# PollutionMapper: Identifying Global Air Pollution Sources

DHRUV AGARWAL, Cornell University, United States
SRINIVASAN IYENGAR, Microsoft Corporation, India
PANKAJ KUMAR, CORAL, IIT Kharagpur, India

Air pollution adversely impacts public health. The National Capital Region (Delhi-NCR) is among the most polluted urban areas in the world. One component of air pollution is PM2.5, which accounts for around 80% of deaths due to air pollution. Solutions for lowering PM2.5 levels in Delhi have been ineffective due to their unscientific design. In this article, we build a mixed-methods model that captures the interplay of various factors—geographical, chemical, meteorological—that contribute to the concentration of PM2.5. Using domain knowledge and KDE sampling from NASA's GEOS-CF dataset, we identify the major sources of each of the seven constituents of PM2.5. From the 68 sources thus selected, we run the NOAA's HYSPLIT wind dispersion model to track the movement of released particles to the sink, i.e., Delhi. Using the concentration of pollutants at the sources and by tracking their movement, we can predict the PM2.5 levels at the sink and identify polluting sources. Our model performed significantly better than the baseline fixed-effects model and captured seasonal variations in all seven constituents of PM2.5. It also uncovered the impact of polluting sources hundreds of kilometers away on the air of Delhi. Policymakers can use such a model to design data-driven policy interventions.

CCS Concepts: • **Applied computing → Environmental sciences;**

Additional Key Words and Phrases: Air pollution, particulate matter, source identification, source apportionment, mixed-effects model

**ACM Reference format:**
Dhruv Agarwal, Srinivasan Iyengar, and Pankaj Kumar. 2024. PollutionMapper: Identifying Global Air Pollution Sources. *ACM J. Comput. Sustain. Soc.* 2, 1, Article 7 (January 2024), 23 pages.
https://doi.org/10.1145/3617129

## 1 INTRODUCTION

Air pollution is an increasingly dangerous health hazard. It causes short-term effects such as headache, throat inflammation, and skin irritation, as well as long-term effects including respiratory and cardiovascular diseases. Globally, 4–10 million deaths are attributed to air pollution each year, making it the largest environmental hazard causing premature deaths [17, 29]. More than 90% of these pollution-related deaths happen in low- and middle-income countries [29]. In India alone, over 1.5 million deaths are attributed to air pollution each year [29]—greater than the official number of deaths caused by Malaria, Tuberculosis, AIDS, and COVID-19 combined. The

Work done while D. Agarwal was at Microsoft Research India.
Authors' addresses: D. Agarwal, Information Science, Cornell University, Ithaca, NY, United States; e-mail: da399@cornell.edu; S. Iyengar, Microsoft Corporation, Bangalore, Karnataka, India; e-mail: sriyengar@microsoft.com; P. Kumar, CORAL, IIT Kharagpur, Kharagpur, West Bengal, India; e-mail: pankaj.kmr1990@gmail.com.

2834-5533/2024/1-ART7 $15.00
https://doi.org/10.1145/3617129

capital of India, Delhi, lies in the Indo-Gangetic plain, which is among the most polluted areas in the world [4].

Being densely populated and the capital of India, **Delhi and the surrounding National Capital Region (Delhi-NCR)** has attracted considerable national and international media attention year after year, highlighting the city's poor air quality [7, 40, 49]. Due to this public health hazard, governments and civic authorities in the state have initiated and encouraged a host of air pollution control measures. For instance, the Delhi government introduced the Odd-Even vehicle rule in 2016 that allows only vehicles with even number plates to ply the road on even calendar dates, and similarly for odd dates [42]. Similarly, local politicians have installed large-scale air purifiers in densely populated areas of the city [13, 35]. While these interventions are made in good faith, they have not been successful in reducing pollution levels in the city, leading to public debates about their effectiveness. For example, the Odd-Even traffic rule was judged ambiguous [28] and ineffective [8] because the main source of pollution was not vehicular traffic but other activities such as crop residue burning and construction. Similarly, it is widely accepted that air purifiers are useful for improving air quality in indoor spaces but are unscientific and ineffective when deployed to clean open outdoor spaces [22, 39]. To create effectual change, there is a need to use data-driven methods to not just evaluate interventions post hoc but to design them in the first place. A plethora of past work has designed models to forecast air quality [4, 10, 11, 33], but these models do not identify polluting sources, rendering them futile for designing interventions. Other works [19, 45, 50] identify polluting sources but do not provide a model to design policy.

To design policy interventions, it is important to identify polluting sources [32]. This requires modeling the interplay of the various factors affecting air quality in a city. We classify these factors as: (1) geography, (2) chemistry, and (3) meteorology. The rise (or fall) in air pollution can be a result of a change in any of these factors. For example, in July–October, monsoon winds blow into India from the southwest, thereby carrying dust from the Thar Desert into Delhi. As a result, during these months, PM2.5 concentration in Delhi is primarily due to its geography and meteorology (the effect of wind). Our model is a cross-disciplinary collaboration between environmental scientists who understand the factors affecting air pollution and computer scientists who can access and analyze large amounts of remote sensing data efficiently.

In this article, we present PollutionMapper,[1] a source-identification model that tracks pollutant particles from their sources and captures the geographical, chemical, and meteorological factors affecting the PM2.5[2] concentration at a "sink" location. In this article, we define our sink as Delhi, but with local knowledge, our approach can be extended to any other city in the world. We focus on modeling PM2.5 over other air pollutants (e.g., PM10), because it is responsible for the majority of deaths from air pollution. PollutionMapper uses a mixed-effects model so we can account for group identities in PM2.5 data occurring due to seasonality and meteorology. We use NASA's GEOS-CF dataset [27] to identify 68 sources of pollution and use NOAA's HYSPLIT wind dispersion model [47] to track the movement of particles from the sources to the sink (Delhi). We build seven such models, one for each of the seven constituents of PM2.5. Each model predicts the hourly concentration of the PM2.5 constituent at Delhi. While other models may predict/forecast PM2.5 better, our main contribution is that PollutionMapper can identify polluting sources, which allows policymakers to focus their attention on those sources.

Our results show that the mixed-effects model generally outperforms the baseline fixed-effects model and is able to capture seasonal variations in PM2.5 variations. A feature importance analysis of the model reveals the effect of sources such as Singrauli and Saharanpur, which are hundreds of

[1]Our code is available at: https://github.com/agdhruv/PollutionMapper

[2]PM2.5 refers to particulate matter that has an aerodynamic diameter <2.5 microns.

kilometers away, on the PM2.5 concentration of Delhi. Such a model can help policymakers test interventions in a way that considers all the important factors affecting air pollution. For example, before implementing the Odd-Even vehicle rule, the Delhi government could preemptively test its impact on the city's air quality. Moreover, such a model could help reveal novel interventions by uncovering the impact of unexpected polluting sources on the city.

The rest of this article is organized as follows: In Section 2, we provide some background by explaining the geographical, chemical, and meteorological factors affecting air pollution. In Section 3, we use this background to build our source-identification model. We then explain the evaluation methodology for the model in Section 4 and present results in Section 5. Finally, we discuss related work and conclude with a summary and the limitations of our work.

## 2 BACKGROUND

PM2.5 is a combination of seven constituents: sulphates, nitrates, **organic carbon (OC), secondary organic aerosols (SOA), black carbon (BC), dust (Du)**, and **sea salt (SS)**. We use the following respective shorthand for these constituents, in part to emphasize throughout the article that these are constituents of PM2.5: $PM_{Su}$, $PM_{Ni}$, $PM_{OC}$, $PM_{SOA}$, $PM_{BC}$, $PM_{Du}$, $PM_{SS}$. We separately model the three factors—geography, chemistry, and meteorology—that affect the concentration of these PM2.5 constituents. For this, we use domain knowledge and use knowledge from atmospheric chemistry. The plots in this section are derived from various datasets (e.g., GEOS-CF [27], GPW [12], NPP [36]); all these datasets are described in Section 3.1.

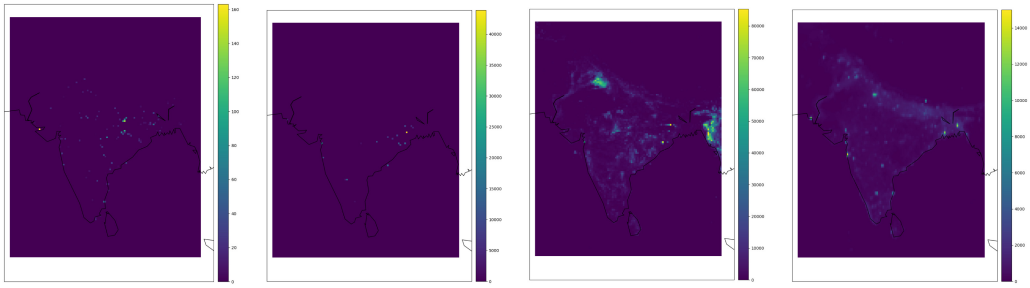
### 2.1 Geography

India generates most of the air pollution that plagues it. This is because of two main reasons. First, India is a peninsula—the southern half of the country is surrounded by ocean on three sides and no air pollution originates in the ocean. Second, the Great Himalayan range, which runs throughout India's north and northeast, acts as a barrier against wind and pollution originating in the industry-heavy Republic of China. In fact, even before being blocked by the Himalayas, the wind from China is blocked by the Tibetan plateau due to its average elevation of 4 km above sea level. The remaining part of the Indian subcontinent between the Himalayas and the oceans consists of a few other countries apart from India—Nepal, Bhutan, Bangladesh, Pakistan, and Sri Lanka. These countries are small in comparison to India and are agro-economies, so they do not produce a lot of air pollution.

Using domain knowledge, it is possible to identify the sources responsible for each of the seven constituents of PM2.5. Because India produces its own pollution, this means that to model air pollution in India, we need to identify the polluting sources mostly within India (and a few neighboring countries at most). Broadly, we identify four main polluting sources in India: thermal power plants, steel plants and other industries, biomass burning, and human activity (e.g., traffic, diesel generators). These four types of sources are shown in Figure 1 and explained in detail below.

*2.1.1 Thermal Power Plants.* Long-running (legacy) thermal power plants in India use coal to generate power. Most of them use Indian coal, which is sulphur-rich and low in carbon content. Hence, many of these power plants in India are located in and around the mineral-rich Chhota Nagpur region in central India. Proximity to minerals reduces the transport cost of raw materials required to generate electricity. Some of the newer thermal power plants in India use coal imported from Australia and Indonesia. Hence, these are located near the coasts for easy access to ports. They are less polluting because they use carbon-rich imported coal.

*2.1.2 Steel Plants and Other Industries.* Other polluting industries such as steel and cement plants also depend on mineral ore for raw materials. So, they are also located close to the Chhota



(a) Thermal Power (b) Steel Plants (c) Biomass Burning (d) Human Activity

Fig. 1. The four major polluting sources modeled in our article. For power plants, we have plotted the thermal power production data across India on a random day [36]. For steel plants, we have plotted the production capacity of steel plants in India [34]. For biomass burning, we have plotted active fire data from NASA VIIRS data. And for human activity, we have used population density data [12] in 2020 as a proxy.

