IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

AI-Based Air Quality Prediction and Satellite-Assisted Pollution Monitoring System for India

Kavin Raj S.B[1], Bhuvan MRJ[2], S Poornima[3]

Dept. of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus Chennai, India[1]

Dept. of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India[2]

Dept. of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus Chennai, India[3]

Abstract: India's official network of continuous ambient air quality monitoring stations is sparse and concentrated in large cities, and its readings reach the public through OpenAQ up to three days late. This paper presents a system that (i) estimates the daily mean concentrations of $PM_{2.5}$, PM_{10} and NO_2 for every 0.1° cell (about 11 km) of India, including cells with no monitor, and (ii) forecasts them one to three days ahead at monitored cells. Both outputs are converted into the CPCB National Air Quality Index. The estimator (Model A) is a LightGBM regressor whose inputs are Sentinel-5P trace-gas columns, satellite fire counts, meteorology and land-use layers computed in Google Earth Engine, plus an inverse-distance interpolation of monitor readings from three days earlier. The forecaster (Model B) is a LightGBM model with quantile bands, and an LSTM was evaluated as a challenger. Both are trained on 560,592 cleaned cell-days of Central Pollution Control Board measurements from July 2018 to September 2026. On whole cities held out from training, Model A reaches R^2 of 0.63, 0.63 and 0.23 for $PM_{2.5}$, PM_{10} and NO_2, and its 80% intervals cover 79–81% of held-out days. On a held-out year forecast with the weather forecasts that were actually issued, Model B improves on persistence by a skill score of +0.24 to +0.45 and outperforms the LSTM (mean skill +0.345 against +0.317). Removing every satellite input lowers R^2 by only 0.00–0.02, and we report this result plainly. A one-command daily refresh and an interactive web globe deliver the results for 594 Indian cities.

Keywords: air quality, $PM_{2.5}$, NO_2, machine learning, LightGBM, LSTM, Sentinel-5P, Google Earth Engine, AQI forecasting, spatial cross-validation, India

I. INTRODUCTION

Air pollution is among the largest health risks in India. The Global Burden of Disease study attributes 1.67 million deaths in India in 2019, or 17.8% of all deaths, to air pollution [1], and ambient particulate matter alone accounts for 0.98 million of them [1]. In ten Indian cities, each 10 µg/m^3 rise in daily $PM_{2.5}$ raises daily mortality by 1.42% [2]. The World Health Organization's 2021 guidelines lower the recommended annual $PM_{2.5}$ level to 5 µg/m^3 [3], far below what Indian monitors record.

Managing this burden requires knowing what people breathe, where they live, every day. The reference measurements in India come from the Central Pollution Control Board's (CPCB) continuous ambient air quality monitoring stations. These stations are expensive and concentrated in large cities [4]. Only 371 of the 30,113 0.1° cells that cover India, about 1.2%, contain a station with usable data. The readings also arrive late: OpenAQ [5], the open platform through which we access them, publishes CPCB's readings in batches up to three days after measurement. On the evening of 3 October 2026, the newest available measurement was from 30 September. For most of the country, therefore, nobody knows today's air quality, let alone tomorrow's.

The cost of this gap is clearest in an emergency. On 2 March 2023 a fire broke out at the Brahmapuram municipal waste dump near Kochi, Kerala, and burned for about twelve days; schools were closed and the Kerala High Court took up the matter on its own motion [6]. The highest $PM_{2.5}$, 679.7 µg/m^3 on 8 March, was recorded by a field campaign that began five days after the fire started [7]. The gap is also routine: across India, 316 of the 594 towns on our map, with about 75 million people, have no CPCB monitor of their own in the OpenAQ data.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

This paper describes a working system that addresses both gaps with machine learning, satellite observations and weather data, and reports its accuracy on tests that match how it is used. The estimates are converted into the official CPCB National Air Quality Index (AQI) [8], so that the output is directly interpretable. Our contributions are:

- A cleaned ground-truth table of 760,637 station-days from 2018 to 2026 that merges two copies of CPCB's records, links stations by their data rather than their published coordinates, and settles the unit of NO_2 readings day by day.
- **Model A**, an estimator of today's $PM_{2.5}$, PM_{10} and NO_2 at any cell in India, evaluated on whole cities it never saw, with intervals calibrated by distance to the nearest monitor.
- **Model B**, a 1–3 day forecaster for monitored cells, evaluated on an unseen year with the weather forecasts that were actually issued, and compared with persistence, climatology, ridge regression and an LSTM.
- A plain account of what each component contributes: the satellite inputs, live rather than reanalysis weather, and fresher measurements.
- An operational system: a daily refresh, an API and a web globe with estimates, forecasts, explanations and health guidance for 594 cities.

II. RELATED WORK

Health need and India-wide estimates. De Bont *et al.* linked daily 1-km $PM_{2.5}$ from a machine-learning model to 3.6 million deaths in ten Indian cities [2]. Its exposure model, by Mandal *et al.*, combines deep learning, random forests and gradient boosting on satellite aerosol, ERA5, land-use and fire inputs, and reaches a daily cross-validated R^2 of 0.86 with whole monitors held out [9]. Wang *et al.* reconstructed daily $PM_{2.5}$ and PM_{10} for India for 1980–2022 at 10 km with LightGBM trained on CPCB data, with R^2 0.70 at held-out sites [10]. Kawano *et al.* used Sentinel-5P NO_2 and CO in a two-stage model for daily 10-km $PM_{2.5}$ over India; their spatial cross-validated R^2 is 0.67, against 0.85 for a random split [11].

Satellite trace gases and machine learning. Kang *et al.* found that adding Sentinel-5P NO_2, CO and the aerosol index to an XGBoost estimator of $PM_{2.5}$ reduced its mean absolute error by about 14% [12]. Griffin *et al.* estimated surface NO_2 from TROPOMI columns with a random forest, with a daily R^2 of 0.79 against 0.38 for traditional profile scaling [13]. Mamić *et al.* predicted $PM_{2.5}$ and PM_{10} from Sentinel-5P gases, GFS weather and terrain retrieved through Google Earth Engine [14], the same open stack used here. Gradient-boosted decision trees, in particular LightGBM [15], handle missing values natively and train quickly on large tables.

