Enhancing Air Pollutant Tracking with ERA5 for Underserved Regions

Devaansh Gupta
devaansh@cs.ucla.edu

Michael Simon
mlsimon@cs.ucla.edu

Eshanika Ray
eshanika@ucla.edu

Abstract

Air pollution is a critical global health concern, particularly in densely populated countries like India. However, comprehensive air quality datasets for such regions are often temporally and spatially sparse, lack physical variables, and are not collated for use in machine learning models. This paper presents an approach to enhance air quality monitoring and prediction in underserved regions by integrating the ERA5 reanalysis dataset with India's Historical Ambient Air Quality (HAAQ) Data. We propose a data processing pipeline that standardizes the HAAQ dataset to a 5.625° grid system and combines it with ERA5 meteorological parameters, creating a more comprehensive dataset that spans from 1987 to 2014. The resulting dataset includes measurements for key pollutants (PM_{10}, SPM, NO_2, SO_2) along with essential meteorological variables such as temperature, humidity, wind components, and cloud cover. We evaluate the spatial and temporal characteristics of our enhanced dataset, addressing challenges such as missing values and varying measurement densities across different regions. Our work provides a standardized framework for integrating historical air quality data with meteorological parameters, potentially enabling more sophisticated air quality assessment and forecasting capabilities in regions with limited data reporting infrastructure. Our code can be found at `https://github.com/devaansh100/cs269/tree/main`.

1 Introduction

Particulate matter and air pollution is the leading cause of respiratory diseases among adults in the US [8]. This makes it imperative to monitor and predict air pollution levels to preemptive action. In recent years, air pollution has grown exponentially due to increasing industrialization and an exploding population with increasingly higher demands of energy [14]. Since air pollution is correlated with anthropogenic factors [18], predicting future levels can effectively be posed as a time-series forecasting problem for machine learning models. While air quality is most highly correlated with anthropogenic activity, prior approaches ignore physical atmospheric variables as features. We posit that these variables are also correlated with pollution; for instance, higher precipitation would reduce the amount of nitrous oxide and sulfur dioxide in the air, wind patterns would help eliminate suspended particulate matter, and higher temperatures would increase ozone concentration in the air [20]. A major reason why existing models are unable to use these additional variables is the lack of a comprehensive dataset combining these atmospheric chemistry variables with the physical ones. In this paper, we aim to propose such a dataset by combining air pollution datasets with ERA5.

For some of the most populous countries of the world, specifically India, air pollution readings are quite sparse. While they do exist, they are usually of a much lower temporal resolution with multiple missing values, are not collated and/or standardized, and again, do not contain physical variables [16]. While some datasets are prepared bu the government, they are often behind broken APIs, lack geographical coordinates, again, do not contain physical variables and are not standardised for

machine learning training and evaluation. In this work, we aim to mitigate these issues by providing an air pollutant prediction dataset for India, which is enhanced by physical vairables from ERA5.

2 Related Works

2.1 Machine Learning for Air Pollutant Prediction

Prediction of air pollution has been attempted with various models; [4, 3] use deep learning-based approaches, while [10, 2, 15, 9] use statistical models. While most models do not use other physical variables as features, [4] does leverage them. However, that's solely enabled by the associated weather station reporting additional physical variables, which is often not the case. [9] also makes use of physical variables during prediction, however, it uses an independently learned model to capture these correlations. On the contrary, we aim to create a dataset to facilitate the use of a single model for this task. A significant amount of research has been conducted on air pollution prediction in the United States. These works typically use subsets of the Air Quality System (AQS) dataset [17], which is a comprehensive repository maintained by the US Environmental Protection Agency. Such an extensive dataset is lacking for most underserved regions.

2.2 Air Pollution Datasets for India

The air pollution datasets for India are typically provided by government agencies but have certain limitations. The Central Pollution Control Board (CPCB) Dataset[16] provides state-wise and district-wise data for India from 2015 onwards. While it offers daily Air Quality Index (AQI) values, it lacks specific pollutant concentrations such as $PM_{2.5}$, PM_{10}, SO_2 and NO_2. The World Health Organization(WHO) dataset[19] also contains pollutant values for Indian cities, but does not extend to rural areas. Datasets covering the entire country are available, but have poor temporal resolution [5], or are estimated by numerical models[6]. For our work, we choose to combine the Historical Ambient Air Quality (HAAQ) Dataset[12], since it contains district wise pollutant values, assimilated daily - providing a sufficient temporal and spatial resolution.