Nagpur region in central India. Additionally, these industries tend to be closer to the coasts to reduce their transportation costs of importing other raw materials or exporting finished products. Hence, these industries are concentrated slightly east of the Chhota Nagpur region, closer to India's east coast (bordering the Bay of Bengal).

*2.1.3 Biomass Burning.* Biomass burning is prevalent in mainly two regions. Punjab in the northwest is an agriculture-heavy state. Farmers burn crop residue to clear their lands at the culmination of each harvesting season. Due to Punjab's proximity to Delhi, this **crop residue burning (CRB)** has a major impact on Delhi's pollution and has attracted much research attention from environmental scientists [4]. Biomass burning is also common in the northeastern states where dense forests are victim to forest fires.

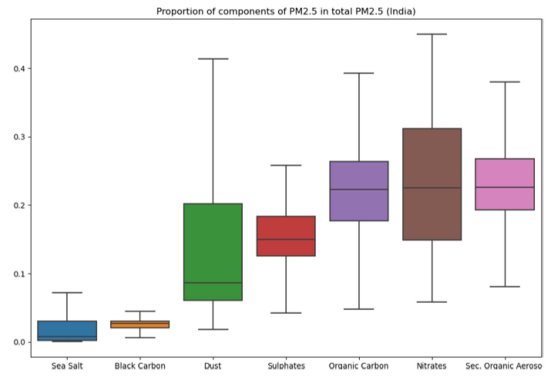
*2.1.4 Human Activity.* Finally, human activity also causes pollution. In urban areas, the common sources are petrol and diesel vehicles, diesel generators during power cuts, and construction activity. In urban areas, people burn firewood (biomass) to cook food or keep their houses warm in the winter. Since it is difficult to find data sources for such quotidian tasks, we use population density as a proxy for human activity. Densely populated cities are spread throughout the country with no pattern.

## 2.2 Chemistry

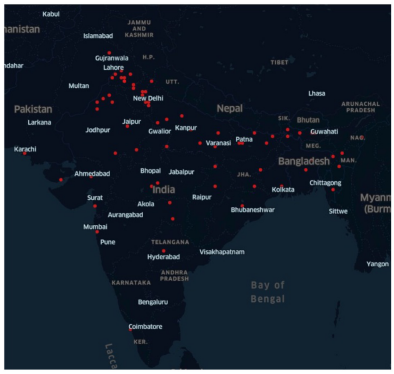
Polluting sources usually release primary pollutants such as gases (e.g., $SO_2$, $NO_2$) and hydrocarbons. Primary pollutants are short-lived and react in the atmosphere to form secondary pollutants such as $PM_{Su}$ and $PM_{Ni}$. Secondary pollutants are solid, persist in the atmosphere for longer durations, and travel long distances with the wind. These reactions happen all the time and are complex to model from first principles. In this section, we describe our method of modeling each of the seven constituents of PM2.5 using domain knowledge from atmospheric chemistry (later summarized in Table 1).

Figure 2(a) shows the proportions of the seven constituents of PM2.5 in the India++ region.[3] Overall, just four constituents ($PM_{Ni}$, $PM_{Su}$, $PM_{SOA}$, $PM_{OC}$) account for 80%–85% of the PM2.5 concentration in the region. $PM_{Du}$ and $PM_{SS}$ are seasonal, as they are brought in by the monsoon winds,

[3]In this article, our study area is the India++ region, a rectangular region consisting of the Indian sub-continent and southern China. The corners of this region in the Mercator projection are: llcrnrlat=4.75, llcrnrlon=67.0, urcrnrlat=38.0, urcrnrlon=97.25.

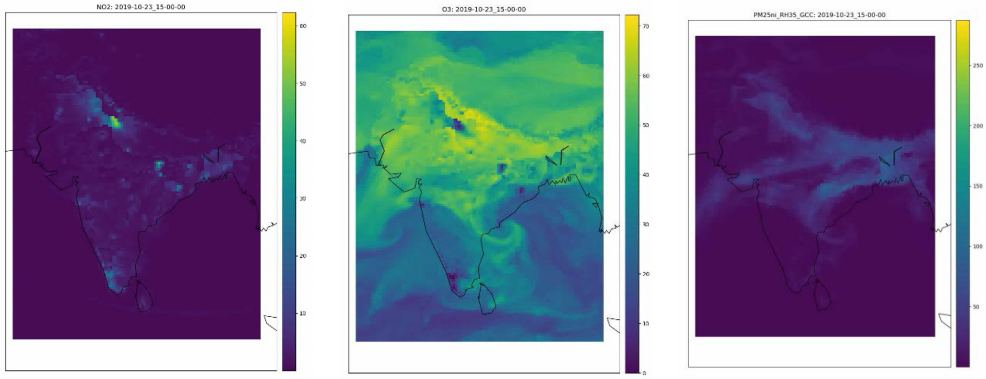



(a) PM2.5 Constituents




(b) 68 polluting sources selected

Fig. 2. (a) shows the proportions of the seven constituents of PM2.5 in the India++ region (see footnote 3). (b) shows the 68 locations selected as polluting sources, as explained in Section 2.1.



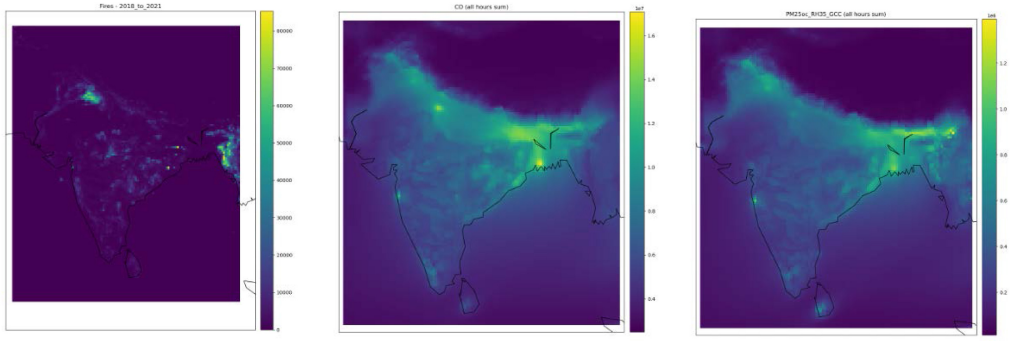

(a) $NO_2$ (b) $O_3$ (c) Nitrates

Fig. 3. This figure shows the concentration of $NO_2$, $O_3$, and nitrates in the India++ region at the same time. (a) shows that $NO_2$ is concentrated in cities and industrial areas. (b) shows that the concentration of $O_3$ is lower wherever $NO_2$ is higher—the brightest spots in (a) match the darkest spots in (b). This depicts the oxidation of $NO_2$ by ozone to form nitrates. (c) shows that, unlike $NO_2$, nitrates are widespread, because they are solid particles and are carried away by wind.

hence the high variability in the box plot for these two constituents. They are also comparatively stable and less harmful. $PM_{BC}$ has a very low proportion but is very harmful.

*2.2.1 Sulphates and Nitrates.* $SO_2$ and $NO_2$ are released in the atmosphere by polluting sources, which oxidize in the presence of atmospheric **oxygen ($O_2$)** and **ozone ($O_3$)** to form $PM_{Su}$ and $PM_{Ni}$, respectively. Fortunately, sources of $SO_2$ and $NO_2$ are concentrated and easily identifiable. $SO_2$ is released primarily by thermal power plants that use sulphur-rich Indian coal. $NO_2$ is released by diesel engines used in industries and diesel generators in cities. As discussed in Section 2.1, the locations of these sources are well-defined. Since $PM_{Su}$ and $PM_{Ni}$ behave similarly in the atmosphere, we explain their chemistry using the example of nitrates.

Visualization of GEOS-CF data in Figure 3(a) confirms that $NO_2$ is concentrated in cities and industrial regions. However, *nitrates* are spread across the country. This is because, unlike $NO_2$,



(a) Fires (b) Carbon Monoxide (CO) (c) Organic Carbon ($PM_{OC}$)

Fig. 4. This figure shows fires over three years and the sum of CO and $PM_{OC}$ concentration over three years in the India++ region. Since $PM_{OC}$ is released directly into the atmosphere during the combustion of biomass or diesel, we see that the $PM_{OC}$ plot is bright where either the fires plot or the CO plot is bright.

nitrates are chemically inert solid particles that travel long distances, depending on the direction and intensity of wind (Figure 3(c)). Figure 3(b) shows the concentration of ozone on the same day. Note that the concentration of ozone is low where the concentration of $NO_2$ is high.[4] This behavior shows that nitrates are formed when $NO_2$ oxidizes in the presence of ozone.

Hence, using the concentration of $SO_2$ and $NO_2$ at their source and wind dispersion pattern to track their travel, we can predict the concentration of $PM_{Su}$ and $PM_{Ni}$ at the sink.

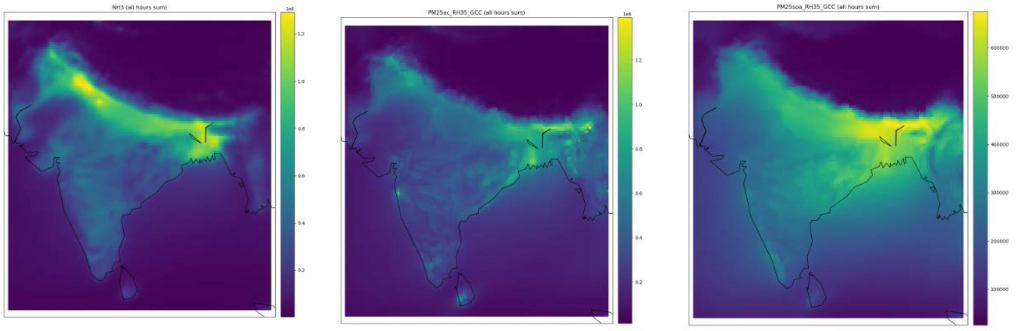
*2.2.2 Organic Carbon.* Unlike other constituents of PM2.5, $PM_{OC}$ is released directly into the atmosphere. It is not a secondary pollutant formed after reacting in the atmosphere. There are two major sources that release $PM_{OC}$ into the atmosphere: biomass burning and combustion of diesel, e.g., diesel engines used in agriculture.

We use NASA Fires data to conveniently identify biomass burning. Identifying diesel combustion is more complicated, because there is no good proxy to identify the combustion of diesel. Fortunately, there is a good proxy for combustion—**carbon monoxide (CO)**, which is formed due to incomplete combustion. The presence of CO signifies combustion, which in turn suggests that $PM_{OC}$ may also have been released during that combustion event. Note that carbon monoxide does not react to form $PM_{OC}$, instead, it is (positively) correlated to $PM_{OC}$. Hence, we use it as a proxy to identify the release of $PM_{OC}$ into the atmosphere. Figure 4 shows the concentration of CO, $PM_{OC}$, and fires summed over three years. Notice that the $PM_{OC}$ plot is brighter where fires or CO plots are brighter.

Hence, using fire data and concentration of CO, and wind dispersion to track travel, we can predict the concentration of $PM_{OC}$ at the sink.