Forecasting and validation. Méndez *et al.* survey 155 machine-learning air-quality forecasting studies [16]. Bernacki and Scherer find that no single model dominates, that tree ensembles are strong on structured data, and that generalisation to new cities and uncertainty quantification remain open [17]. Long short-term memory (LSTM) networks [18] are the usual deep-learning baseline, and we test one as a challenger. When data are spatially structured, random cross-validation lets a model test on places it has effectively seen; Roberts *et al.* recommend blocking by the structure instead [19], and Linnenbrink *et al.* show that random k-fold cross-validation fails for clustered samples [20]. Table I summarises these studies.

TABLE I. LITERATURE SURVEY: STUDIES CLOSEST TO THIS WORK

Ref.	Approach and data	Key result
[2]	ML 1-km $PM_{2.5}$ linked to mortality, ten Indian cities	+1.42% daily deaths per 10 µg/m^3
[9]	Ensemble of DL, RF and GBM; AOD, ERA5, land use, fires	Daily CV R^2 0.86, monitors held out
[10]	LightGBM; ERA5-Land, emissions, MERRA-2; CPCB labels	R^2 0.70 at held-out sites
[11]	Two-stage ML; MODIS AOD, Sentinel-5P NO_2, CO	Spatial R^2 0.67; random split 0.85
[12]	XGBoost + SHAP; Sentinel-5P gases, ERA5	Gases cut MAE by about 14%
[13]	Random forest on TROPOMI NO_2, ERA5	Daily R^2 0.79 vs 0.38 for scaling
[14]	Random forest; Sentinel-5P, GFS, terrain via Earth Engine	r up to 0.82 for $PM_{2.5}$

[20]	kNN distance-matching cross-validation	Random k-fold fails on clustered samples
[17]	Review of data-driven forecasters	New-city tests and uncertainty still open
This work	LightGBM on Sentinel-5P, fires, weather, land and 3-day-old monitors; 1–3 day forecaster	Held-out-city R^2 0.63/0.63/0.23; skill +0.24 to +0.45

The India-wide datasets [9]–[11] are retrospective research products that end between 2020 and 2023. Our work differs in three ways: it uses same-day Sentinel-5P trace-gas columns rather than AOD, so that it can run every day; it is evaluated on whole held-out cities and on a held-out year; and the system's operational constraints, three-day-old measurements and forecast rather than reanalysis weather, are built into both training and testing.

III. DATA

A. Ground Truth: CPCB Station Measurements

Daily means of $PM_{2.5}$, PM_{10} and NO_2 were downloaded per sensor from the OpenAQ v3 API [5]. OpenAQ's copy has no CPCB records at all from November 2022 to January 2025.

A second copy of CPCB's records, Project Parisar's 15-minute station files (Kaggle, CC BY-SA 4.0), runs to 31 December 2024 and fills the days OpenAQ lacks [21]. Parisar's station list has no names, so the two copies were linked station by station *by their data*: we required a median difference of at most 10% over at least 30 shared values, clearly better than any other nearby station. This linked 328 stations, whose daily values differ by a median of 2.1%. Coordinates were not used, because the two lists place the same instrument a median of 7.2 km apart. Duplicate OpenAQ sites within 150 m (or 2 km when one name contains the other) were merged, and each station's city was taken from its CPCB name, matched to GeoNames [22].

Every sensor-day passed the seven quality rules of Fig. 1; a flagged value was dropped, never repaired. The detection limits are those of beta attenuation monitors for PM and chemiluminescence analysers for NO_2. Of 1,367,999 OpenAQ sensor-days, 1,014,396 passed every rule; most removals are days with fewer than 16 hours of data, followed by flat-lined and stuck sensors. Of Parisar's 2,040,537 sensor-days, 1,862,324 passed. A further 4,417 station-days reported more $PM_{2.5}$ than PM_{10}, which is physically impossible, so both values were dropped. NO_2 that OpenAQ labels "ppb" is sometimes really µg/m³; each day the unit was settled by the analyser identity $NO_x = NO + NO_2$, checked within 10%. Of the sensor-days labelled ppb, 64,328 were confirmed as ppb, 134,145 were really µg/m³, and 58,964 that fitted neither were dropped.

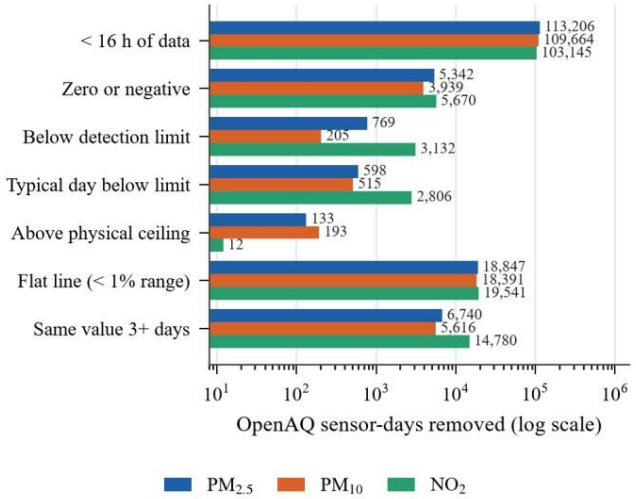


Fig. 1. OpenAQ sensor-days removed by each quality rule, per pollutant (log scale). Of 1,367,999 sensor-days, 1,014,396 passed every rule.

The result is 760,637 clean station-days from 1 July 2018 to 1 October 2026: 719,000 for $PM_{2.5}$, 697,017 for PM_{10} and 666,701 for NO_2. They come from 529 stations in 31 states and union territories (Fig. 2). Of these station-days, 367,203 come from OpenAQ alone, 389,489 from Parisar alone, and 3,945 combine the two (Fig. 3). January 2025 is the only month with no clean data.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

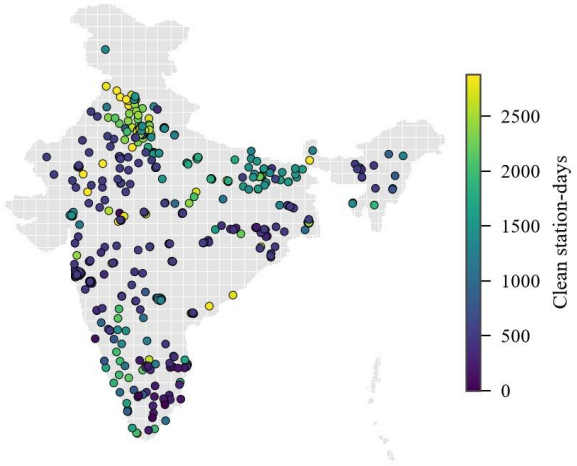


Fig. 2. CPCB stations with clean data, coloured by the number of clean station-days. 525 of the 529 stations are shown; four lie more than 40 km from the city in their own name and were treated as having wrong coordinates. The grey background is the 0.1° grid of India.

