expected under “business-as-usual” conditions, the model predictions would still be a better estimate than using historical trends as proxy for “business-as-usual conditions.” To test our hypothesis, we trained the model using the data from Jan 2017 – Dec 2019. Figure S27 compares the diurnal trends for (1) observed concentrations in Jan 2020, (2) model predictions in Jan 2020 and, (3) historical averages for Jan 2017, 2018, and 2019. As shown in Figure S27, the model predictions capture the trends better than historical averages.

As shown in Figure 4, there are general agreements between the two approaches, such as greater differences during Phase 1 (P1), especially for PM_1, Org, and nitrate. During Phase 1, activities such as transportation and construction were severely restricted, and several factories and businesses were shut down, which explain reduction in these species. The reduction in nitrate was lower than the reduction in NO_x (which observed a “% difference” less than −100%; see comparison with NO_x in Patel et al. 2021b), suggesting complex nitrate chemistry in Delhi. “%

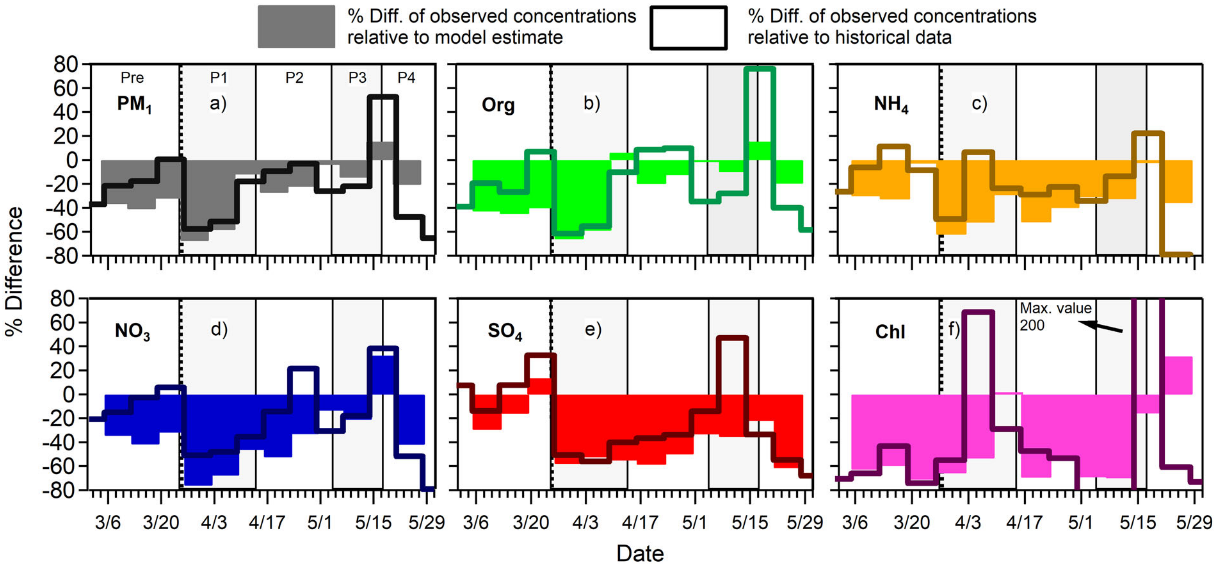


Figure 4. Comparison of percentage difference of observed concentrations relative to predicted (model) concentrations versus observed concentrations relative to historical concentrations (2017–2019) for (1) PM_1, (b) Org, (c) NH_4, (d) NO_3, (e) SO_4, and (f) Chl. Data shown for weekly median concentrations. P1–P4 are the four phases of the lockdown. Lockdown restrictions were the strictest during P1 and P2.

difference" relative to the model was lower even before the lockdown for most species (Figure 4), suggesting lower activity (and therefore emissions) than expected before the lockdown. This is consistent with the mobility trends observed during the period, which showed reduced mobility even before the lockdown (Patel et al. 2021b). The trends in sulfate were similar during P1 and P2 for the two approaches suggesting reduced emissions during the period. These trends are also consistent with the reduced electricity generation during this period (Patel et al. 2021b).