*2.2.3 Secondary Organic Aerosols.* $PM_{SOA}$ is a secondary pollutant that is formed when organic carbon ($PM_{OC}$) travels over long distances and, over time, reacts to form $PM_{SOA}$. One of the main chemicals it reacts with, to form $PM_{SOA}$, is **ammonia ($NH_3$)**. Hence, the presence of both $NH_3$ and $PM_{OC}$ is needed for the formation of $PM_{SOA}$. Figure 5 shows the concentration of $NH_3$, $PM_{OC}$, and $PM_{SOA}$ summed over three years. Note that $PM_{SOA}$ is bright in the northeast region where both $NH_3$ and $PM_{OC}$ are bright. However, the concentration of $NH_3$ is very high in the Punjab region, but, since $PM_{OC}$ is lesser in that area, the resulting concentration of $PM_{SOA}$ is also lesser.

[4]We use this simplified inverse view of the relationship between $NO_2$ and $O_3$. In reality, this relationship is complex and depends on the relative availability of $NO_2$ and volatile organic compounds.



(a) Ammonia ($NH_3$) (b) Organic Carbon ($PM_{OC}$) (c) Sec. Org. Aerosols ($PM_{SOA}$)

Fig. 5. This figure shows the sum concentration of $NH_3$, $PM_{OC}$, and $PM_{SOA}$ over three years in the India++ region. $PM_{OC}$ reacts with ammonia to form $PM_{SOA}$. Hence, we see that $PM_{SOA}$ (sub-figure c) is bright when *both* $NH_3$ and $PM_{OC}$ are bright, for example, in the northeast region.

Hence, using the concentration of $PM_{OC}$ and $NH_3$ at the source, and wind dispersion to track travel, we can predict the concentration of $PM_{SOA}$ at the sink.

*2.2.4 Dust, Sea Salt.* $PM_{Du}$ and $PM_{SS}$ are chemically stable, so they do not react during travel from one location to another. They are simply picked up by the wind from one location and deposited in another location. Figure 6(a) shows that during the monsoon season (June–September), winds blowing from the southwest carry $PM_{Du}$ from the Thar Desert (Rajasthan) into Delhi/NCR. These monsoon winds also carry $PM_{SS}$ from the ocean into mainland India. Figure 6(b) shows the $PM_{SS}$ concentrations just before the monsoon season, and Figure 6(c) shows $PM_{SS}$ concentrations 12 days later at the onset of monsoon. These two figures show the movement of $PM_{SS}$ from the oceans to India's mainland with the monsoon winds.

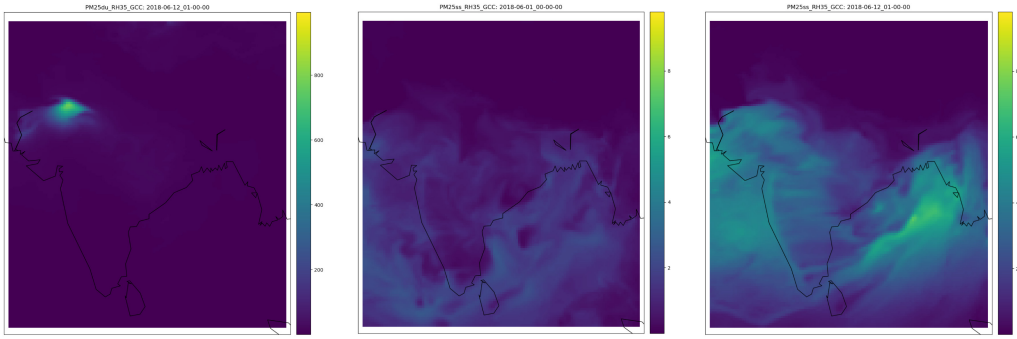
Because of their chemically stable nature, $PM_{Du}$ and $PM_{SS}$ can be very directly modeled using their concentration and wind dispersion data to track their movement.

*2.2.5 Black Carbon.* $PM_{BC}$ is a harmful pollutant that has similar sources as $PM_{OC}$. $PM_{BC}$ is released when $PM_{OC}$ is released, albeit in much lesser quantity. Figure 7(a) shows the strong correlation between $PM_{BC}$ and $PM_{OC}$ (Pearson's $r = 0.99$). This is good news, because now we do not need to model $PM_{BC}$ separately. Hence, we can model $PM_{BC}$ in the same way that we model $PM_{OC}$— using fire data and concentration of CO.

## 2.3 Meteorology

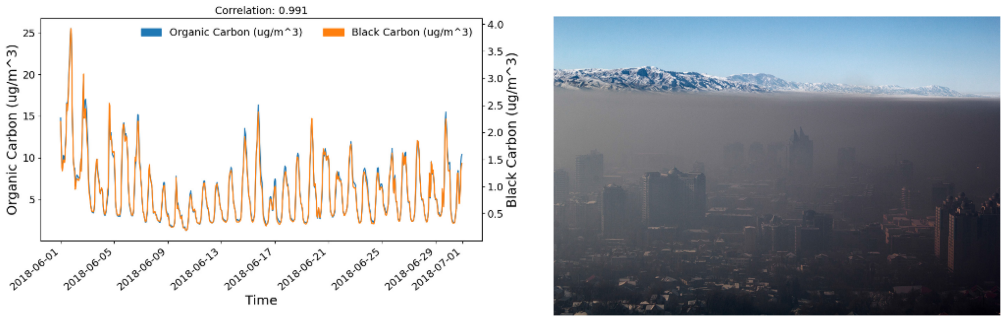
Air movement in the earth's atmosphere is caused due to gradients. Horizontal movement is generally caused due to a pressure gradient, causing winds to blow from high-pressure to low-pressure areas. Winds carry solid **particulate matter (PM)** over long distances, affecting the air quality of locations on the way. Vertical air movement is caused due to temperature gradient. These vertical convection currents dilute the concentration of PM near the earth's surface. In special atmospheric conditions called temperature inversion, vertical currents are impeded, causing PM to be trapped near the surface. Below, we explain our modeling of temperature inversion (vertical) and winds (horizontal).

*2.3.1 Temperature Inversion.* In normal atmospheric conditions, the air gets cooler as we go higher up from the earth's surface. This temperature gradient causes convection currents, which diffuse the particulate matter near the earth's surface over a large volume of air. However, during



(a) Dust (b) Sea Salt (pre-monsoon) (c) Sea Salt (12 days later)

Fig. 6. This figure shows the concentration of dust ($PM_{Du}$) and sea salt ($PM_{SS}$) in the India++ region during India's monsoon season. (a) shows the monsoon winds blowing from the southwest carrying $PM_{Du}$ particles from India's Thar Desert (the bright spot) into Delhi. (b) shows the concentration of $PM_{SS}$ before the onset of monsoon. (c) shows the higher concentration of $PM_{SS}$ in mainland India just 12 days later, once the monsoon season has begun.




(a) $PM_{OC}$-$PM_{BC}$ Correlation (b) Smog trapped

Fig. 7. (a) Correlation between Organic Carbon and Black Carbon over a random month in our study period. (b) shows smog trapped close to the earth's surface in Almaty, Kazakhstan, during a temperature inversion. Photo credits to Igors Jefimovs on Wikipedia under CC BY 3.0 license.

the winter season (and at night), the earth cools very quickly due to its higher thermal mass than air. Air near the earth's surface gets cooler and is trapped by a layer of warmer air above it. This causes an inversion in the temperature gradient, shutting off convection currents. This phenomenon is called a temperature inversion.

Temperature inversion has a multiplicative effect on the PM concentration of a particular location. It causes an "envelope" to form over the earth's surface, which traps all air and its PM near the surface, as shown in Figure 7(b). Since temperature inversions are more common during the winter months, the air pollution problem in Delhi intensifies during the winter months. In PollutionMapper, we account for this phenomenon by incorporating the **planetary boundary layer height (PBLH)**, which is the height in meters of the first temperature inversion in the atmosphere. The lower the PBLH, the lesser volume there is for particulate matter to diffuse in, and therefore there is a higher PM concentration closer to the earth's surface. Figures 8(a) and (b) show the sum of PM2.5 concentration and mean PBLH for a winter month (January 2018). On comparing the two

images, we see that PM2.5 is lower wherever the PBLH is higher. Most noticeably, a higher PBLH in the Tibetan plateau (north of India) causes a low PM2.5 concentration in the region.

*2.3.2 Wind Dispersion.* As discussed in Section 2.2, given the concentration of pollutants at their sources of release, we can track their transport through the atmosphere to the sink. To track pollutant particles through wind, we use NOAA's[5] HYSPLIT model [47], which uses Lagrangian simulations to track the air particle movement. HYSPLIT has two models: trajectory and dispersion. The *trajectory* model determines the trajectory of air parcels through the atmosphere. The *dispersion* model, which we use, computes the dispersion of pollutant particles through the ambient atmosphere. HYSPLIT's dispersion mode is often used to track volcanic ash and wildfire smoke to plan relief efforts. Further, the HYSPLIT dispersion model allows *forward* and *backward* simulations. A forward run in dispersion mode computes how air particles released from a source location will disperse in the atmosphere in the hours to come. A backward run computes the opposite: how the air particles arriving at a receptor location (the sink) were dispersed in the atmosphere on their way to this location, hours before they actually reached the sink.

In our formulation, we run the HYSPLIT dispersion model in both forward and backward modes. We run it in the forward mode from the 68 selected polluting sources in the India++ region; we release particles from each source every day and track the dispersion of particles for the next 48 hours (2 days). We also run the dispersion model in backward mode at the sink (Delhi) every hour and track where the particles there are coming from over the last 168 hours (7 days).[6][7] These HYSPLIT runs output the position (latitude, longitude, height) of air particles leaving the sources (forward mode) and those coming into the sink (backward mode).

Then, we look for instances of *confluence*: an event where a particle originating from any polluting source (identified from HYSPLIT forward run data) collides with a particle flowing into the sink (identified from HYSPLIT backward run data); this is depicted in Figure 8(c). Note that, since it is computationally infeasible to identify exact collisions of particles (at the exact latitude, longitude, and height), we "gridifiy" the atmosphere by using discretizations specified in the GFS dataset.[8] The existence of a confluence then suggests that a pollutant particle released from a source traveled to the sink. In this setup, a pollutant may travel up to nine days after its release from a source to reach the sink (after two days of forward dispersion, the particle may collide with a particle from backward dispersion, which may travel up to seven days to reach the sink).

## 3 POLLUTIONMAPPER: GLOBAL MODEL

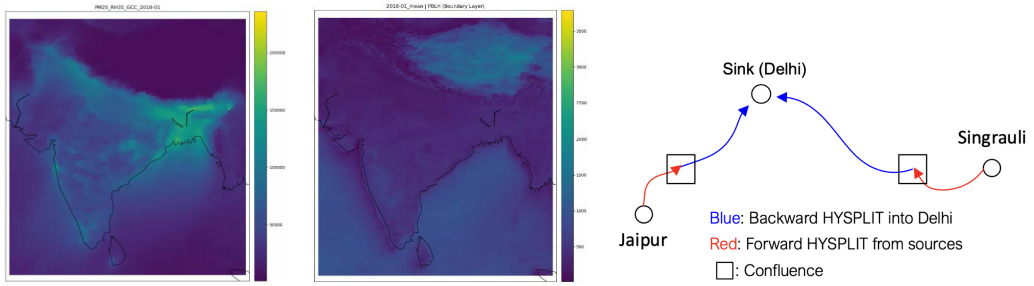
In the previous section, we discussed the geographical, chemical, and meteorological factors affecting air pollution. We observed that India has limited pollution sources, which we classified as thermal power plants, steel plants, biomass burning, and human activity. Then, we identified the seven constituents of PM2.5 and described how we can model each of them based on the concentration of pollutants at the polluting sources. Finally, we described the meteorology—wind and temperature inversion—that affects PM concentration on the earth's surface.