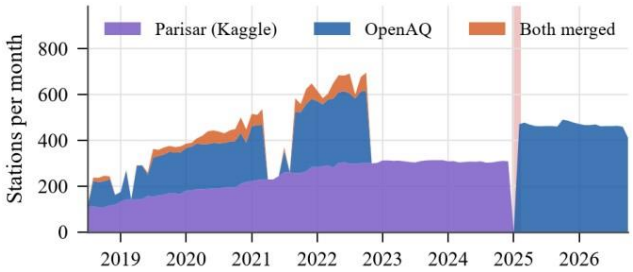


Fig. 3. Stations with at least one clean daily value per month, by source. Parisar's copy fills OpenAQ's gap from November 2022 to December 2024; the shaded band is January 2025, the only month with no clean data.

B. Pollution Levels in the Data

The clean data show a strong seasonal cycle (Fig. 4). Across all stations, the monthly median of daily $PM_{2.5}$ falls from 76 µg/m^3 in November to 22 µg/m^3 in July and August, during the monsoon; PM_{10} moves from 150 to 52 µg/m^3, and NO_2 from 28 in December to 14 in August. This cycle, driven by rain, the depth of the boundary layer and crop-residue burning after the monsoon, is the largest single source of variation the models must explain.

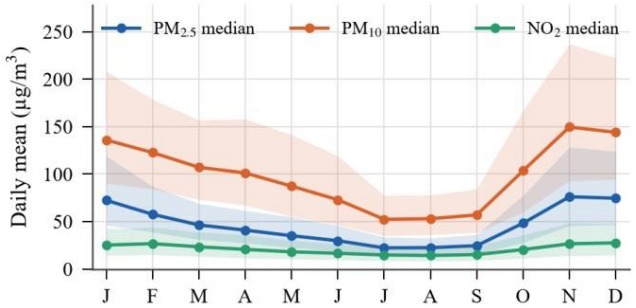


Fig. 4. Seasonal cycle of the clean CPCB data: median daily concentration of all station-days in each calendar month, with the interquartile range shaded.

Converted to CPCB sub-indices (Fig. 5), 15.3% of $PM_{2.5}$ station-days are Poor, Very Poor or Severe, and 36.2% of PM_{10} station-days are Moderate. NO_2 is Good on 80.9% of station-days and almost never worse than Moderate, so in daily means the AQI is set by particulate matter almost everywhere.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

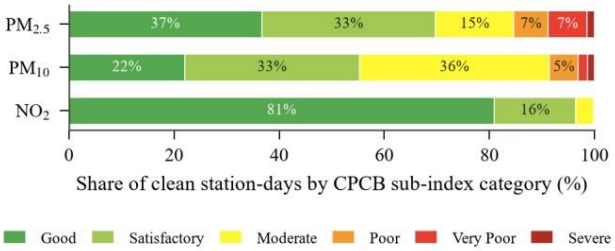


Fig. 5. Share of clean station-days in each CPCB sub-index category, per pollutant (719,000, 697,017 and 666,701 station-days).

C. Inputs: Satellite, Fire, Weather and Land

All inputs are computed in Google Earth Engine [23] on the same 0.1° grid, with one code path for training and for the live map (Table II). The near-real-time products of the TROPOMI instrument on Sentinel-5P [24] give tropospheric NO_2, CO, SO_2, formaldehyde and the UV aerosol index. FIRMS active-fire detections from MODIS [25] are counted within 25, 50 and 100 km over two days. Training uses the ERA5 reanalysis [26] and the live system uses NOAA's GFS forecasts [27]. The static layers are elevation from GMTED2010 [28], population density from GPW v4.11 [29] and land-cover fractions from MODIS MCD12Q1 [30]. Fig. 6 shows four of these inputs for one live day.

TABLE II. MODEL INPUTS

Group	Variables	Source	Latency
Satellite	NO_2, CO, SO_2, HCHO, aerosol index	Sentinel-5P NRTI	same day
Fires	counts within 25/50/100 km	FIRMS (MODIS)	same day
Weather	BLH (mean, min), T, RH, wind, u, v, rain, cloud	ERA5 / GFS	~6 d / same day
Land	elevation (mean, s.d.), population, urban, crop, forest, water	GMTED2010, GPW, MCD12Q1	static
Time, place	day of year, weekday, lat, lon	–	–

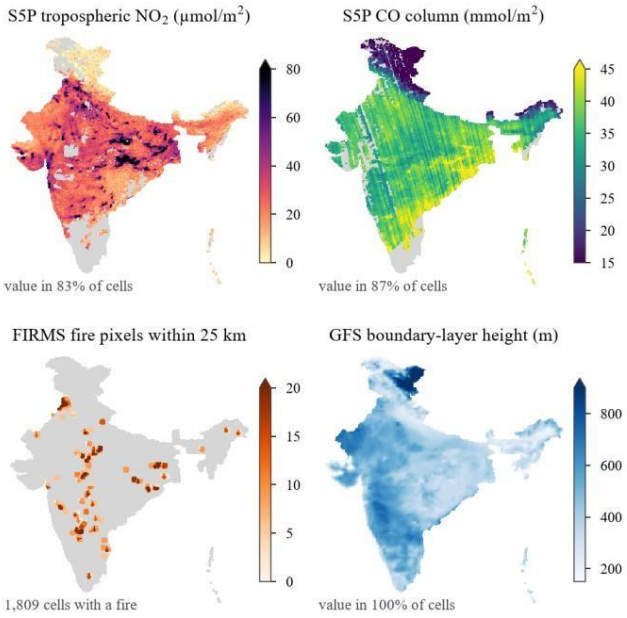


Fig. 6. Four of the model's inputs on 14 October 2026 for all 30,113 cells of India. Grey: no satellite value that day (cloud or no valid retrieval). NO_2 has a value in 83% of cells and CO in 87%; 1,809 cells have a fire within 25 km.