While there are similarities between the trends generated using the two methods, there are also notable differences. For example, higher concentrations relative to historical concentrations were observed during Phase 3 (P3) for most species (Figure 4). However, these concentrations were explained by the model. Thus, while these observations were "atypical" relative to what had been observed historically (2017–2019), they were within the expected concentrations based on meteorology and emission proxies. This demonstrates the advantage of quantifying the influence of atypical events using the model predictions, mainly because the models can account for the influence of changing meteorology and certain emission proxies. In this case, the increased concentrations were likely due to NW winds and agricultural fires observed during the period. Interestingly, the increase in nitrate concentrations was not captured by model predictions in P3 (Figure 4d), indicating the influence of other factors such as additional formation pathways (e.g., organic nitrate) not accounted for by the model. Nevertheless, this analysis shows that the influence of atypical events may be quantified by random forest model predictions, given there is enough data available to train these models.

4. Conclusions

In summary, we used over three years of PM_1 data measured as part of the Delhi Aerosol Supersite study to build machine learning models using random forest regression to estimate PM_1 and its constituents by using meteorological parameters and emission proxies. Overall, parameters such as T, H, WD, RH, MOY and fire counts can capture the majority of variability in PM_1 and its constituent concentrations, indicating the importance of meteorology (specifically T, H, WD), agricultural burning (fire counts) and season-specific sources (e.g., domestic biomass burning) in governing the temporal variability of PM_1 constituent concentrations.

We used sensitivity analysis and feature ranking to understand the influence of individual factors and demonstrate that emission sources in winter months play an important role in contributing to high PM_1 concentrations. We also show that cloud cover and PM_1 concentrations are anti-correlated. It may be partly due to the semi-direct aerosol effect on clouds and co-variance of cloudy periods with higher temperatures. We encourage future studies to investigate aerosol-cloud interactions and their influence on air pollution in polluted cities. Through the variability of PM_1 with HOD and Org with T, we demonstrate the importance of secondary aerosol in Delhi. Further, our analysis is consistent with the influence of various emission sources on chloride, including industrial emissions from the NW, agricultural burning emissions and trash burning. Our results are also consistent with the influence of power plants and aqueous phase chemistry on sulfate concentrations. These analyses re-iterate the need for multi-sectoral and multi-regional policies to tackle air pollution in Delhi. Further, we utilized the predictive capability of the models to quantify the influence of the COVID-19 lockdown in 2020 on the air quality and demonstrated that the advantage of using this method over historical concentrations is that it can account for changing meteorology and certain emission sources. Overall, our analysis provides robust and detailed insights into the factors influencing PM_1 in Delhi and shows the applicability of machine learning methods in atmospheric science applications.

Acknowledgements

We thank Dr. Ann Marie Carlton for fruitful discussions about MODIS cloud data and Dr. Barron Henderson for providing the scripts to extract MODIS cloud data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This manuscript is based on work supported in part by the Welch Foundation under Grants F-1925-20170325 and F-1925-20200401 and the National Science Foundation under Grant 1653625. It is also supported by the University of Austin Cockrell School of Engineering Graduate Continuing Fellowship. We thank the funding agencies for their support.

ORCID

Kanan Patel http://orcid.org/0000-0002-2368-8624
Sahil Bhandari http://orcid.org/0000-0002-9822-420X
Shahzad Gani http://orcid.org/0000-0002-6966-0520
Purushottam Kumar http://orcid.org/0000-0002-7692-095X
Nisar Baig http://orcid.org/0009-0004-3889-2697
Joshua Apte http://orcid.org/0000-0002-2796-3478
Lea Hildebrandt Ruiz http://orcid.org/0000-0001-8378-1882

References

An, W. J., R. K. Pathak, B. H. Lee, and S. N. Pandis. 2007. Aerosol volatility measurement using an improved thermodenuder: Application to secondary organic aerosol. *J Aerosol Sci.* 38 (3):305–14. doi:10.1016/j.jaerosci.2006.12.002.