2.3 Data Reporting for Underserved Countries

To address the limitations of collection of air pollution data in underserved countries, low-cost sensor technologies are being deployed to expand coverage in resource-constrained areas [1, 13]. NASA's GEOS-CF model combines satellite observations, modeling, and super computing to generate global air quality forecasts, helping to fill gaps between ground monitors [11]. Machine learning techniques are also being employed to improve the accuracy of these models and to extrapolate data to areas with limited monitoring [7].

3 Historical Ambient Air Quality (HAAQ) Data For India

We would work towards creating an enhanced air pollutant dataset for India. Details on the raw data source is provided in this section.

Statistic	**Value**
# Data Points (Readings)	259817
Time Range	1987 - 2014
Temporal Resolution	Daily (45% missing days)
# Cities/Towns	261
Measured Pollutants	**# Missing Values**
Nitrogen Dioxide (NO_2)	9789
Sulfur Dioxide (SO_2)	21009
Particulate Matter (RSPM/PM_{10})	13253
Suspended Particulate Matter (SPM/$PM_{2.5}$)	131895

Table 1: Dataset Statistics for the Station-level HAAQ Dataset.

This dataset contains measured district-wise data for India, provided by the Central Pollution Control Board, for approximately 28 years. For each district, data is collected by various weather stations, focusing on both metropolitan and non-metropolitan areas. A major limitation that has prevented wide scale usage of this data is the fact that the files have not been collated and standardized, with challenges like a broken API and lack of standardized coordinates for measurement locations.

We convert this dataset to a 5.625° grid system. This is followed by aggregating data from different weather stations within the same grid cell. A major shortcoming of the HAAQ dataset is the inconsistent sampling rate, with approximately 45% of days having missing data despite the intended daily measurement frequency. To circumvent this issue, we design our task to make predictions independent of the time step.

A summary of the variables present in the original HAAQ dataset, along with other statistics is provided in Table 1. Note that PM_{10} refers to particulate matter smaller than 10 micrometers (μm) in diameter, SPM is a broader metric for particulate matter of any size, and NO_2 and SO_2 are greenhouse gasses. However, in some measurement localities, the SPM data column is reported as a $PM_{2.5}$ value, making it ambiguous whether the SPM is referring to the standard definition, or the more stringent $PM_{2.5}$ which covers matter smaller than 2.5 μm in diameter. We list the # of missing entries to showcase the incompleteness of the HAAQ dataset.

4 Task Formulation

Let $\mathcal{G} = \{g_1, \ldots, g_N\}$ denote the set of geographical grid points at 5.625° resolution, where N is the total number of grid points. At each time step t and grid point g_i, we have, (i) input physical variables from ERA5, $\mathbf{X}_{t,g_i} \in \mathbb{R}^{Cin}$, where C_{in} is the number of ERA5 variables, and (ii) target air pollutant measurements, $\mathbf{y}_{t,g_i} \in \mathbb{R}^{Cout}$, where C_{out} is the number of pollutant variables.

Our prediction task can be formulated as learning a function $f(\cdot)$, parametrized by θ such that $f_\theta : \mathbb{R}^{C_{in}} \to \mathbb{R}^{C_{out}}$. For a given timestep (day) t and grid point g_i, the model predicts $\hat{\mathbf{y}}_{t,g_i} = f_\theta(\mathbf{X}_{t,g_i})$.

In other words, given the physical variables for a day and location, the model predicts the corresponding pollutant variables. While this is not a time-series prediction task, it can be reformulated as one, however, there would be additional challenges with respect to sparsity of the data (both, spatially and temporally) in that case. With the current formulation, f_θ can be chained with weather prediction models to predict pollutant values.

5 Data Processing Pipeline

HAAQ Dataset Processing The HAAQ Dataset required extensive preprocessing. First, we implemented a custom solution to hack through the broken API interface. Following this, we dropped values that were assimilated monthly to maintain daily granularity. Then, town-wise assimilation was performed by combining data from different weather stations within the same area. We then added coordinates for each town using GeoPy[1] and converted the absolute coordinates to a 5.625° grid system to match WeatherBenchV2's resolution. Finally, we index the data based on sampling date and ERA5 coordinates. The date format is parsed to a standard dd-mm-yyyy format.