MODIS MAIAC AOD [31] was evaluated only as an ablation, because it arrives about a week late. Missing satellite values stay missing, and LightGBM learns what a gap means. Coverage is incomplete (Fig. 7): over the training table, NO_2 is present on 71% of cell-days, CO on 74%, SO_2 on 69%, formaldehyde on 77% and the aerosol index on 98%, while MAIAC AOD would be present on only 54%.

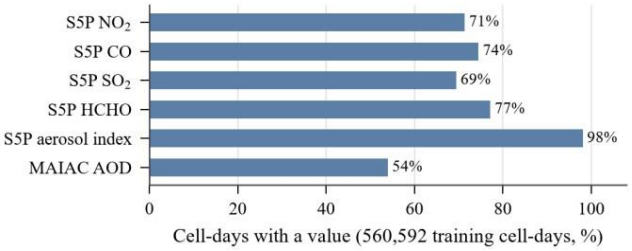


Fig. 7. Share of the 560,592 training cell-days on which each satellite input has a value.

On its own, each input tracks the ground readings only moderately (Fig. 8). The strongest rank correlations with $PM_{2.5}$ are rain (−0.54), cloud (−0.51), the lowest-hour boundary layer (−0.47), Sentinel-5P CO (+0.47) and Sentinel-5P NO_2 (+0.45); MAIAC AOD reaches only +0.27. Weather explains as much as the satellite gases, a first sign of the result in Section V-C.

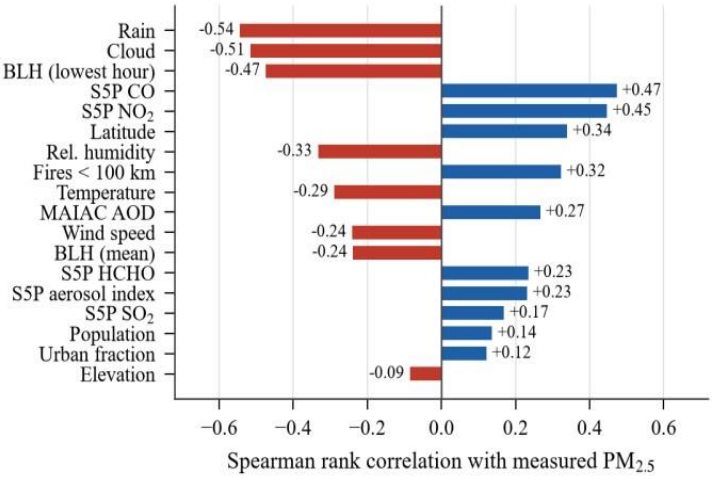


Fig. 8. Spearman rank correlation of each input with measured $PM_{2.5}$ over the training cell-days where both are present. Blue: higher input, more $PM_{2.5}$; red: the opposite.

D. The Training Table

The unit of analysis is the 0.1° cell on one day. A cell's value is the mean of its stations, so a city with forty stations does not count forty times. The training table has 560,592 cell-days in 371 cells and 274 cities, from 1 July 2018 to 26 September 2026, with $PM_{2.5}$ targets on 533,185 cell-days, PM_{10} on 520,992 and NO_2 on 500,216.

E. Mapping Live Weather onto the Training Weather

ERA5 is the better weather record, but it arrives about six days late, so the live system must use GFS. We sampled both at the station cells on 233,120 cell-days between 15 January 2025 and 26 September 2026. They move together, with correlations of 0.66 for rain and 0.80–0.95 for the other variables, but differ in scale. Compared with ERA5, GFS's boundary layer is 26% deeper on average, its air 17 percentage points drier and its wind 29% stronger, so a model trained on ERA5 would see cleaner-looking weather than there is. GFS is therefore mapped onto ERA5's scale per variable, with both linear fits and empirical quantile mapping [32] (Fig. 9); Section V-D measures which one preserves more accuracy.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

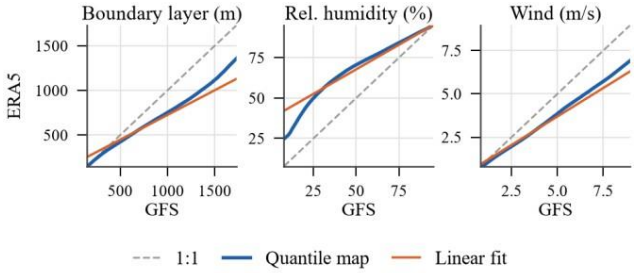


Fig. 9. GFS against ERA5 at the station cells, 15 January 2025 to 26 September 2026. The quantile map (blue) and the linear fit (orange) both pull GFS's deeper boundary layer, drier air and stronger wind onto ERA5's scale; the dashed line is 1:1.

IV. METHODOLOGY

A. System Overview

The system has three layers (Fig. 10). Four sources of data feed a Python pipeline that runs once a day from one command. It cleans the station readings, builds the inputs for every cell, runs both models and computes the AQI and its reasons. The results are saved as files, served by a read-only API and shown on a web globe.

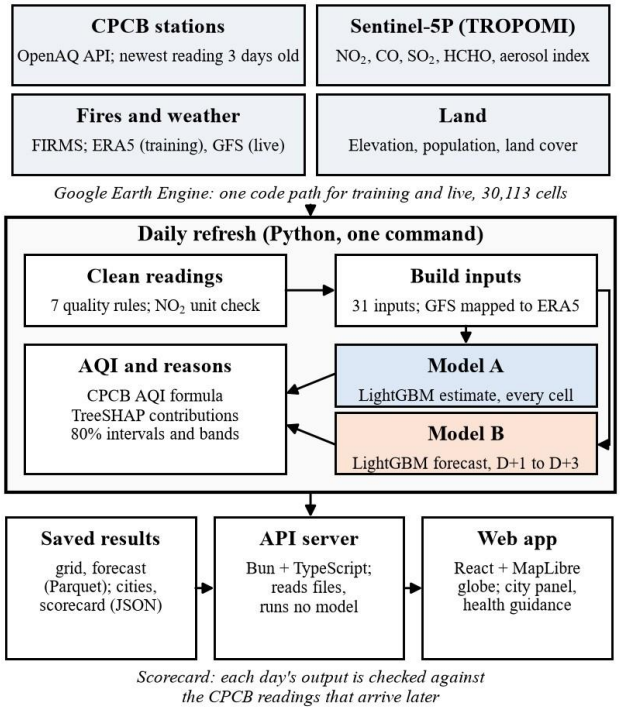


Fig. 10. System architecture. The models run once a day in the refresh; the server only reads the saved results.