Andrew, R. 2020. India's daily electricity generation. CICERO Center for International Climate Research. Accessed August 24, 2022. https://folk.universitetetioslo.no/roberan/t/POSOCO.shtml

Arlot, S., and M. Lerasle. 2014. Why V = 5 is enough in V-fold cross-validation. *hal-00743931v2*. Accessed August 24, 2022. https://hal.archives-ouvertes.fr/hal-00743931v2/document

Arub, Z., S. Bhandari, S. Gani, J. S. Apte, L. Hildebrandt Ruiz, and G. Habib. 2020. Air mass physiochemical characteristics over New Delhi: Impacts on aerosol hygroscopicity and cloud condensation nuclei (CCN) formation. *Atmos. Chem. Phys.* 20 (11):6953–71. doi:10.5194/acp-20-6953-2020.

Automated Surface Observing System (ASOS). 1998. User's guide, national oceanic and atmospheric administration, Department of Defense Federal Aviation Administration, United States Navy. Accessed August 24, 2022. https://www.weather.gov/media/asos/aum-toc.pdf

Baig, N. A., M. Yawar, K. Jain, G. Singh, S. Singh, D. Siddharthan, and G. Habib. 2020. Association between traffic emissions mixed with resuspended dust and heart rate variability among healthy adults in Delhi. *Air Qual. Atmos. Health* 13 (3):371–8. doi:10.1007/s11869-020-00800-2.

Baltensperger, U., R. Chirico, P. F. DeCarlo, J. Dommen, K. Gaeggeler, M. F. Heringa, M. Li, A. S. H. Prévôt, M. R. Alfarra, D. S. Gross, et al. 2010. Recent developments in the mass spectrometry of atmospheric aerosols. *Eur J Mass Spectrom (Chichester)* 16 (3):389–95. doi:10.1255/ejms.1084.

Bengio, Y., and Y. Grandvalet. 2004. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* 5:1089–105. Accessed August 24, 2022. https://dblp.org/rec/journals/jmlr/BengioG04.

Bhandari, S., S. Gani, K. Patel, D. S. Wang, P. Soni, Z. Arub, G. Habib, J. S. Apte, and L. Hildebrandt Ruiz. 2020. Sources and atmospheric dynamics of organic aerosol in New Delhi, India: Insights from receptor modeling. *Atmos. Chem. Phys.* 20 (2):735–52. doi:10.5194/acp-20-735-2020.

Bosilovich, M. G., R. Lucchesi, and M. Suarez. 2016. MERRA-2: File Specification. GMAO Office Note No. 9 (Version 1.1). Accessed August 24, 2022. http://gmao.gsfc.nasa.gov/pubs/office_notes

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

Brook, R. D., S. Rajagopalan, C. A. Pope, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, American Heart Association Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, and Council on Nutrition, Physical Activity and Metabolism, et al. 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the

American Heart Association. *Circulation* 121 (21):2331–78. doi:10.1161/CIR.0b013e3181dbece1.

Campbell-Lendrum, D., and A. Prüss-Ustün. 2019. Climate change, air pollution and noncommunicable Diseases. *Bull. World Health Organ.* 97 (2):160–1. doi:10.2471/BLT.18.224295.

Chauhan, A., and R. P. Singh. 2020. Decline in $pm_{2.5}$ concentrations over major cities around the world associated with covid-19. *Environ. Res.* 187:109634. doi:10.1016/j.envres.2020.109634.

Christiansen, A. E., A. G. Carlton, and B. H. Henderson. 2020. Differences in fine particle chemical composition on clear and cloudy days. *Atmos. Chem. Phys.* 20 (19):11607–24. doi:10.5194/acp-20-11607-2020.

Christopher, S. A., and P. Gupta. 2010. Satellite remote sensing of particulate matter air quality: The cloud-cover problem. *J. Air Waste Manag. Assoc.* 60 (5):596–602. doi:10.3155/1047-3289.60.5.596.

Christopoulos, C. D., S. Garimella, M. A. Zawadowicz, O. Möhler, and D. J. Cziczo. 2018. A machine learning approach to aerosol classification for single-particle mass spectrometry. *Atmos. Meas. Tech.* 11 (10):5687–99. doi:10.5194/amt-11-5687-2018.

Cohen, A. J., M. Brauer, R. Burnett, H. R. Anderson, J. Frostad, K. Estep, K. Balakrishnan, B. Brunekreef, L. Dandona, R. Dandona, et al. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* 389 (10082):1907–18. doi:10.1016/S0140-6736(17)30505-6.