[5]National Oceanic and Atmospheric Administration.

[6]The reason to do both forward and backward runs to track particle dispersion from the sources to the sink is the computational complexity of HYSPLIT simulations. The trivial way would be to release particles at the source and see if they end up at the sink. However, due to the volume of the atmosphere, the chances of this happening in a simulation with a small number of particles are very low.

[7]The atmospheric residence time of PM2.5 is 3–5 days [48]. So, we allowed the particles emitted from the source to travel for two days (their "peak," after which they would start to get removed from the atmosphere) and then gave a long 7-day buffer for these particles to travel to the sink.

[8]$0.25^\circ \times 0.25^\circ$ for latitude and longitude and a non-uniform discretization for altitude/height.




(a) Combined PM2.5 (b) Mean PBLH (c) Confluence of particles

Fig. 8. (a) shows the sum of PM2.5 concentration (combined across the seven constituents) in the India++ region in January 2018. (b) shows the mean planetary boundary layer height (PBLH) in the India++ region in January 2018. Noticeably, the PM2.5 concentration is high (bright) where the boundary layer in low (dark) and vice versa. (c) depicts a *confluence* of particles originating from two polluting sources (Jaipur and Singrauli) with the particles flowing into the sink (Delhi).

Table 1. The First Two Columns Show the PM2.5 Components and Their Predictors

| Constituent | Predictors (at 68 sources) | No. of Features |
|---|---|---|
| Sulphate ($PM_{Su}$) | $SO_2$ | $1 \times 68 \times 9 = 612$ |
| Nitrate ($PM_{Ni}$) | $NO_2$ | |
| Dust ($PM_{Du}$) | Dust | |
| Sea salt ($PM_{SS}$) | Sea salt | |
| Organic Carbon ($PM_{OC}$) | CO<br>Active Fires | $2 \times 68 \times 9 = 1{,}224$ |
| SOA ($PM_{SOA}$) | $NH_3$<br>OC | |
| Black Carbon ($PM_{BC}$) | CO<br>Active Fires | |

Predictors are the pollutants used in our regression model to predict the respective PM2.5 constituent. The last column shows the number of features used in the training data for the respective constituent. Number of features = number of predictors × number of polluting sources (68) × number of days of wind dispersion (9).

In this section, we assemble the three factors discussed in the previous section into a source-identification model. We build the model in three steps. First, we describe **data sources** for the predictor pollutants (e.g., $SO_2$, $NO_2$) or their proxies (e.g., population density, fires). Then, we sample from this data to **select source locations**. Finally, we put everything together to **define the feature set** for predicting PM2.5 constituents at the sink. In the next section, we input these features into different models to predict PM2.5 at the sink (Delhi).

### 3.1 Pollutant Data Sources

In Sections 2.1 and 2.2, we defined predictors for each of the seven constituents of PM2.5. The constituents and their predictors are summarized in Table 1. We used NASA's GEOS-CF forecast dataset [27] for most of our predictor pollutants. GEOS-CF is an atmospheric composition forecast

from NASA, which uses an atmospheric chemistry transport model called GEOS-Chem for generating the forecasts. It is commonly used for estimating pollutants (e.g., Reference [31]). This data is available at an hourly temporal granularity and $0.25^\circ \times 0.25^\circ$ spatial granularity ($\sim$ 25 km $\times$ 25 km at the equator). The only predictor data we use that is not measured by GEOS-CF is active-fires data, for which we use NASA's VIIRS dataset. Finally, to compute wind dispersion, we use NASA's HYSPLIT model [47], which internally uses meteorology data from the **GFS (Global Forecast System)** model from the United States National Weather Service. The GFS dataset also provides us with PBLH data that we use to model temperature inversion.

In addition to the data for the predictor pollutants, we used other data sources for selecting the 68 source locations. To identify the power plants in India (sources of $SO_2$), we fetched data from the National Power Portal of India [36]. To identify the major steel manufacturing plants in India (sources of $NO_2$), we used the Global Energy Monitor data [34]. To identify the population density as a proxy for human activity (source of $NO_2$), we used the NASA **Gridded Population of the World (GPW)** v4 dataset [12]. Some of these datasets are available at different spatial granularities. Hence, to make these datasets usable in conjunction with GEOS-CF data, we re-aligned all of them to the same $0.25^\circ \times 0.25^\circ$ spatial grid used by GEOS-CF. Next, we explain how we use the data from these data sources to select the 68 polluting source sites.

### 3.2 Source Locations Selection

A key observation in our work is that there are few sites in the country that are responsible for releasing the most pollutants. In this section, we describe our process of selecting these polluting sources around the country. We used a two-dimensional probability density function on the data sources described above, supplemented by domain knowledge described in Section 2.1, to select pollution sources in the India++ region. For example, to select major sources of $PM_{Ni}$, we plotted the gridded data for the predictors and proxies of nitrates—$NO_2$, population density, and steel plants data—and sampled the "bright spots" from these plots. However, since we needed to perform forward HYSPLIT runs for all the selected sources, which is computationally expensive, instead of sampling blindly, we applied domain knowledge in the sampling process. We preferred sites that were closer to Delhi. We also preferred sites that came up as sources of multiple pollutants (e.g., thermal power plants release both $PM_{Su}$ and $PM_{Ni}$). We also manually discarded sites that were spatially close to other sites that we had already sampled, because the wind blowing in these locations would have similar dispersion patterns. We repeated the sampling process for all seven constituents of PM2.5, selecting source sites for each pollutant. We stopped sampling when we thought that the selected sites adequately covered the main polluting regions, giving us a total of 68 source locations shown in Figure 2(b).

### 3.3 Building the Feature Set

Now, we discuss the feature set for predicting the seven constituents of PM2.5 at the sink. After selecting the source locations and computing the confluence of wind particles originating from these locations and flowing into the sink, we used this data to create our feature set. As summarized in Table 1, the number of features for each of the constituents is defined by three quantities: the number of predictors for the constituent, the number of source locations (68), and the number of days for which a particle can travel from the source to the sink (9 days). We have different features for different travel durations (up to 9 days), because some pollutants take time to react and form particulate matter. For example, air particles that traveled for nine days may have more $PM_{Su}$ than air particles that traveled for one day, which may have more $SO_2$.

So, for instance, there are $2 \times 68 \times 9 = 1,224$ features for predicting $PM_{SOA}$ at the sink, as there are two predictors for SOA: ammonia ($NH_3$) and organic carbon ($PM_{OC}$). Air particles carrying

$NH_3$ and $PM_{OC}$ travel from the 68 selected sources for up to 9 days to reach the sink. Hence, the features for the $PM_{SOA}$ model are:

```
OC_source1_day1,\ldots, OC_source1_day9,\ldots, OC_source68_day1,\ldots, OC_source68_day9,
NH3_source1_day1,\ldots, NH3_source1_day9,\ldots, NH3_source68_day1,\ldots, NH3_source68_day9
```

In our prediction task, we predict the concentration of each PM2.5 constituent at Delhi (the sink) every hour. So, the feature set has a value for each of these features every hour for three years. The value of a feature `predictor_source`$S$`_day`$D$ at time $t$ is defined as follows: If we found a confluence (i.e., if particles originating from $S$ traveled to the sink in $h$ hours such that $\lceil h/24 \rceil = D$ and the particle reached the sink at time $t$), then the value of the feature is the concentration of the predictor $D$ days before $t$ at location $S$. If a confluence was not found, then the value of the feature is zero. Intuitively, this means that the value of the feature is non-zero if the wind carried the pollutants from the source $S$ to the sink over $D$ days.

## 4 EVALUATION METHODOLOGY

In this section, we explain the experimental setup used to train the models and the evaluation metric used to compare the models.

### 4.1 Experimental Setup

We created a regression task to predict the concentration of each of the seven constituents of PM2.5 at the sink at an hourly granularity using the features explained previously. We used four regression algorithms to train our model: **linear regression (LR), Light Gradient-Boosting Machine (LGBM)** [26], **k-Nearest Neighbors (KNN)**, and **Random Forest (RF)**. We used the implementations provided by Scikit-learn [38] with default parameters for LR, KNN, and RF (200 estimators). For LGBM, we used the implementations provided by the `lightgbm` Python package with default parameters.

Next, we note that it is not enough to look at the individual predictors at the source locations: We must also account for the group identities within that data. In other words, we must account for the existence of random effects in this data. Hence, in addition to the baseline fixed-effects models, for each algorithm, we also train two mixed-effects models, one with the month identifier as the group and another with the **planetary boundary layer height (PBLH)** as the group. The month group accounts for the seasonality of PM2.5 pollutants. The PBLH group accounts for the effect of temperature inversion on the dispersion of air pollution. Since PBLH is a continuous height value, we discretize it into eight buckets. We used the `MERF` Python package, which uses an expectation-maximization algorithm, to train the mixed-effects models. To improve the robustness of our predictions, we trained each model using 10-fold cross-validation.

### 4.2 Evaluation Metric

We used **Mean Absolute Percentage Error (MAPE)** to evaluate our models. MAPE is a commonly used metric to measure the prediction accuracy of regression models. It represents the average of the absolute percentage errors made on each prediction. MAPE is a popular metric because it scales the error to percentage units, which makes it intuitive to interpret (lower MAPE is better). MAPE for a regression model is calculated as follows:

$$\text{MAPE} = \frac{100\%}{n} \cdot \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{\overline{y}} \right|,$$

where, $y_t$ is the actual value, $\hat{y}_t$ is the predicted value, $\overline{y}$ is the mean of actual values, and $n$ is the number of predictions. Note that this is actually a modified version of MAPE called weighted

Table 2. MAPE (in %) for the Four Regression Algorithms with Only Fixed-effects and Two Kinds of Mixed-effects (Month and PBLH)

| | Fixed Effects | | | | | | | Mixed Effects (Group: Months) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ni** | **Su** | **OC** | **SOA** | **Du** | **SS** | **BC** | **Ni** | **Su** | **OC** | **SOA** | **Du** | **SS** | **BC** |
| **LR** | 56.11 | 26.25 | 36.86 | 30.27 | 60.46 | 40.48 | 41.62 | 50.57 | 25.23 | 33.69 | 28.38 | 56.26 | 36.34 | 38.91 |
| **LGBM** | 33.86 | 15.87 | 23.33 | 19.35 | 27.77 | 19.66 | 27.61 | 33.01 | 15.72 | 22.99 | 19.04 | 27.35 | 18.89 | 27.07 |
| **KNN** | 24.93 | 9.94 | 17.39 | 12.44 | 13.57 | 9.73 | 20.62 | 25.21 | 10.0 | 17.49 | 12.52 | 13.92 | 9.72 | 20.73 |
| **RF** | 26.38 | 9.91 | 17.59 | 13.83 | 17.08 | 13.11 | 20.72 | 26.12 | 9.87 | 17.49 | 13.82 | 17.39 | 12.85 | 20.55 |
| | Mixed Effects (Group: PBLH) | | | | | | | | | | | | | |
| | **Ni** | **Su** | **OC** | **SOA** | **Du** | **SS** | **BC** | | | | | | | |
| **LR** | 46.95 | 26.18 | 28.0 | 26.82 | 59.33 | 40.19 | 33.41 | | | | | | | |
| **LGBM** | 31.74 | 15.81 | 20.33 | 18.09 | 27.71 | 19.72 | 24.88 | | | | | | | |
| **KNN** | 22.89 | 9.87 | 15.56 | 11.99 | 17.5 | 11.04 | 18.92 | | | | | | | |
| **RF** | 23.34 | 9.89 | 15.18 | 12.63 | 18.24 | 13.57 | 18.65 | | | | | | | |

We use KNN for the rest of our experiments. For KNN, the better-performing mixed-effects group for each constituent is highlighted in green. KNN's MAPE for the "overall" PM2.5 (i.e., sum across all seven constituents) is 14.74%. The variance of the chosen models (green cells) in %: Ni: 8.54, Su: 4.16, OC: 2.88, SOA: 2.86, Du: 59.32, SS: 16.63, BC: 4.11.