ERA5 Dataset Integration The ERA5 dataset processing involved several key steps to ensure compatibility with HAAQ. We first aggregated the hourly readings to daily readings, to match HAAQ's temporal resolution. An inner join was performed with HAAQ, which eliminated various time steps and locations where air pollution data was not available. While all atmospheric variables can be combined with HAAQ, we show our results with a subset of these variables to curtail computational costs. Additionally, we subsampled ERA5 to a 5.625° grid. The processed dataset was then saved to disk in a csv format.

Train-Validation-Test Split The dataset was temporally divided to ensure proper evaluation, with the training set spanning 1987-2012, validation set covering 2013 and test set comprising 2014. Since our task is independent of the location, we could have randomly sampled all the locations as well. However, splitting on time steps is consistent with notion of future values prediction, motivating the method of chaining with weather forecasting models.

[1] https://pypi.org/project/geopy/

The processed dataset stats are reported in Table 2.

Variable	# Samples	# Missing	# Samples	# Missing	# Samples	# Missing
	Train		Validation		Test	
NO_2	41831	615	4570	17	4626	9
SO_2		1544		17		7
RSPM		1250		11		4
SPM		8903		3675		3549

Table 2: Dataset Statistics for the 5.625° gridded HAAQ Dataset, enhanced with ERA5. # Samples denotes the total number of entries for that variables, including missing values. For each entry, all corresponding phhysical variables from ERA5 are also present.

6 Baselines

Input Variables Due to computational constraints, we use the following physical variables; temperature, u-component of wind at 10m, v-component of wind at 10m, relative humidity, specific humidity, total precipitation and total cloud cover. For variables with values at multiple pressure levels, we choose the value at pressure-level = 1000hPa, which is the value near the ground.

Output Variables We only predict the values of NO_2, SO_2 and RSPM. While we provide the data for SPM, we do not use it in the model due to the large number of missing values.

Metric We report the baselines on the conventional RMSE metric.

Models We provide results on two baselines, (i) climatology and (ii) three-layer MLP. Climatology always predicts the value equal to the mean of the output variable in the training dataset. We use a shallow MLP to demonstrate the incorporation of physical variables improves the prediction of pollutant values. It consists of 75 learnable parameters, a skip connection and a relu nonlinearity after after each layer (except the last). It is trained for 10 epochs, with the Adam optimizer and a learning rate of $1e-2$. We also implement early stopping while training the network, choosing the model with the least validation RMSE. Training is performed with the L2 loss, with a batch size of 32.

Data transforms To fill in the missing values, we perform two transformations, (i) we assume the datapoints are uniformly spaced and interpolate between them, (ii) missing values at the start of the dataset are backfilled with first available observation. Additionally, to stabilise training, the input and output variables are normalised to a standard normal distribution, using the means and standard deviation from the training split.

Results The baseline results are shown in Table 3. The 3-layer MLP, with only 75 parameters, outperforms the climatology baseline on all metrics, except RMSE_{SO_2}. This is interesting, since it implies that the prediction of SO_2 is not very dependent on physical variables. We study this more in depth, in section 7.

Model	$\textbf{RMSE}_{NO_2}$	$\textbf{RMSE}_{SO_2}$	$\textbf{RMSE}_{RSPM}$	RMSE
Climatology	0.8675	**0.7408**	0.7943	0.8023
3-layer MLP	**0.8147**	0.7439	**0.7101**	**0.7552**

Table 3: Baseline results on the HAAQ dataset, augmented with ERA5. For RMSE, lower scores imply better performance. The best performing models are in **bold**.

7 Data Analysis

To better understand the spatial and temporal characteristics of our integrated dataset, we present several key visualizations that highlight different aspects of the data distribution and relationships between variables.

Data Coverage We plot the locations of the towns for which we provide data in Figure 1 - specifically, we plot the average number of daily readings for NO_2. Other variables follow similar trends. Notably, the provided dataset extensively covers all of India, including the North-East, which is often underrepresented in datasets.

Correlations between physical and chemistry variables We visualize the correlation matrix between selected physical atmospheric variables from ERA5 and the paired chemistry data from the HAAQ dataset in Figure 4. The weakest correlation between any of the chemistry variables and the physical variables from ERA5 is SO_2, which is consistent with the fact we were unable to outperform Climatology with a trained MLP. We hope such a treatment of variables better motivates our choice of creating a combined dataset. Future works can use such correlations insights to perform enhanced feature engineering for other underserved regions, where data is limited.