Durre, I., R. S. Vose, and D. B. Wuertz. 2006. Overview of the integrated global radiosonde archive. *J Clim* 19 (1):53–68. doi:10.1175/JCLI3594.1.

Faulhaber, A. E., B. M. Thomas, J. L. Jimenez, J. T. Jayne, D. R. Worsnop, and P. J. Ziemann. 2009. Characterization of a thermodenuder-particle beam mass spectrometer system for the study of organic aerosol volatility and composition. *Atmos. Meas. Tech.* 2 (1):15–31. doi:10.5194/amt-2-15-2009.

Feingold, G., H. Jiang, and J. Y. Harrington. 2005. On smoke suppression of clouds in Amazonia. *Geophys. Res. Lett.* 32 (2):1–4. doi:10.1029/2004GL021369.

Feng, R., H. j Zheng, H. Gao, A. r Zhang, C. Huang, J. x Zhang, K. Luo, and J. r Fan. 2019. Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: A case study in Hangzhou, China. *J Clean Prod* 231:1005–15. doi:10.1016/j.jclepro.2019.05.319.

Fire Information for Resource Management System (FIRMS). 2022. Accessed August 24, 2022. https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms

Gani, S., Bhandari, S. Patel, K. Seraj, S. Soni, P. Arub, Z. Habib, G. Hildebrandt Ruiz, L. Apte, and J. S. 2020. Particle number concentrations and size distribution in a polluted megacity: The Delhi Aerosol Supersite study. *Atmos. Chem. Phys.* 20 (14):8533–49. doi:10.5194/acp-20-8533-2020.

Gani, S., S. Bhandari, S. Seraj, D. S. Wang, K. Patel, P. Soni, Z. Arub, G. Habib, L. Hildebrandt Ruiz, and J. S. Apte. 2019. Submicron aerosol composition in the world's most polluted megacity: The Delhi Aerosol Supersite study. *Atmos. Chem. Phys.* 19 (10):6843–59. doi:10.5194/acp-19-6843-2019.

Gautam, S. 2020. The influence of covid-19 on air quality in india: a boon or inutile. *Bull. Environ. Contam. Toxicol.* 104 (6):724–6. doi:10.1007/s00128-020-02877-y.

Gelaro, R., W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, et al. 2017. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Climate* 30 (14):5419–54. doi:10.1175/JCLI-D-16-0758.1.

Giglio, L., W. Schroeder, and C. O. Justice. 2016. The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sens. Environ.* 178:31–41. doi:10.1016/j.rse.2016.02.054.

Grange, S. K., J. D. Lee, W. S. Drysdale, A. C. Lewis, C. Hueglin, L. Emmenegger, and D. C. Carslaw. 2021. COVID-19 lockdowns highlight a risk of increasing ozone pollution in European urban areas. *Atmos. Chem. Phys.* 21 (5):4169–85. doi:10.5194/acp-21-4169-2021.

Grieshop, A. P., J. M. Logue, N. M. Donahue, and A. L. Robinson. 2009. Laboratory investigation of photochemical oxidation of organic aerosol from wood fires 1: measurement and simulation of organic aerosol evolution. *Atmos. Chem. Phys* 9:1263–77. doi:10.5194/acp-9-1263-2009.

Guttikunda, S. K., and G. Calori. 2013. A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India. *Atmos. Environ.* 67:101–11. doi:10.1016/j.atmosenv.2012.10.040.

Guttikunda, S. K., and B. R. Gurjar. 2012. Role of meteorology in seasonality of air pollution in megacity Delhi, India. *Environ. Monit. Assess.* 184 (5):3199–211. doi:10.1007/s10661-011-2182-8.

Hansen, J., M. Sato, and R. Ruedy. 1997. Radiative forcing and climate response. *J. Geophys. Res.* 102 (D6):6831–64. doi:10.1029/96JD03436.

Health Effects Institute (HEI). 2020. Health Effects Institute: State of global air 2020: special report. Accessed August 24, 2022. https://www.stateofglobalair.org/

Hou, L., Q. Dai, C. Song, B. Liu, F. Guo, T. Dai, L. Li, B. Liu, X. Bi, Y. Zhang, et al. 2022. Revealing drivers of haze pollution by explainable machine learning. *Environ. Sci. Technol. Lett.* 9 (2):112–9. doi:10.1021/acs.estlett.1c00865.