MAPE. The original MAPE formula uses $y_t$ (the actual value) in the denominator, whereas we use $\overline{y}$ (mean of actual values) in the denominator. This avoids a division by zero error if $y_t = 0$ for some $t$.

## 5 RESULTS

In this section, we present the findings from our analysis. While PollutionMapper shows promising results, we emphasize that it is not designed to be a PM2.5 forecast model but a source identification model. Hence, the validation results shown in the first three subsections below are only intended to establish the legitimacy of the model. With a strong model, we then demonstrate in the last subsection how our model can be used to identify pollution sources and therefore affect policy.

### 5.1 Comparison with Fixed-effects Model

Table 2 shows the MAPE for the different regression algorithms with **fixed-effects (FE)** and two kinds of mixed-effects (ME-PBLH and ME-Month). First, we note that linear regression generally performs the worst, and random forest performs the best (lowest MAPE) for both FE and ME models. However, random forest is computationally very expensive and infeasible to train with such a large feature set. Hence, we use the KNN model for the rest of our experiments, whose MAPE is comparable to that of random forest, and sometimes even lower.

We found that the ME models generally outperform the FE model, validating our hypothesis of the existence of mixed effects in the concentration of PM2.5 constituents. Further, we observe that the ME-PBLH model outperforms the ME-Month model for all constituents except $PM_{SS}$ and $PM_{Du}$. This is due to the seasonal nature of sea salt and dust, which are blown into mainland India with the monsoon winds. For other constituents, the effect of PBLH is more pronounced than the effect of the month. This shows the importance of accounting for temperature inversions when designing policy to combat air pollution caused by $PM_{Su}$, $PM_{Ni}$, $PM_{OC}$, $PM_{SOA}$, and $PM_{BC}$.

The caption of Table 2 also shows the variance of absolute percentage errors of the chosen models for each constituent. We see a small variance in all constituents except $PM_{SS}$ and $PM_{Du}$. Unlike other constituents that have shorter trend cycles (e.g., daily, monthly) these two constituents have a seasonal nature. This makes it hard for the model to identify trends and therefore results in a higher variance.

Table 3 shows the p-value and effect size of the improvement in MAPE from FE to ME models. Due to the aforementioned seasonal nature of $PM_{SS}$ and $PM_{Du}$, we compare the FE model with the ME-Month model for these two constituents, and with the ME-PBLH model for all the other

Table 3. P-value and Effect Size of Improvement in MAPE across FE and ME Models

| | p-value | | | | | | | Cohen's d | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ni** | **Su** | **OC** | **SOA** | **Du** | **SS** | **BC** | **Ni** | **Su** | **OC** | **SOA** | **Du** | **SS** | **BC** |
| **LR** | 0.0 | 0.168 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.07 | 0.67 | 8.11 | 5.61 | 3.34 | 3.72 | 14.49 |
| **LGBM** | 0.0 | 0.236 | 0.0 | 0.0 | 0.050 | 0.0 | 0.0 | 2.96 | 0.15 | 7.53 | 3.14 | 0.50 | 0.90 | 6.38 |
| **KNN** | 0.0 | 0.037 | 0.0 | 0.0 | 0.008 | 0.613 | 0.0 | 3.36 | 0.13 | 5.83 | 2.18 | −0.24 | 0.03 | 3.84 |
| **RF** | 0.0 | 0.347 | 0.0 | 0.0 | 0.110 | 0.019 | 0.0 | 4.37 | 0.11 | 7.01 | 3.17 | −0.20 | 0.64 | 6.02 |

For Sea Salt and Dust, we compare the MAPE of the FE model against that of the ME model with month group. For the rest of the pollutants, we compare the MAPE of the FE model against that of the ME model with PBLH group.

constituents. We see that the improvement in model performance from FE to ME models is statistically significant and the effect size is also large for all constituents except $PM_{SS}$ and $PM_{Du}$. For these two constituents, the FE model outperforms the ME-Month model, indicating that a FE model would be enough to predict their concentration at the sink.

### 5.2 Comparison against Measured Data

We now compare our model's predictions against the PM2.5 data recorded by the **Central Pollution and Control Board (CPCB)**, the apex government authority responsible for pollution control in India. For this comparison, we downloaded data from 35 CPCB sensors spread all over Delhi for a period of two years (2019–2020). We compare our predictions against the mean of these 35 sensors. However, CPCB sensors report only the "overall" PM2.5 concentration, not the seven constituents individually. Hence, we sum the predictions for the individual constituents to get our "overall" PM2.5 prediction. We then compare the mean PM2.5 of the 35 CPCB sensors against this sum.

When compared against CPCB data, the MAPE reported by our model is 56.78%. While the error may seem like a lot, it is essential to note that CPCB sensors are on-ground and hence measure local pollution values affected by local traffic, small fires, and so on. However, GEOS-CF forecasts global pollution values, which renders this comparison unfair. This explanation is supported by the fact that the MAPE between actual GEOS-CF values (not our predictions) and CPCB sensors *itself* is 58.98%! Yet, we have presented this result for transparency and completeness. We re-emphasize we are not trying to forecast PM2.5 in Delhi. There are better forecast models available, but they do not identify specific pollution sources, which is the main contribution of our article.

### 5.3 Group-level PM2.5 Constituent Distribution

Figure 9 plots all the predictions from the ME-Month model for each constituent grouped by month. In the plot, circles represent the mean, and error bars indicate ± 1 standard deviation. For most constituents ($PM_{Ni}$, $PM_{OC}$, $PM_{SOA}$, and $PM_{BC}$), we see a clear trend where their concentration in Delhi increases in the winter months (October, November, and December). This is due to the effect of temperature inversions in the winter months, which traps pollutants closer to the earth's surface, not allowing them to diffuse into the upper atmosphere.

The trend for $PM_{Su}$ is unexpectedly different. Their concentration rises sharply in June and slightly in the months of January and September. Upon investigating this behavior, we suspect that this is due to the higher electricity consumption in these months. June is extremely hot with temperatures going up to 39 °C (102 °F); the onset of the monsoon makes the weather hot and humid, causing people to use air coolers and air conditioners for long durations. September is the withdrawal of the monsoon causing very high humidity, and January is the peak of winter in Delhi. These conditions cause an increase in electricity demand during these months, which requires the thermal power plants around Delhi to function at a higher capacity. This hypothesis is further supported by power production data from the National Power Portal of India. The mean

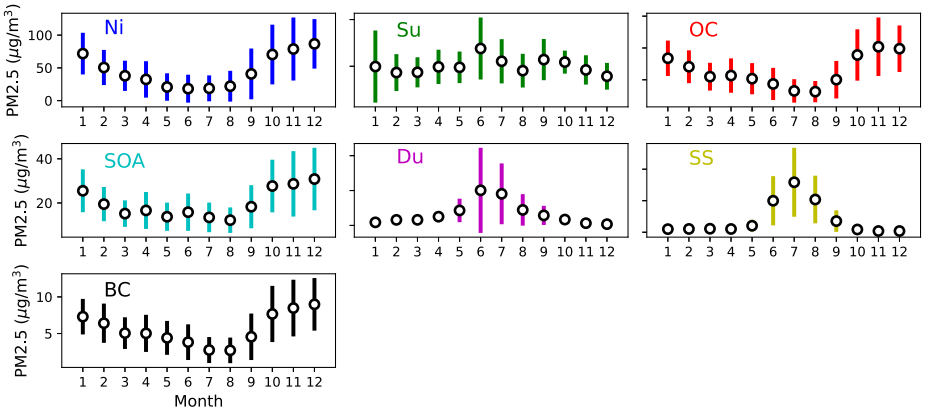



Fig. 9. Predictions from ME model (Group: Month).

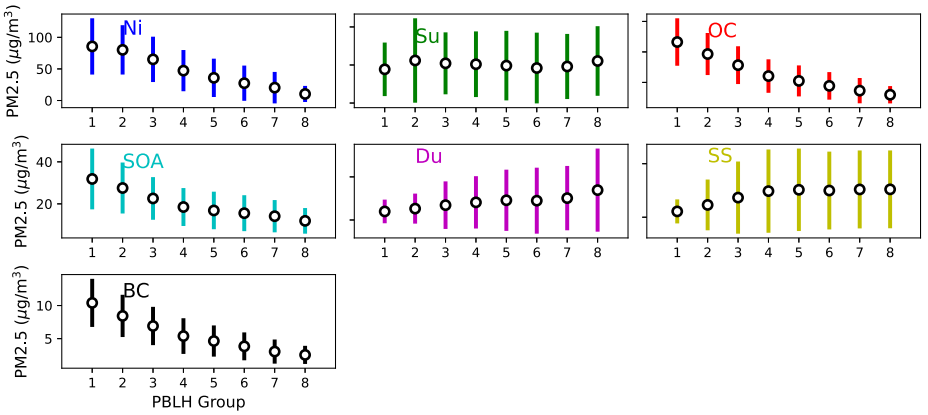



Fig. 10. Predictions from ME model (Group: PBLH).

production by power plants around Delhi across three years is 3.27 **mega units per day (MU/day)**, but the mean in June is 4.41 MU/day, in September is 3.96 MU/day, and in January is 2.82 MU/day. While the production in January is lesser than the yearly mean, the concentration of $PM_{Su}$ in January remains high due to a lower PBLH in the winter.

For $PM_{Du}$ and $PM_{SS}$, we see a spike in June, July, and August, which are the main monsoon months in India. The spike is because monsoon winds originate from the Arabian Sea in the south-west and cross the Thar Desert on their way to Delhi, therefby carrying dust into Delhi. Similarly, sea salt is carried from the oceans to Delhi by monsoon winds.

These sets of results show that our ME-Month model and its feature set are able to capture the seasonal variations in PM2.5 constituents such as sulphates, dust, and sea salt in Delhi. Interestingly, the month group is also able to capture the effects of temperature inversion, likely due to the correlation between the month of the year and the PBLH.