B. Model A: Estimating Today's Pollution Anywhere

Model A uses one LightGBM regressor [15] per pollutant, with a learning rate of 0.05, 63 leaves, at least 50 samples per leaf, row and column subsampling of 0.8, an L2 penalty of 1.0 and early stopping after 100 rounds. The target is the concentration itself, not its logarithm; a log target makes the model estimate a typical day rather than the mean, and it biased the estimates low in our tests (Section V-C). The *hybrid* inputs add what monitors elsewhere measured: the inverse-distance-weighted (IDW) mean [33] of the readings at the eight nearest *other* cells, and the distance d_1 to the nearest of them:

$$\hat{v} = \Sigma_k\, w_k\, y_k \,/\, \Sigma_k\, w_k, \quad w_k = d_k^{-2} \qquad (1)$$

The readings y_k are taken from *three days before* the target day, the newest measurement the live system can always count on given OpenAQ's delay. A cell's own reading is never its input.

C. Evaluation Protocol

The map's job is to estimate places with no monitor, so the headline test holds out whole cities. The 274 cities are split into five groups, and each group is estimated by models trained on the other four. The neighbour inputs, too, come only from the training cities, so a held-out city is not a monitor at all [19]. A random split by row is reported only to show how much it flatters a model. The comparisons are IDW interpolation of the monitors alone (same day, the day before and three days before), ridge regression and a random forest [34], and LightGBM without the monitor inputs, without the satellite inputs, with MAIAC AOD, and with a log target. Accuracy is reported as R^2, RMSE and MAE in $\mu g/m^3$.

D. Uncertainty Intervals

Each estimate carries an 80% interval whose width is set from held-out errors in log space at the same distance from the nearest monitor, in bands of 0–10, 10–25, 25–50, 50–100 and over 100 km. With $q_{0.1}$ and $q_{0.9}$ the band's error quantiles, the interval is $[\hat{y}\, e^{q0.1}, \hat{y}\, e^{q0.9}]$. Coverage is checked by cross-conformal calibration [35]: each fold's intervals come only from the other folds' errors.

E. AQI and Explanations

The CPCB AQI [8] converts each pollutant's concentration C_p into a sub-index by linear interpolation between breakpoints (Table III):

$$I_p = I_{lo} + (I_{hi} - I_{lo})(C_p - B_{lo}) / (B_{hi} - B_{lo}) \qquad (2)$$

The AQI is the largest I_p, and that pollutant is reported as dominant. CPCB requires at least three pollutants, one of them $PM_{2.5}$ or PM_{10}, and our three estimates satisfy this. CPCB leaves the Severe band open-ended, so its upper concentrations follow a common convention. The reasons shown for each estimate are TreeSHAP contributions [36], [37], which LightGBM computes exactly; each is how far an input moves the estimate from the model's average.

TABLE III. CPCB AQI BREAKPOINTS USED (24-HOUR MEANS, $\mu G/M^3$)

Category	AQI	$PM_{2.5}$	PM_{10}	NO_2
Good	0–50	0–30	0–50	0–40
Satisfactory	51–100	31–60	51–100	41–80
Moderate	101–200	61–90	101–250	81–180
Poor	201–300	91–120	251–350	181–280
Very Poor	301–400	121–250	351–430	281–400
Severe	401–500	250+	430+	400+

F. Model B: Forecasting One to Three Days Ahead

Model B forecasts $PM_{2.5}$, PM_{10} and NO_2 at monitored cells for leads of 1, 2 and 3 days, using only what is known when the forecast is made:

the cell's own measurements, the newest three days old as on the live system: per pollutant the newest value and its 7- and 30-day means, plus the forecast pollutant's last seven days one by one;

what monitors in other cells measured, and that day's satellite, fire and weather inputs;

the weather forecast for the target day, the calendar, land and position.

Training uses ERA5 for the target day's weather ("perfect prognosis"), because GFS forecasts in Earth Engine reach back only to 2025. The last 365 days, 27 September 2025 to 26 September 2026, are the test year, and the year before it is used for early stopping. On the test year the inputs are the GFS forecasts that were actually issued, mapped with quantile fits learned only from earlier days. Persistence repeats the newest measurement and is the standard bar for a forecast. Skill is measured against it [38]:

$$SS = 1 - MSE_{model} / MSE_{persistence} \qquad (3)$$

A positive SS beats persistence, and 1 is perfect. The other baselines are climatology (the cell's usual level for the time of year), ridge regression, LightGBM without the weather forecast, LightGBM with the weather that actually happened, and LightGBM with one-day-old measurements. LightGBM quantile regressions [39] at $\alpha = 0.1$ and 0.9 give an 80% band, making 27 final models.

G. LSTM Challenger

An LSTM [18], implemented in PyTorch [40] on the CPU, reads each cell's recent days as a sequence. Its final hidden state is concatenated with the non-sequential inputs and passed to a dense head of 128 ReLU units that forecasts all nine values at once. It gets the same information, training years and test year as LightGBM. Training uses Adam with a batch size of 1024 and stops when the validation loss has not improved for four epochs. Four settings of window, hidden units and learning rate were tried, and the one with the lowest loss on the validation year was chosen; the test year played no part in this choice.

V. RESULTS

All numbers below are held-out results: predictions made by models that never saw the row being predicted.

A. Model A on Unseen Cities

Fig. 11 compares all models on whole held-out cities, and Table IV gives their errors. LightGBM on satellite, weather and land inputs alone reaches $R^2 = 0.62$ for $PM_{2.5}$, while ridge regression reaches 0.38, so the relationships are strongly nonlinear. Interpolating same-day monitor readings is as good as LightGBM for particulate matter, but those readings are not available live: interpolation of three-day-old readings drops to $R^2 = 0.37$. Among the models the live system can feed every day, ranked by R^2 averaged over the three pollutants, the hybrid with three-day-old monitors is best (0.49), ahead of LightGBM without monitors (0.48). It is the model the map uses.