International Satellite Cloud Climatology Project (ISCCP). 2022. NASA. Accessed August 24, 2022. https://isccp.giss.nasa.gov/role.html

Jain, S., and T. Sharma. 2020. Social and travel lockdown impact considering coronavirus disease (Covid-19) on air quality in megacities of India: Present benefits, future challenges, and way forward. *Aerosol Air Qual. Res.* 20 (6):1222–36. doi:10.4209/aaqr.2020.04.0171.

Jaiprakash, A. Singhai, G. Habib, R. S. Raman, and T. Gupta. 2017. Chemical characterization of $PM_{1.0}$ aerosol in Delhi and source apportionment using positive matrix factorization. *Environ. Sci. Pollut. Res. Int.* 24 (1):445–62. doi:10.1007/s11356-016-7708-8.

Justice, C. O., E. Vermote, J. R. G. Townshend, R. Defries, D. P. Roy, D. K. Hall, V. V. Salomonson, J. L. Privette, G. Riggs, A. Strahler, et al. 1998. The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Trans. Geosci. Remote Sensing* 36 (4):1228–49. doi:10.1109/36.701075.

Kallos, G., M. Astitha, P. Katsafados, and C. Spyrou. 2007. Long-range transport of anthropogenically and naturally produced particulate matter in the Mediterranean and North Atlantic: Current state of knowledge. *J Appl Meteorol Climatol* 46 (8):1230–51. doi:10.1175/JAM2530.1.

Kampa, M., and E. Castanas. 2008. Human health effects of air pollution. *Environ. Pollut.* 151 (2):362–7. doi:10.1016/j.envpol.2007.06.012.

Koren, I., Y. J. Kaufman, L. A. Remer, and J. V. Martins. 2004. Measurement of the effect of amazon smoke on inhibition of cloud formation. *Science* 303 (5662):1342–5. https://www.science.org/doi/10.1126/science.1089424.

Kumar, S. 2020. Effect of meteorological parameters on spread of covid-19 in india and air quality during lockdown. *Sci. Total Environ.* 745:141021. doi:10.1016/j.scitotenv.2020.141021.

Kumar, P., S. Hama, H. Omidvarborna, A. Sharma, J. Sahani, K. V. Abhijith, S. E. Debele, J. C. Zavala-Reyes, Y. Barwise, and A. Tiwari. 2020. Temporary reduction in fine particulate matter due to anthropogenic emissions switch-off during covid-19 lockdown in Indian cities. *Sustain. Cities Soc.* 62:102382. doi:10.1016/j.scs.2020.102382.

Louvaris, E. E., E. Karnezi, E. Kostenidou, C. Kaltsonoudis, and S. N. Pandis. 2017. Estimation of the volatility distribution of organic aerosol combining thermodenuder and isothermal dilution measurements. *Atmos. Meas. Tech.* 10 (10):3909–18. doi:10.5194/amt-10-3909-2017.

Lovrić, M., K. Pavlović, M. Vuković, S. K. Grange, M. Haberl, and R. Kern. 2021. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. *Environ. Poll.* 274:115900. doi:10.1016/j.envpol.2020.115900.

Mahato, S., S. Pal, and K. G. Ghosh. 2020. Effect of lockdown amid covid-19 pandemic on air quality of the megacity Delhi, India. *Sci. Total Environ.* 730:139086. doi:10.1016/j.scitotenv.2020.139086.

Manchanda, C., M. Kumar, V. Singh, M. Faisal, N. Hazarika, A. Shukla, V. Lalchandani, V. Goel, N. Thamban, D. Ganguly, et al. 2021. Variation in chemical composition and sources of pm2.5 during the covid-19 lockdown in Delhi. *Environ. Int.* 153:106541. doi:10.1016/j.envint.2021.106541.