Figure 10 shows all the predictions from the ME-PBLH model for each constituent grouped by the eight PBLH discretizations. For most constituents ($PM_{Ni}$, $PM_{OC}$, $PM_{SOA}$, and $PM_{BC}$), the concentration in Delhi is higher when the PBLH is lower, due to the envelope effect of a low PBLH. $PM_{Du}$ and $PM_{SS}$ show an opposite trend where their concentration increases when the PBLH is higher. This is because dust and sea salt are seasonal PM2.5 constituents that arrive in

Delhi during June–September when the temperature is warmer and the PBLH is higher. Predictions from the ME-PBLH model show no trend for $PM_{Su}$. This may be because, as described above, due to increased electricity production in summer and monsoon months, the concentration of $PM_{Su}$ increases, which cancels out the effect of a higher PBLH. This nature of $PM_{Su}$ is captured by the ME-Month model, which explains why the MAPE for $PM_{Su}$ does not improve much from the ME-Month to the ME-PBLH models.

### 5.4 Feature Importance: Identifying Major Pollution Sources

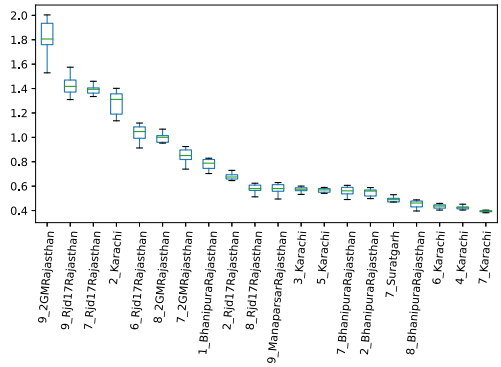
It is important to not just predict the air quality in Delhi but to also know what are the major sources responsible for this pollution. To find this, we performed a feature importance analysis of our model, PollutionMapper. We used the SHAP Python library's "KernelExplainer" class to conduct our feature importance analysis. Acknowledging the seasonal nature of PM2.5 constituents as found above, we conducted this analysis for different seasons, focusing on seasons in which Delhi sees the most air pollution. Figure 11 shows the SHAP values of the top 20 features and is explained below.

*5.4.1 Dust in June.* Figure 11(a) shows the feature importance for predicting the concentration of $PM_{Du}$ in Delhi during the onset of the monsoon winds (June). We observe that *all* the top 20 most important features are in the Thar Desert (Rajasthan) or in Karachi (Pakistan), both of which are very dusty areas with dry climates. Both these areas are also to the southwest of Delhi, the same direction from which the strong monsoon winds blow in June. Further, there are other sites in our feature set that are southwest of Delhi but are not as dusty (e.g., Mundra in Gujarat) and hence do not appear in the top 20 most important features. Conversely, there are dusty sites (e.g., Jaipur in Rajasthan) that do not appear, as they are not in the path of the dominant direction of the monsoon winds. These results show that incorporating the wind dispersion modeling using HYSPLIT was useful to improve the predictive power of PollutionMapper.

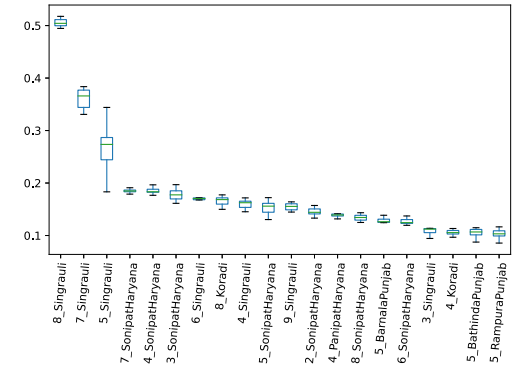
*5.4.2 Sulphates in Winter.* Figure 11(b) shows the feature importance for predicting $PM_{Su}$ in Delhi during the winter months. We see that 7 of the top 20 most important features are from Singrauli (in Jharkhand, central-eastern India). Singrauli and its surrounding region have many thermal power plants and other industries due to easy access to coal and minerals in that area. Interestingly, our model was able to identify Singrauli even though it is more than 700 km away from Delhi. Similarly, Koradi (in Maharashtra, central India) is also identified as an important feature even though it is more than 800 km away from Delhi. Koradi also has a major thermal power plant station. These results show that polluting sources even hundreds of kilometers away have an impact on the PM2.5 concentration of Delhi.

*5.4.3 Nitrates in Winter.* Figure 11(c) shows the feature importance for $PM_{Ni}$ in the winter season. We find that 10 of the top 20 features are from in and around Delhi itself, suggesting that Delhi is responsible for a majority of its own concentration of nitrates. This indicates that local policies such as the Odd-Even rule, suspension of construction activities, and so on, would be useful in protecting the air quality of Delhi in this season. Surprisingly, there are six features from Behat (in district Saharanpur, Uttar Pradesh), which is 170 km north of Delhi. On digging further, we found that this region suffers from chronic power cuts and is a major agricultural belt, due to which there is high usage of diesel generators in the area. There are also paper industries in the area that release NOx (e.g., nitric oxide and nitrogen dioxide) into the atmosphere, which oxidizes to form $PM_{Ni}$.

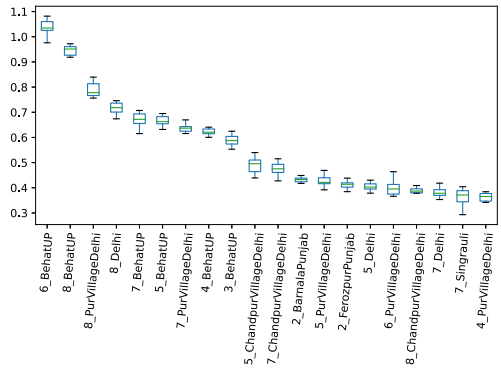
*5.4.4 Effect of Lingering Secondary Pollutants.* We observe in Figure 11(b) that 8 of the top 10 features for $PM_{Su}$ represent pollutants that have traveled to Delhi over a long duration (five or



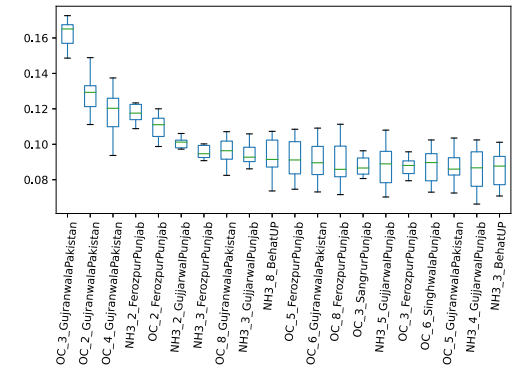

(a) Dust ($PM_{Du}$) in June



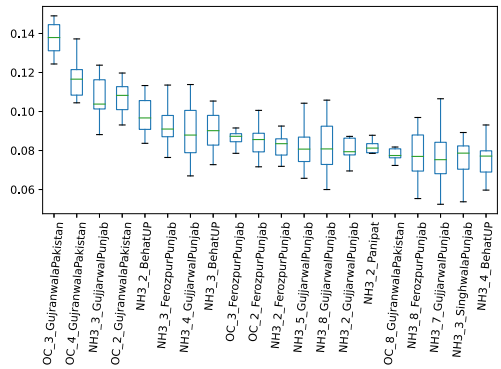

(b) Sulphates ($PM_{Su}$) in winter



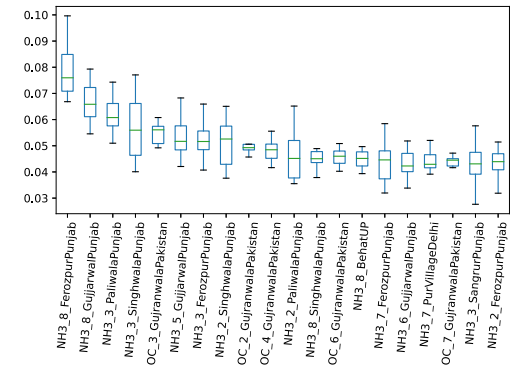

(d) Nitrates ($PM_{Ni}$) in winter




(d) SOA ($PM_{SOA}$) in winter




(e) SOA ($PM_{SOA}$) in October (end of Kharif)




(f) SOA ($PM_{SOA}$) in May (end of Rabi)

Fig. 11. Feature importance (SHAP values) for different PM2.5 constituents in different seasons. Winter is defined as November, December, and January. "Kharif" and "Rabi" are the two major cropping seasons in India. Kharif lasts from July to October, and Rabi from October to March. The beginning of each season is preceded by crop residue burning to clear the land after harvesting the crop from the previous season.

more days), even from sources that are close to Delhi (such as Sonipat, Haryana). This means that pollutants that took $\geq 5$ days to reach Delhi were more responsible for the $PM_{Su}$ concentration in Delhi than the pollutants that took fewer days. This shows that the longer $SO_2$ remained in the atmosphere after being released from the source, the more harmful it became by turning into $PM_{Su}$.

This observation is also true for $PM_{Ni}$ in Figure 11(c), where also eight of the top 10 features represent long travel times. These results show that PollutionMapper is able to capture the chemical reactions taking place in the atmosphere.

*5.4.5 Secondary Organic Aerosols after Crop Harvesting.* Figures 11(e) and (f) show the feature importance of predicting $PM_{SOA}$ in October and May, the end of the two major crop harvesting seasons in India. We chose to analyze feature importance in these seasons because crop harvests are followed by CRB to clear the land for the next sowing season, leading to intense pollution episodes in Delhi. In both these plots, we see that an overwhelming majority of the top 20 features come from the agriculture-heavy state of Punjab (both in the Indian and Pakistani sides[9]), where crop residue is burned during these months.

Recall that $PM_{SOA}$ is formed by the combination of a reacting gas like ammonia ($NH_3$) and organic carbon ($PM_{OC}$). Interestingly, we observe that features from India in both plots are almost all ammonia and rarely $PM_{OC}$. However, features from Pakistan are exclusively $PM_{OC}$ and *never* ammonia. This is hardly a coincidence and is likely because the Indian government provides significant farm subsidies on ammonia-rich fertilizers as compared to Pakistan. The overuse of these fertilizers releases ammonia, which reacts with other organic matter to produce $PM_{SOA}$.

## 6 RELATED WORK

Research around air pollution in computer science and related communities has revolved around two broad themes: sensing and modeling. Sensing methods are aimed at collecting air pollution data using static or dynamic sensor deployments. On the sensing side, past literature has focused on several research problems that arise during the deployment of such sensor networks. These problems include sensor placement [3, 37], calibration [14, 41], crowd-sourcing [1, 2, 30], and so on. The data collected from such sensor networks are then used to model and predict air pollution (e.g., References [4, 10, 11, 24]). Our work focuses on this latter thread of research, which we summarize below.

### 6.1 Data-driven Pollution Modeling

Air pollution predictions are important for both government policy-making (e.g., designing policies for traffic control, banning construction activities) and everyday human decisions (e.g., whether to exercise or not) [33]. As a result, modeling air pollution has been an important research endeavor in the environmental sciences. Prior literature classifies air pollution models as one of three types [10, 30, 33]: deterministic, deep-learning based, and statistical.