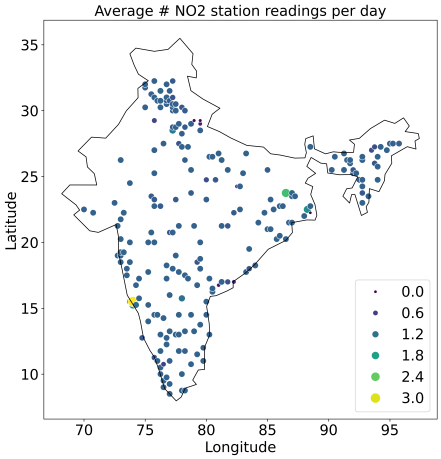

Figure 1: A map showing the average number of NO2 readings per day. Evidently, we provide data for most regions of the country, including the North-east, which is often underrepresented.

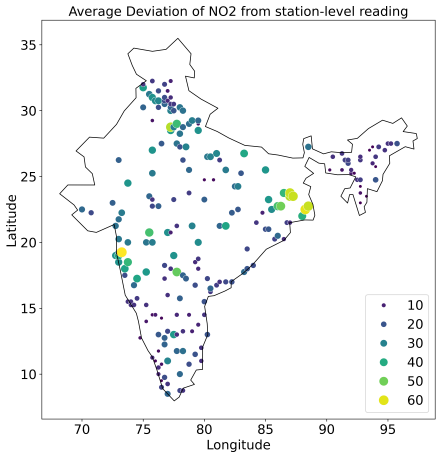

Figure 2: Average deviations between gridded and station level NO_2 values in the HAAQ dataset, for the 0.25° grid. The absence of a marker denotes no information loss.

Data loss due to gridding To determine the degree of data loss due to spatial aggregation by gridding, we calculate the average deviation in the grid values from the station values for NO_2 concentration across all time steps, for each geographic location in Figure 2. This is done for 0.25° grid system. As expected, we lose the most granularity near metropolitan areas with more stations - corresponding to Mumbai, Delhi and Kolkata. There is an assumption here, which assigns the same pollutant value to the entire grid cell, which would be more evident in lower resolutions. Other variables also show similar trends.

Variation of data across timesteps To visualize any seasonality to the chemical variables, we visualize one of the pollutants, NO_2 over time in Figure 3. We observe increased variance during the beginning of the sampling period in the years following 1987, followed by more regular seasonal cycles of NO_2 concentration. One potential explanation for this could be that data collection for air pollutants was catalyzed by heavy pollution in the late 1980s and early 1990s. The cyclical nature of the pollutant roughly aligns with increased A/C and energy usage during the summer months.

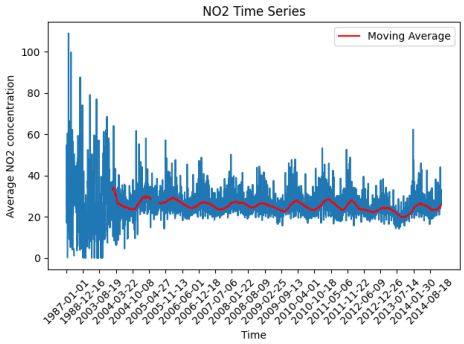

Figure 3: Average NO_2 concentration (PPM) across India over time

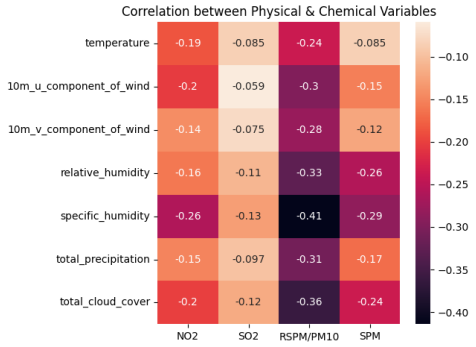

Figure 4: Pearson's Correlation Coefficient Matrix between Physical and Chemical variables

8 Conclusion

This paper presents a significant contribution to air quality monitoring and prediction in underserved regions through the creation of an integrated ERA5-HAAQ dataset. Our work addresses several critical gaps in existing air quality data for India by combining historical pollution measurements with comprehensive meteorological parameters from ERA5. The resulting dataset contains 51,027 data points spanning from 1987 to 2014, covering majority of the geographical area of India.

Our data processing pipeline demonstrates a practical approach to combining disparate datasets while maintaining data integrity and temporal consistency. This methodology can be extended to other underserved regions facing similar challenges in air quality monitoring. Furthermore, the standardized format of our dataset makes it readily accessible for machine learning applications, potentially enabling better pollution forecasting and policy decisions. Future work could focus on enahcing this data with additional variables, coming from sources like satellite data.