Masson-Delmotte, V., P. Zhai, A. Pirani, S. Connors, L. Péan, C. Berger, S. Caud, N. Chen, Y. Goldfarb, L. Gomis, et al. 2022. IPCC: Climate Change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change. *Cambridge University Press*. Accessed August 24, 2022. https://www.ipcc.ch/report/ar6/wg1/#FullReport.

Mccarty, W., L. Coy, R. Gelaro, A. Huang, D. Merkova, E. B. Smith, M. Sienkiewicz, K. Wargan, and R. D. Koster. 2016. Technical Report Series on Global Modeling and Data Assimilation, Volume 46 MERRA-2 Input Observations: Summary and Assessment. Accessed August 24, 2022. https://gmao.gsfc.nasa.gov/pubs/docs/McCarty885.pdf

MERRA-2 website. 2022. Accessed August 24, 2022. https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/

Misra, P., R. Imasu, S. Hayashida, A. A. Arbain, R. Avtar, and W. Takeuchi. 2020. Mapping brick kilns to support environmental impact studies around Delhi using sentinel-2. *ISPRS Int J Geoinf* 9 (9):544. doi:10.3390/ijgi9090544.

Molina, L. T. 2021. Introductory lecture: air quality in megacities. *Faraday Discuss.* 226:9–52. doi:10.1039/D0FD00123F.

Morakinyo, O. M., M. I. Mokgobu, M. S. Mukhola, and R. P. Hunter. 2016. Health outcomes of exposure to biological and chemical components of inhalable and respirable particulate matter. *IJERPH.* 13 (6):592. doi:10.3390/ijerph13060592.

Muñoz-Sabater, J., E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, et al. 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* 13 (9):4349–83. doi:10.5194/essd-13-4349-2021.

Nair, A. A., F. Yu, P. Campuzano-Jost, P. J. DeMott, E. J. T. Levin, J. L. Jimenez, J. Peischl, I. B. Pollack, C. D. Fredrickson, A. J. Beyersdorf, et al. 2021. Machine learning uncovers aerosol size information from chemistry and meteorology to quantify potential cloud-forming particles. *Geophys Res Lett* 48:e2021GL094133. doi:10.1029/2021GL094133.

NASA SC11. 2022. Biomass burning aerosol effects on clouds and precipitation. *Our Planet*. Accessed August 24, 2022. https://www.nas.nasa.gov/SC11/demos/demo18.html

Ng, N. L., S. C. Herndon, A. Trimborn, M. R. Canagaratna, P. L. Croteau, T. B. Onasch, D. Sueper, D. R. Worsnop, Q. Zhang, Y. L. Sun, et al. 2011. An Aerosol Chemical Speciation Monitor (ACSM) for routine monitoring of the composition and mass concentrations of ambient aerosol. *Aero Sci. Tech.* 45 (7):780–94. doi:10.1080/02786826.2011.560211.

Oktas. 2022. World weather information service. Accessed August 24, 2022. https://worldweather.wmo.int/oktas.htm

Pande, P., M. Shrivastava, J. E. Shilling, A. Zelenyuk, Q. Zhang, Q. Chen, N. L. Ng, Y. Zhang, M. Takeuchi, T. Nah, et al. 2022. Novel application of machine learning techniques for rapid source apportionment of aerosol mass spectrometer datasets. *ACS Earth Space Chem.* 6 (4):932–42. doi:10.1021/acsearthspacechem.1c00344.

Pant, P., S. K. Guttikunda, and R. E. Peltier. 2016. Exposure to particulate matter in India: A synthesis of findings and future directions. *Environ. Res.* 147:480–96. doi:10.1016/j.envres.2016.03.011.

Pant, P., and R. M. Harrison. 2012. Critical review of receptor modelling for particulate matter: A case study of India. *Atmos. Environ.* 49:1–12. doi:10.1016/j.atmosenv.2011.11.060.

Pant, P., A. Shukla, S. D. Kohl, J. C. Chow, J. G. Watson, and R. M. Harrison. 2015. Characterization of ambient $PM_{2.5}$ at a pollution hotspot in New Delhi, India and inference of sources. *Atmos. Environ.* 109:178–89. doi:10.1016/j.atmosenv.2015.02.074.