TABLE IV. MODEL A ON HELD-OUT CITIES: MAE / RMSE (μG/M^3)

Model	$PM_{2.5}$	PM_{10}	NO_2
LGBM: sat., weather, land	18.6 / 30.0	35.6 / 53.5	12.6 / 19.3
LGBM + monitors, same day†	17.0 / 27.7	34.1 / 52.2	12.4 / 19.0
LGBM + monitors, day before†	17.6 / 28.6	35.5 / 54.2	12.5 / 19.1
LGBM + monitors, 3 days old	**18.3 / 29.3**	**34.7 / 52.0**	**12.5 / 19.1**
Random forest	18.6 / 30.6	36.2 / 54.9	12.5 / 19.2
Ridge regression	24.1 / 38.1	45.3 / 66.3	13.0 / 19.7
IDW, same day†	18.0 / 30.0	35.2 / 53.7	13.6 / 20.9
IDW, day before†	20.3 / 33.7	39.7 / 60.0	13.8 / 21.2
IDW, 3 days old	23.3 / 38.3	45.6 / 68.0	14.1 / 21.6

† Upper bounds: such fresh readings are not available live. Bold: the map's model.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

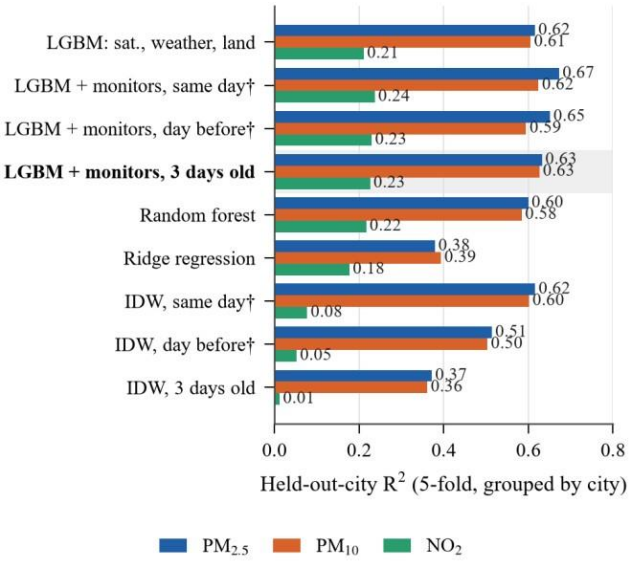


Fig. 11. Held-out-city R^2 of Model A and its baselines (5-fold, grouped by city). Shaded: the model the map uses. † Upper bound: such fresh readings are not available live.

Fig. 12 plots the map model's held-out estimates against the measurements. They follow the diagonal for $PM_{2.5}$ and PM_{10} but are compressed at the high end, a common trait of regression models that minimise squared error. NO_2 is the hard case: its best held-out R^2 is 0.23, and interpolating monitors explains almost nothing (0.01–0.08). This is consistent with NO_2's short lifetime and strong local gradients near traffic.

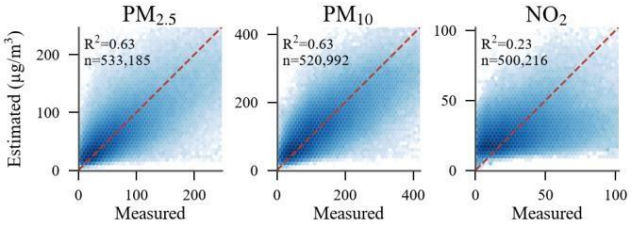


Fig. 12. Map model estimates against measurements on held-out cities (µg/m^3); darker hexagons hold more cell-days, axes end at the 99th percentile, and the dashed line is perfect agreement.

B. Accuracy Against Distance to a Monitor

Accuracy falls with the distance from the held-out place to the nearest training monitor (Fig. 13). For $PM_{2.5}$, the map's model goes from R^2 = 0.77 within 10 km to 0.31 beyond 100 km, while interpolation of three-day-old readings goes from
0.37 to 0.07. Beyond 10 km the map's model, which uses only three-day-old readings, beats interpolation of *same-day* readings for both particulate pollutants (PM_{10} beyond 100 km: 0.34 against 0.19). This is where the satellite, weather and land inputs earn their place.

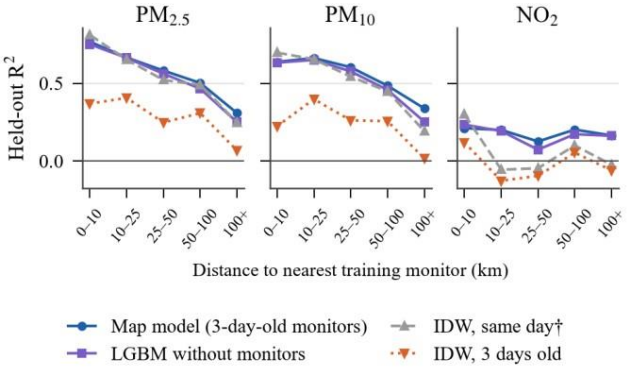


Fig. 13. Held-out R^2 by distance to the nearest monitor in the training cities. † Upper bound, not available live.

C. Ablations: Random Split, Satellite, AOD and Log Target

Fig. 14 shows how each variant changes held-out R^2. A random row split inflates R^2 by 0.17 for $PM_{2.5}$ and PM_{10} and by 0.28 for NO_2 (from 0.21 to 0.49): a model can learn a cell's neighbouring days and be tested on the same cell, which would badly mislead a map of unmonitored places [19]. Removing every satellite input (Sentinel-5P and fires) lowers R^2 by 0.00–0.02, and adding MAIAC AOD adds at most 0.01. The satellite therefore adds little skill of its own: weather, land use, position and season carry most of it.

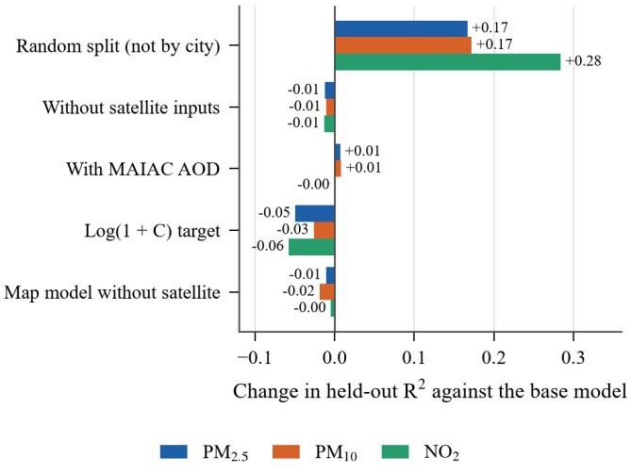


Fig. 14. Change in held-out-city R^2 for each variant of Model A, against LightGBM on satellite, weather and land inputs (first four rows) or against the map model (last row).