Deterministic models predict air pollution based on the first principles of atmospheric chemistry. They use physical equations to model the emission, transport, diffusion, and chemistry of air pollution [30, 33]. These models, such as CMAQ [11], are useful to understand the first principles of air pollution. However, they have to account for many factors and therefore tend to be complex, incomplete, inaccurate, and computationally heavy [10, 33] for prediction.

Recently, deep-learning-based methods are being utilized for air-quality prediction. These methods have been found useful to exploit the spatial and temporal correlations in the features using **convolutional neural networks (CNNs)** [16] and gated **deep neural networks (DNNs)** [33]. **Recurrent neural networks (RNNs)** with **long-short-term memory (LSTM)** units have also been popular to better utilize historical data in forecasting future air quality [10, 18, 24]. However, because deep-learning models are hard to interpret, such models are useful for air-quality

[9]The state of Punjab was divided into its Indian and Pakistani sides during the partition of the Indian subcontinent in 1947.

forecasts, but not so much for source identification. However, our aim is to identify pollution sources so policymakers can design directed policies to counter those sources.

Our work lies in the realm of statistical models. Statistical models use parametric or non-parametric statistical or classical machine learning methods to predict air pollution. Their predictions are often based on observed measurements such as remote sensing datasets [5], weather conditions, historical air pollution [6], topological features, and so on [30, 33]. Such models have been popular due to the availability of such data and their interpretability. However, many of these works examine the impact of a single source of pollution. For example, a rich body of work has quantified the impact of crop burning and biomass burning on the air quality of Delhi (e.g., References [4, 9, 15, 32]). Other work considers multiple sources but models air pollution *within* the city [24, 25]. Instead, we consider multiple sources such as biomass burning, thermal power plants, steel plants, and human activity from all over India. This is similar to source apportionment, which we discuss next.

### 6.2 Source Apportionment Studies

As explained in Guttikunda's excellent primer [21] on **source apportionment (SA)**, such studies (reviewed in References [45, 50]) focus on isolating the sources of pollution in a city. SA studies can be classified as "top-down" or "bottom-up" [21]. The top-down approach involves collecting air samples and analyzing them in a laboratory to assess the contributions of different sources. The bottom-up approach, which our study also follows, utilizes existing data to estimate pollution sources.

However, most bottom-up SA studies in Delhi consider sources within or around the city [23, 46]. For example, Guttikunda et al. [23] estimate the impact of various sources such as vehicular emissions, domestic activities, and power plants, but they only consider intra-urban sources, whereas we model a multitude of pollution sources across the country. Other apportionment studies [20, 43, 44, 50] focus on the *chemical* characterization of PM2.5. Such studies reveal generic pollution sources such as biomass burning or industries but do not identify which farmlands or factories specifically are responsible, thereby impeding targeted policy design.

Ghosh et al. [19] present the closest work to ours in identifying subcontinental and regional sources affecting air quality in Delhi. The main differences between our works lie in the methods. First, they use HYSPLIT in the trajectory mode rather than the dispersion mode, which causes information to be lost about the travel of particles from the source to the sink. Second, they present an "analysis" that is useful to identify past sources but cannot be generalized to test the efficacy of future policy interventions. In comparison, PollutionMapper allows a what-if analysis of policy interventions by tweaking relevant features and measuring their impact on PM2.5 at the sink.

## 7 CONCLUSION AND LIMITATIONS

Air pollution is a major public health hazard responsible for 4–10 million deaths globally each year. Given ineffective past interventions, there is a need for data-driven ways of designing policy interventions. In this article, we presented PollutionMapper, a source-identification model that captures the geographical, chemical, and meteorological factors affecting PM2.5 concentration at the sink location. We focused our analysis on PM2.5 in Delhi but with little effort, this model is generalizable to other cities in India and around the world. The model tracked the release of pollutants from their source to the sink using the HYSPLIT wind dispersion model and predicted the concentration of the seven PM2.5 constituents at the sink with low MAPE. Instead of a standard fixed-effects model, we used a mixed-effects model that was able to capture the seasonality of the

various constituents of PM2.5. A feature importance analysis uncovered the impact of far-away polluting sources on the PM2.5 concentration in Delhi.

Policymakers can use such a model to test the efficacy of policy interventions before spending time, effort, and public money in deploying the interventions. For example, before implementing the Odd-Even traffic rule in Delhi, policymakers can test the expected effect of this intervention on the air quality of the city. They can use PollutionMapper and artificially lower the concentration of $NO_2$ in the feature set of $PM_{Ni}$ model and then run the model to predict the $PM_{Ni}$ in Delhi. If there is no significant change, then that would mean that the intervention is ineffective.

**Limitations and Future Work:** First, this work focuses on modeling the global causes of air pollution that affect the entire city. However, it is well-known that PM2.5 concentration is a local phenomenon that changes every few meters. Modeling hyper-local factors is not the focus of this work. In future work, we aim to build a similar model to identify to hyper-local pollution sources that affect the air quality within a small locality, such as garbage dumps and small fires. Second, this article presents a comprehensive analysis of our model and its predictions. Despite our best attempts, it is possible that the model may be picking up spurious correlations, and we are working towards conducting a causal analysis of PollutionMapper.

## ACKNOWLEDGMENTS

We would like to thank Amit Sharma for his help with the feature importance analysis and Tanuja Ganu for her insightful suggestions during the analysis phase. We would also like to acknowledge the open datasets made available by NOAA and NASA, without which such an analysis would not have been possible.

## REFERENCES

[1] Ismi Abidi, Sagar Ravi Gaddam, Saswat Kumar Pujari, Chinmay Shirish Degwekar, and Rijurekha Sen. 2022. Complexity of factor analysis for particulate matter (PM) data: A measurement based case study in Delhi-NCR. In *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS'22)*. Association for Computing Machinery, New York, NY, 45–56. DOI: https://doi.org/10.1145/3530190.3534808

[2] Dhruv Agarwal, Srishti Agarwal, Vidur Singh, Rohita Kochupillai, Rosemary Pierce-Messick, Srinivasan Iyengar, and Mohit Jain. 2021. Understanding driver-passenger interactions in vehicular crowdsensing. *Proc. ACM Hum.-comput. Interact.* 5, CSCW2, Article 482 (Oct. 2021), 24 pages. DOI: https://doi.org/10.1145/3479869

[3] Dhruv Agarwal, Srinivasan Iyengar, Manohar Swaminathan, Eash Sharma, Ashish Raj, and Aadithya Hatwar. 2020. Modulo: Drive-by sensing at city-scale on the cheap. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS'20)*. Association for Computing Machinery, New York, NY, 187–197. DOI: https://doi.org/10.1145/3378393.3402275

[4] Meghna Agarwala and Abhinav Chandel. 2020. Temporal role of crop residue burning (CRB) in Delhi's air pollution. *Environ. Res. Lett.* 15, 11 (Oct. 2020), 114020. DOI: https://doi.org/10.1088/1748-9326/abb854

[5] Pinhas Alpert, Olga Shvainshtein, and Pavel Kishcha. 2012. AOD trends over megacities based on space monitoring using MODIS and MISR. *Am. J. Clim. Change* 01, 03 (2012), 117–131. DOI: https://doi.org/10.4236/ajcc.2012.13010

[6] Joshua S. Apte, Kyle P. Messier, Shahzad Gani, Michael Brauer, Thomas W. Kirchstetter, Melissa M. Lunden, Julian D. Marshall, Christopher J. Portier, Roel C. H. Vermeulen, and Steven P. Hamburg. 2017. High-resolution air pollution mapping with Google Street View cars: Exploiting big data. *Environ. Sci. Technol.* 51, 12 (June 2017), 6999–7008. DOI: https://doi.org/10.1021/acs.est.7b00891

[7] BBC. 2021. Air Pollution May Reduce Life Expectancy of Indians by Nine Years, Says Study. Retrieved from https://www.bbc.com/news/world-asia-india-58405479

[8] Aparajita Bhattacharya. 2016. The Data Is Unambiguous: The Odd-even Policy Failed to Lower Pollution in Delhi. Brookings. Retrieved from https://www.brookings.edu/opinions/the-data-is-unambiguous-the-odd-even-policy-failed-to-lower-pollution-in-delhi/

[9] Srinivas Bikkina, August Andersson, Elena N. Kirillova, Henry Holmstrand, Suresh Tiwari, A. K. Srivastava, D. S. Bisht, and Örjan Gustafsson. 2019. Air quality in megacity Delhi affected by countryside biomass burning. *Nat. Sustain.* 2, 3 (Feb. 2019), 200–205. DOI: https://doi.org/10.1038/s41893-019-0219-0

[10] Tien-Cuong Bui, Van-Duc Le, and Sang-Kyun Cha. 2018. A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM. arXiv:cs.LG/1804.07891

[11] Daewon Byun and Kenneth L. Schere. 2006. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Appl. Mechan. Rev.* 59, 2 (03 2006), 51–77. DOI: https://doi.org/10.1115/1.2128636

[12] Center For International Earth Science Information Network-CIESIN-Columbia University. 2018. Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11. DOI: https://doi.org/10.7927/H4JW8BX5

[13] Ritika Chaturvedi. 2021. Gautam Gambhir Installs 3 Giant Air Purifiers in Delhi, but Experts Call It "Unscientific." Retrieved from https://theprint.in/environment/gautam-gambhir-installs-3-giant-air-purifiers-in-delhi-but-experts-call-it-unscientific/550732/

[14] Sukwon Choi, Nakyoung Kim, Hojung Cha, and Rhan Ha. 2009. Micro sensor node for air pollutant monitoring: Hardware and software issues. *Sensors* 9, 10 (2009), 7970–7987. DOI: https://doi.org/10.3390/s91007970

[15] Daniel H. Cusworth, Loretta J. Mickley, Melissa P. Sulprizio, Tianjia Liu, Miriam E. Marlier, Ruth S. DeFries, Sarath K. Guttikunda, and Pawan Gupta. 2018. Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India. *Environ. Res. Lett.* 13, 4 (Mar. 2018), 044018. DOI: https://doi.org/10.1088/1748-9326/aab303

[16] Ebrahim Eslami, Yunsoo Choi, Yannic Lops, and Alqamah Sayeed. 2019. A Real-time Hourly Ozone Prediction System using Deep Convolutional Neural Network. arXiv:physics.ao-ph/1901.11079

[17] Richard Fuller, Philip J. Landrigan, Kalpana Balakrishnan, Glynda Bathan, Stephan Bose-O'Reilly, Michael Brauer, Jack Caravanos, Tom Chiles, Aaron Cohen, Lilian Corra, Maureen Cropper, Greg Ferraro, Jill Hanna, David Hanrahan, Howard Hu, David Hunter, Gloria Janata, Rachael Kupka, Bruce Lanphear, Maureen Lichtveld, Keith Martin, Adetoun Mustapha, Ernesto Sanchez-Triana, Karti Sandilya, Laura Schaefli, Joseph Shaw, Jessica Seddon, William Suk, Martha María Téllez-Rojo, and Chonghuai Yan. 2022. Pollution and health: A progress update. *Lancet Planet. Health* 6, 6 (01 Jun. 2022), e535–e547. DOI: https://doi.org/10.1016/S2542-5196(22)00090-0

[18] Amir Ghaderi, Borhan M. Sanandaji, and Faezeh Ghaderi. 2017. Deep Forecast: Deep Learning-based Spatio-temporal Forecasting. arXiv:cs.LG/1707.08110

[19] Saikat Ghosh, Jhumoor Biswas, Sarath Guttikunda, Soma Roychowdhury, and Mugdha Nayak. 2014. An investigation of potential regional and local source regions affecting fine particulate matter concentrations in Delhi, India. *J. Air Waste Manag. Assoc.* 65, 2 (Nov. 2014), 218–231. DOI: https://doi.org/10.1080/10962247.2014.982772

[20] Hao Guo, Sri Harsha Kota, Shovan Kumar Sahu, Jianlin Hu, Qi Ying, Aifang Gao, and Hongliang Zhang. 2017. Source apportionment of PM2.5 in north India using source-oriented air quality models. *Environ. Pollut.* 231 (Dec. 2017), 426–436. DOI: https://doi.org/10.1016/j.envpol.2017.08.016

[21] Sarath Guttikunda and Puja Jawahar. 2004. A Primer on Source Apportionment. Retrieved from https://urbanemissions.info/wp-content/uploads/docs/What_is_Source_Apportionment.pdf Accessed: 8-4-2023.