Patel, K., S. Bhandari, S. Gani, M. J. Campmier, P. Kumar, G. Habib, J. Apte, and L. Hildebrandt Ruiz. 2021a. Sources and dynamics of submicron aerosol during the autumn onset of the air pollution season in Delhi. *ACS*

Earth Space Chem. 5 (1):118–28. doi:10.1021/acsearthspacechem.0c00340.

Patel, K., M. J. Campmier, S. Bhandari, N. Baig, S. Gani, G. Habib, J. S. Apte, and L. Hildebrandt Ruiz. 2021b. Persistence of primary and secondary pollutants in Delhi: Concentrations and composition from 2017 through the COVID pandemic. *Environ. Sci. Technol. Lett.* 8 (7):492–7. doi:10.1021/acs.estlett.1c00211.

Patel, K., D. Wang, P. Chhabra, J. Bean, S. V. Dhulipala, and L. Hildebrandt Ruiz. 2020. Effects of sources and meteorology on ambient particulate matter in Austin, Texas. *ACS Earth Space Chem.* 4 (4):602–13. doi:10.1021/acsearthspacechem.0c00016.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, M. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.* 12 (85): 2825–30. Accessed August 24, 2022. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pratt, K. A., and K. A. Prather. 2012. Mass spectrometry of atmospheric aerosols-recent developments and applications. Part II: On-line mass spectrometry techniques. *Mass Spectrom. Rev.* 31 (1):17–48. doi:10.1002/mas.20330.

Qin, Y., J. Ye, P. Ohno, P. Liu, J. Wang, P. Fu, L. Zhou, Y. J. Li, S. T. Martin, and C. K. Chan. 2022. Assessing the nonlinear effect of atmospheric variables on primary and oxygenated organic aerosol concentration using machine learning. *ACS Earth Space Chem.* 6 (4):1059–66. doi:10.1021/acsearthspacechem.1c00443.

Rosenfeld, D., U. Lohmann, G. B. Raga, C. D. O'Dowd, M. Kulmala, S. Fuzzi, A. Reissell, and M. O. Andreae. 2008. Flood or drought: How do aerosols affect precipitation? *Science* 321 (5894):1309–13. doi:10.1126/science.1160606.

Rosina, K., F. Batista e Silva, P. Vizcaino, M. Marín Herrera, S. Freire, and M. Schiavina. 2020. Increasing the detail of European land use/cover data by combining heterogeneous data sets. *Int J Digit Earth* 13 (5):602–26. doi:10.1080/17538947.2018.1550119.

Rubal, and D. Kumar. 2018. Evolving Differential evolution method with random forest for prediction of air pollution. *Proc. Computer Sci., Elsevier B.V* 132:824–33. doi:10.1016/j.procs.2018.05.094.

Salcedo, D., T. B. Onasch, K. Dzepina, M. R. Canagaratna, Q. Zhang, J. A. Huffman, P. F. Decarlo, J. T. Jayne, P. Mortimer, D. R. Worsnop, et al. 2006. Characterization of ambient aerosols in Mexico City during the MCMA-2003 campaign with Aerosol Mass Spectrometry: results from the CENICA Supersite. *Atmos. Chem. Phys.* 6 (4): 925–46. doi:10.5194/acp-6-925-2006.

Samset, B. H., M. Sand, C. J. Smith, S. E. Bauer, P. M. Forster, J. S. Fuglestvedt, S. Osprey, and C. F. Schleussner. 2018. Climate impacts from a removal of anthropogenic aerosol emissions. *Geophys. Res. Lett.* 45 (2):1020–9. doi:10.1002/2017GL076079.

Schraufnagel, D. E., J. R. Balmes, C. T. Cowl, S. De Matteis, S.-H. Jung, K. Mortimer, R. Perez-Padilla, M. B. Rice, H. Riojas-Rodriguez, A. Sood, et al. 2019. Air pollution and noncommunicable diseases: A review by the forum of international respiratory societies' environmental committee, Part 1: The damaging effects of air pollution. *Chest* 155 (2):409–16. doi:10.1016/j.chest.2018.10.042.

Seinfeld, J. H., and S. N. Pandis. 2006. *Atmospheric chemistry and physics: From air pollution to climate change*. 2nd ed. New Jersey: J. Wiley.