9 Limitations

While our dataset is comprehensive, it does have some limitations. The significant proportion of missing values across different pollutants requires careful consideration during model development. The varying density of measurement stations across different regions may also introduce spatial biases in the processed dataset. Additionally, as shown in Figure 2, converting station level data into the grid-level reduces its granularity causing data loss.

Another inherent limitation of our task is the assumption that the pollutant values only depend on the physical variables of a particular location. However, it is common knowledge that air pollution is heavily influenced by various other human factors as well. While it would be fair to argue that these human factors can manifest themselves in physical variables as well, a robust approach modeling this phenomenon is still unknown and would make an interesting direction for future works.

References

[1] AirGradient. Airgradient project in rwanda, 2024. Accessed November 10, 2024.

[2] U. A. Bhatti, Y. Yan, M. Zhou, S. Ali, A. Hussain, H. Qingsong, Z. Yu, and L. Yuan. Time series analysis and forecasting of air pollution particulate matter (pm 2.5): an sarima and factor analysis approach. *Ieee Access*, 9:41019–41031, 2021.

[3] J. Duan, Y. Gong, J. Luo, and Z. Zhao. Air-quality prediction based on the arima-cnn-lstm combination model optimized by dung beetle optimizer. *Scientific Reports*, 13(1):12127, 2023.

[4] B. S. Freeman, G. Taylor, B. Gharabaghi, and J. Thé. Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8):866–886, 2018.

[5] T. Ganguly, K. L. Selvaraj, and S. K. Guttikunda. National clean air programme (ncap) for indian cities: Review and outlook of clean air action plans. *Atmospheric Environment: X*, 8:100096, 2020.

[6] A. C. A. Group. Satellite-derived pm2.5 dataset. `https://sites.wustl.edu/acag/datasets/surface-pm2-5/#V5.GL.04`. Accessed: 2024-11-06.

[7] C.-Z. Huang, J.-H. Huang, Z.-R. Peng, and D.-X. Yue. An overview of machine learning and its applications in air quality forecasting. *Sustainability*, 13(9):4669, 2021.

[8] S. H. Jeong and S. Y. Kyung. Particulate-matter related respiratory diseases. *Tuberculosis and Respiratory Diseases*, 83(2):116–121, 2020.

[9] Y. Liu, L. Wen, Z. Lin, C. Xu, Y. Chen, and Y. Li. Air quality historical correlation model based on time series. *Scientific Reports*, 14(1):22791, 2024.

[10] E. Marinov, D. Petrova-Antonova, and S. Malinov. Time series forecasting of air quality: a case study of sofia city. *Atmosphere*, 13(5):788, 2022.

[11] NASA. Nasa tempo: Tropospheric emissions: Monitoring of pollution, 2024. Accessed November 10, 2024.

[12] C. P. C. B. of India. Historical daily ambient air quality. `https://www.data.gov.in/catalog/historical-daily-ambient-air-quality-data`. Accessed: 2024-11-06.

[13] A. Oluleye, G. Akinlabi, and K. Owoade. Low-cost air quality monitoring in african cities. *Environmental Monitoring and Assessment*, 195(2):1–15, 2023.

[14] A. Peters, D. W. Dockery, J. E. Muller, and J. Schwartz. Air pollution and population health: A global challenge. *Environmental Health Perspectives*, 116(12):1481–1486, 2008.

[15] M. S. Ramadan, A. Abuelgasim, and N. Al Hosani. Advancing air quality forecasting in abu dhabi, uae using time series models. *Frontiers in Environmental Science*, 12:1393878, 2024.

[16] A. Roychowdhury, A. Somvanshi, and S. Kaur. Status of air quality monitoring in india: Spatial spread, population coverage and data completeness. Technical report, Centre for Science and Environment, New Delhi, July 2023.

[17] US Environmental Protection Agency. Air quality system data mart. Internet database, 2024. Available at `https://www.epa.gov/outdoor-air-quality-data`.

[18] Y. Wang, Q. Ying, J. Hu, and H. Zhang. Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in china during 2013–2014. *Environment International*, 73:413–422, 2014.

[19] World Health Organization. Who global air quality database. Technical report, World Health Organization, 2024.

[20] Z. Zhang, T. Xue, and X. Jin. Impact of meteorological conditions on the covid-19 transmission: A systematic review and meta-analysis. *Science of The Total Environment*, 726:138870, 2020.