[22] Sarath Guttikunda and Puja Jawahar. 2020. Can we vacuum our air pollution problem using smog towers? *Atmosphere* 11, 9 (Aug. 2020), 922. DOI: https://doi.org/10.3390/atmos11090922

[23] Sarath K. Guttikunda and Giuseppe Calori. 2013. A GIS based emissions inventory at 1 km × 1 km spatial resolution for air pollution analysis in Delhi, India. *Atmos. Environ.* 67 (2013), 101–111. DOI: https://doi.org/10.1016/j.atmosenv.2012.10.040

[24] Shiva R. Iyer, Ananth Balashankar, William H. Aeberhard, Sujoy Bhattacharyya, Giuditta Rusconi, Lejo Jose, Nita Soans, Anant Sudarshan, Rohini Pande, and Lakshminarayanan Subramanian. 2022. Modeling fine-grained spatio-temporal pollution maps with low-cost sensors. *npj Clim. Atmos. Sci.* 5, 1 (Oct. 2022). DOI: https://doi.org/10.1038/s41612-022-00293-z

[25] Michael Jerrett, Altaf Arain, Pavlos Kanaroglou, Bernardo Beckerman, Dimitri Potoglou, Talar Sahsuvaroglu, Jason Morrison, and Chris Giovis. 2004. A review and evaluation of intraurban air pollution exposure models. *J. Expos. Sci. Environ. Epidem.* 15, 2 (Aug. 2004), 185–204. DOI: https://doi.org/10.1038/sj.jea.7500388

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[27] Christoph A. Keller, K. Emma Knowland, Bryan N. Duncan, Junhua Liu, Daniel C. Anderson, Sampa Das, Robert A. Lucchesi, Elizabeth W. Lundgren, Julie M. Nicely, Eric Nielsen, Lesley E. Ott, Emily Saunders, Sarah A. Strode, Pamela A. Wales, Daniel J. Jacob, and Steven Pawson. 2021. Description of the NASA GEOS composition forecast modeling system GEOS-CF v1.0. *J. Adv. Model. Earth Syst.* 13, 4 (2021), e2020MS002413. DOI: https://doi.org/10.1029/2020MS002413 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002413

[28] Prashant Kumar, Sunil Gulia, Roy M. Harrison, and Mukesh Khare. 2017. The influence of odd–even car trial on fine and coarse particles in Delhi. *Environ. Pollut.* 225 (June 2017), 20–30. DOI: https://doi.org/10.1016/j.envpol.2017.03.017

[29] Philip J. Landrigan, Richard Fuller, Nereus J. R. Acosta, Olusoji Adeyi, Robert Arnold, Niladri Nil Basu, Abdoulaye Bibi Baldé, Roberto Bertollini, Stephan Bose-O'Reilly, Jo Ivey Boufford, Patrick N. Breysse, Thomas Chiles, Chulabhorn Mahidol, Awa M. Coll-Seck, Maureen L. Cropper, Julius Fobil, Valentin Fuster, Michael Greenstone, Andy Haines, David Hanrahan, David Hunter, Mukesh Khare, Alan Krupnick, Bruce Lanphear, Bindu Lohani, Keith Martin, Karen V. Mathiasen, Maureen A. McTeer, Christopher J. L. Murray, Johanita D. Ndahimananjara, Frederica Perera, Janez Potočnik, Alexander S. Preker, Jairam Ramesh, Johan Rockström, Carlos Salinas, Leona D. Samson, Karti Sandilya, Peter D. Sly, Kirk R. Smith, Achim Steiner, Richard B. Stewart, William A. Suk, Onno C. P. van Schayck, Gautam N. Yadama, Kandeh Yumkella, and Ma Zhong. 2018. The Lancet commission on pollution and health. *Lancet* 391, 10119 (Feb. 2018), 462–512.

[30] Jason Jingshi Li, Boi Faltings, Olga Saukh, David Hasenfratz, and Jan Beutel. 2012. Sensing the air we breathe: The opensense Zurich dataset. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, 323–325.

[31] C. Ling, L. Cui, and R. Li. 2023. Global impact of the COVID-19 lockdown on surface concentration and health risk of atmospheric benzene. *Atmos. Chem. Phys.* 23, 5 (2023), 3311–3324. DOI: https://doi.org/10.5194/acp-23-3311-2023

[32] Tianjia Liu, Miriam E. Marlier, Ruth S. DeFries, Daniel M. Westervelt, Karen R. Xia, Arlene M. Fiore, Loretta J. Mickley, Daniel H. Cusworth, and George Milly. 2018. Seasonal impact of regional outdoor biomass burning on air pollution in three Indian cities: Delhi, Bengaluru, and Pune. *Atmos. Environ.* 172 (2018), 83–92. DOI: https://doi.org/10.1016/j.atmosenv.2017.10.024

[33] Zhipeng Luo, Jianqiang Huang, Ke Hu, Xue Li, and Peng Zhang. 2019. AccuAir: Winning solution to air quality prediction for KDD Cup 2018. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. Association for Computing Machinery, New York, NY, 1842–1850. DOI: https://doi.org/10.1145/3292500.3330787

[34] Global Energy Monitor. 2022. Global Steel Plant Tracker. Retrieved from https://globalenergymonitor.org/projects/global-steel-plant-tracker/download-data/

[35] NDTV. 2021. India's First Smog Tower Inaugurated in Delhi. NDTV. Retrieved from https://www.ndtv.com/delhi-news/indias-first-smog-tower-inaugurated-in-delhi-2516569

[36] Government of India. 2023. National Power Portal. Retrieved from https://npp.gov.in/

[37] Kevin P. O'Keeffe, Amin Anjomshoaa, Steven H. Strogatz, Paolo Santi, and Carlo Ratti. 2019. Quantifying the sensing power of vehicle fleets. *Proc. Nat. Acad. Sci.* 116, 26 (June 2019), 12752–12757. DOI: https://doi.org/10.1073/pnas.1821667116

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.

[39] Shivangi Pradhan. 2021. Delhi's Smog Towers Are a Wasteful Use of Public Funds, as Air Pollution Has Been Politicised. Scroll.in. Retrieved from https://scroll.in/article/1005214/delhis-smog-towers-are-a-wasteful-use-of-public-funds-as-air-pollution-has-been-politicised

[40] Reuters. 2021. New Delhi Is World's Most Polluted Capital for Third Straight Year: IQAir Study. Retrieved from: https://indianexpress.com/article/india/new-delhi-is-worlds-most-polluted-capital-for-third-straight-year-iqair-study-7230892/

[41] Olga Saukh, David Hasenfratz, and Lothar Thiele. 2015. Reducing multi-hop calibration errors in large-scale mobile sensor networks. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks (IPSN'15)*. Association for Computing Machinery, New York, NY, 274–285. DOI: https://doi.org/10.1145/2737095.2737113

[42] Express News Service. 2021. Odd-Even Vehicle Rationing Scheme: History, Efficacy, Expert Views. Retrieved from: https://indianexpress.com/article/cities/delhi/odd-even-vehicle-rationing-scheme-history-efficacy-expert-views-8252090/

[43] S. K. Sharma and T. K. Mandal. 2017. Chemical composition of fine mode particulate matter (PM 2.5 ) in an urban area of Delhi, India and its source apportionment. *Urb. Clim.* 21 (Sept. 2017), 106–122. DOI: https://doi.org/10.1016/j.uclim.2017.05.009

[44] S. K. Sharma, T. K. Mandal, Srishti Jain, Saraswati, A. Sharma, and Mohit Saxena. 2016. Source apportionment of PM2.5 in Delhi, India using PMF model. *Bull. Environ. Contam. Toxicol.* 97, 2 (May 2016), 286–293. DOI: https://doi.org/10.1007/s00128-016-1836-1

[45] Nandita Singh, Vishnu Murari, Manish Kumar, S. C. Barman, and Tirthankar Banerjee. 2017. Fine particulates over south Asia: Review and meta-analysis of PM2.5 source apportionment through receptor model. *Environ. Pollut.* 223 (2017), 121–136. DOI: https://doi.org/10.1016/j.envpol.2016.12.071

[46] Anjali Srivastava, B. Sengupta, and S. A. Dutta. 2005. Source apportionment of ambient VOCs in Delhi city. *Sci. Total Environ.* 343, 1–3 (May 2005), 207–220. DOI: https://doi.org/10.1016/j.scitotenv.2004.10.008

[47] A. F. Stein, R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. D. Cohen, and F. Ngan. 2015. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Am. Meteorol. Soc.* 96, 12 (2015), 2059–2077. DOI: https://doi.org/10.1175/BAMS-D-14-00110.1

[48] Jianjun Wang, Meigen Zhang, Xiaolin Bai, Hongjian Tan, Sabrina Li, Jiping Liu, Rui Zhang, Mark A. Wolters, Xiuyuan Qin, Miming Zhang, Hongmei Lin, Yuenan Li, Jonathan Li, and Liqi Chen. 2017. Large-scale transport of PM2.5 in the lower troposphere during winter cold surges in China. *Sci. Rep.* 7, 1 (Oct. 2017). DOI: https://doi.org/10.1038/s41598-017-13217-2

[49] Jin Wu, Derek Watkins, Josh Williams, Shalini Venugopal Bhagat, Hari Kumar, Jeffrey Gettleman, Rumsey Taylor, Leslye Davis, and Karan Deep Singh. 2020. Who Gets to Breathe Clean Air in New Delhi? Retrieved from https://www.nytimes.com/interactive/2020/12/17/world/asia/india-pollution-inequality.html

[50] Shweta Yadav, Sachchida N. Tripathi, and Maheswar Rupakheti. 2022. Current status of source apportionment of ambient aerosols in India. *Atmos. Environ.* 274 (2022), 118987. DOI: https://doi.org/10.1016/j.atmosenv.2022.118987

Received 4 April 2023; revised 11 July 2023; accepted 30 July 2023