Sharma, M., S. Jain, and B. Y. Lamba. 2020. Epigrammatic study on the effect of lockdown amid covid-19 pandemic on air quality of most polluted cities of Rajasthan (India). *Air Qual. Atmos. Health* 13 (10):1157–65. doi:10.1007/s11869-020-00879-7.

Stirnberg, R., J. Cermak, S. Kotthaus, M. Haeffelin, H. Andersen, J. Fuchs, M. Kim, J. E. Petit, and O. Favez. 2021. Meteorology-driven variability of air pollution (PM_1) revealed with explainable machine learning. *Atmos. Chem. Phys.* 21 (5):3919–48. doi:10.5194/acp-21-3919-2021.

Su, T., Z. Li, Y. Zheng, Q. Luan, and J. Guo. 2020. Abnormally shallow boundary layer associated with severe air pollution during the COVID-19 lockdown in China. *Geophys. Res. Lett.* 47:e2020GL090041. doi:10.1029/2020GL090041.

United Nations: World urbanization prospects. 2018. Accessed August 24, 2022. https://population.un.org/wup/

United States Environmental Protection Agency. 2022. Criteria pollutants. Accessed August 24, 2022. https://www.epa.gov/criteria-air-pollutants

University of Illinois Urbana-Champaign. 2022. The weather world 2010 project, effects of cloud cover on forecasted temperatures. Accessed August 24, 2022. http://ww2010.atmos.uiuc.edu/(Gh)/wwhlpr/fcst_temps_cloud_cover.rxml.

Wang, Y., Y. Wen, Y. Wang, S. Zhang, K. M. Zhang, H. Zheng, J. Xing, Y. Wu, and J. Hao. 2020. Four-month changes in air quality during and after the COVID-19 lockdown in six megacities in China. *Environ. Sci. Technol. Lett.* 7 (11):802–8. doi:10.1021/acs.estlett.0c00605.

Wang, G., R. Zhang, M. E. Gomez, L. Yang, M. L. Zamora, M. Hu, Y. Lin, J. Peng, S. Guo, J. Meng, et al. 2016. Persistent sulfate formation from London Fog to Chinese haze. *Proc. Natl. Acad. Sci. U S A* 113 (48):13630–5. doi:10.1073/pnas.1616540113.

Wang, F., Z. Zhang, G. Wang, Z. Wang, M. Li, W. Liang, J. Gao, W. Wang, D. Chen, Y. Feng, et al. 2022. Machine learning and theoretical analysis release the non-linear relationship among ozone, secondary organic aerosol and volatile organic compounds. *J Environ Sci (China)* 114: 75–84. doi:10.1016/j.jes.2021.07.026.

World Health Organization (WHO). 2016. Ambient air pollution: a global assessment of exposure and burden of Disease 2016. Accessed August 24, 2022. https://apps.who.int/iris/handle/10665/250141

World Health Organization (WHO). 2021. Global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Accessed August 24, 2022. https://apps.who.int/iris/handle/10665/345329

Yang, J., Y. Wen, Y. Wang, S. Zhang, J. P. Pinto, E. A. Pennington, Z. Wang, Y. Wu, S. P. Sander, J. H. Jiang, et al. 2021. From COVID-19 to future electrification: Assessing traffic impacts on air quality by a machine-learning model. *Proc. Natl. Acad. Sci. USA.* 118 (26): e2102705118. doi:10.1073/pnas.2102705118.

Yu, R., Y. Yang, L. Yang, G. Han, and O. A. Move. 2016. RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors (Switzerland)* 16 (1): 86. doi:10.3390/s16010086.

Zhang, Q., J. L. Jimenez, M. R. Canagaratna, J. D. Allan, H. Coe, I. Ulbrich, M. R. Alfarra, A. Takami, A. M. Middlebrook, Y. L. Sun, et al. 2007. Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere mid-latitudes. *Geophys. Res. Lett.* 34:L13801. doi:10.1029/2007GL029979.

Zhang, Z., B. Xu, W. Xu, F. Wang, J. Gao, Y. Li, M. Li, Y. Feng, and G. Shi. 2022. Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on $PM_{2.5}$ pollution. *Environ. Res.* 212 (Pt B):113322. doi:10.1016/j.envres.2022.113322.