Training on log(1 + C) costs 0.03–0.06 R^2 and biases the estimates low (Fig. 15), with a mean error of −6.6, −11.7 and −5.0 µg/m³ for $PM_{2.5}$, PM_{10} and NO_2, against −0.4, −0.8 and +0.1 when trained on the concentration itself.

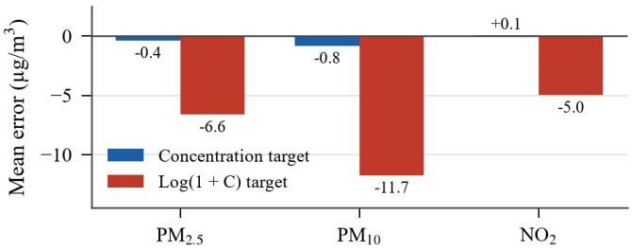


Fig. 15. Mean error (estimate minus measurement) on held-out cities for the two choices of training target.

D. The Cost of Live Weather

The models are trained on ERA5, but the live map has only GFS. Held-out rows from the period both cover were reestimated with GFS weather (Fig. 16). Raw GFS costs up to 0.12 R^2 for PM_{10} (0.53 to 0.41). Quantile mapping recovers most of this (0.49) and is best for both particulate pollutants, so the live system uses it.

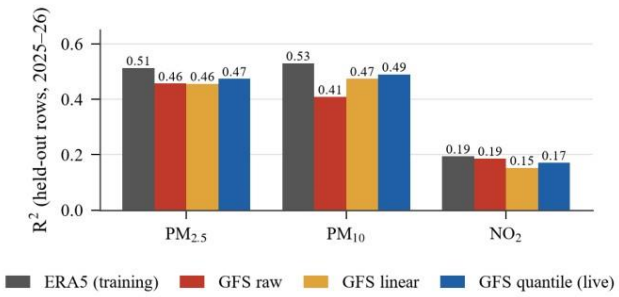


Fig. 16. Map model R^2 on held-out rows from January 2025 to September 2026, with the training weather (ERA5) and with GFS, raw or mapped.

E. Calibration of the Intervals

The cross-conformal 80% intervals contain the measurement on 80% of held-out cell-days in every distance band for $PM_{2.5}$, on 80–81% for PM_{10} and on 79–80% for NO_2 (Fig. 17). They widen with distance, as they should: the $PM_{2.5}$

interval runs from ×0.65–×1.55 of the estimate within 10 km to ×0.51–×1.95 beyond 100 km. NO_2's intervals are the widest, up to ×0.39–×2.57.

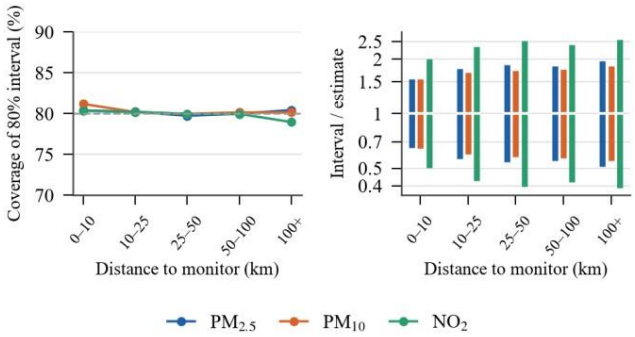


Fig. 17. Left: share of held-out cell-days inside the 80% interval, by distance to the nearest monitor. Right: the interval as a multiple of the estimate (log scale).

F. AQI Category Agreement

On 50,000 held-out cell-days with all three pollutants measured, the CPCB AQI category computed from the map model's estimates matches the category computed from the measurements on 55% of days, and it is at most one category off on 95%.

G. What Drives the Estimates

Fig. 18 shows the inputs with the largest mean absolute TreeSHAP contribution. The interpolated neighbour value leads for all three pollutants. Season, rain, the lowest-hour boundary layer and latitude follow, reflecting the seasonal cycle of Fig. 4 and the north–south gradient. Sentinel-5P CO appears for both particulate pollutants and Sentinel-5P NO_2 for NO_2, so the models do use the satellite, but the other inputs can largely stand in for it. The final models stopped at 341, 1173 and 68 trees respectively.

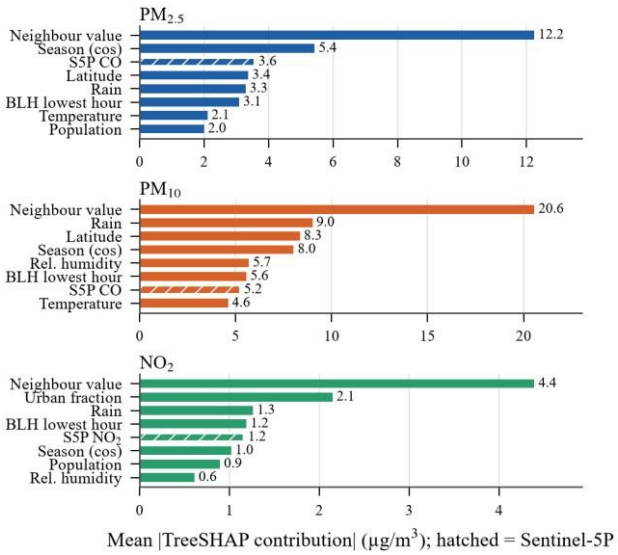


Fig. 18. The eight most influential inputs of each final Model A, by mean absolute TreeSHAP contribution. Hatched bars are Sentinel-5P inputs.

H. Model B on an Unseen Year

Table V and Fig. 19 give the forecasts for the test year. LightGBM beats persistence and climatology at every lead, for every pollutant, with skill from +0.24 to +0.45 (Fig. 20), and it beats ridge regression for both particulate pollutants. Skill is highest for PM_{10} and at three days ahead, where persistence decays most. For NO_2, persistence is already strong ($R^2 = 0.67$ at one day) and ridge regression is as good as LightGBM (skill +0.245 against +0.241 at one day), so the gain from the non-linear model is nil.

TABLE V. MODEL B ON THE TEST YEAR (27 SEP 2025 – 26 SEP 2026), NEWEST MEASUREMENT 3 DAYS OLD

Pollutant	Lead	Pers.	Clim.	Ridge	LGBM	MAE	SS
$PM_{2.5}$	1	0.51	0.40	0.65	**0.68**	12.5	+0.34
	2	0.50	0.41	0.64	**0.66**	12.8	+0.32
	3	0.46	0.40	0.63	**0.66**	13.0	+0.38
PM_{10}	1	0.45	0.35	0.64	**0.68**	24.9	+0.42
	2	0.43	0.35	0.62	**0.67**	25.5	+0.42
	3	0.38	0.34	0.61	**0.66**	26.2	+0.45
NO_2	1	0.67	0.12	0.75	**0.75**	5.7	+0.24
	2	0.65	0.12	0.74	**0.75**	5.9	+0.26
	3	0.63	0.11	0.73	**0.73**	6.0	+0.26

Columns 3–6: R^2. MAE of LightGBM in µg/m^3; SS: skill against persistence, (3). Test forecasts per lead: 82,416–84,308 (PM), 61,348–62,329 (NO_2).

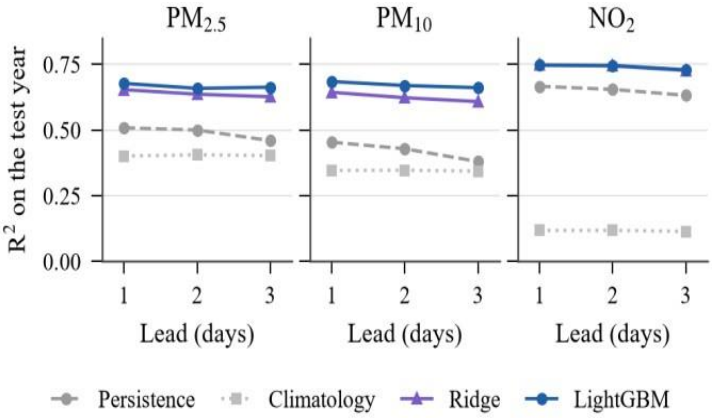


Fig. 19. R^2 of Model B and the reference forecasts on the test year, by lead time.

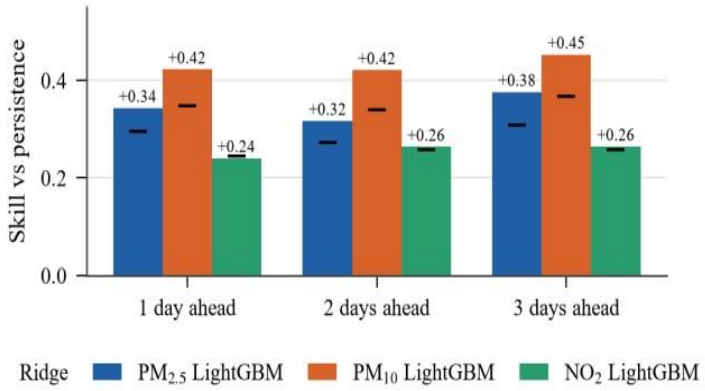


Fig. 20. Skill of LightGBM against persistence, (3), by lead time. The black ticks mark ridge regression.

I. Value of the Weather Forecast and of Fresher Data

Fig. 21 isolates two inputs. Averaged over the three leads, the GFS forecast for the target day adds +0.02 R^2 for $PM_{2.5}$, +0.03 for PM_{10} and −0.00 for NO_2; perfect knowledge of the weather (ERA5) would add +0.06, +0.07 and +0.02. The forecast helps particulate matter most at two and three days ahead. If the newest measurement were one day old instead of three, as CPCB's own real-time feed would allow, the one-day-ahead R^2 would rise from 0.68 to 0.74 ($PM_{2.5}$), from 0.68 to 0.73 (PM_{10}) and from 0.75 to 0.82 (NO_2). OpenAQ's delay therefore costs 0.02–0.07 R^2 across leads.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

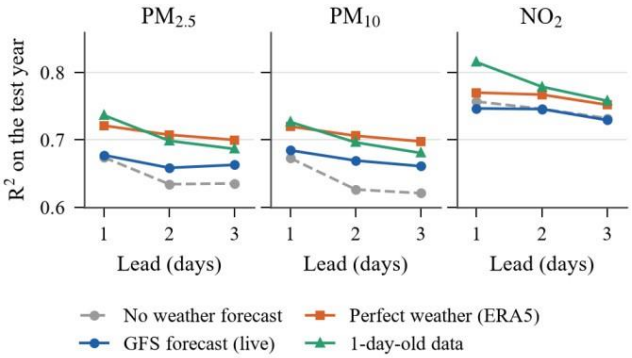


Fig. 21. LightGBM's test-year R^2 without a weather forecast, with the GFS forecast used live, with perfect weather (ERA5), and with the newest measurement one day old instead of three.

J. Forecast Bands and AQI Categories

Fig. 22 follows one monitored Delhi cell, the one with the most test days (338), through the test year. The one-day-ahead forecast tracks the winter build-up and the monsoon minimum with $R^2 = 0.69$, against 0.58 for persistence. At this high-pollution cell its 80% band contains only 61% of the days, so the bands, calibrated over all cells, are too narrow in the most polluted places.

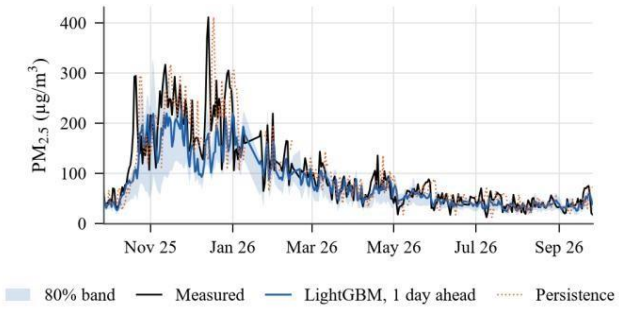


Fig. 22. One-day-ahead $PM_{2.5}$ forecasts for one monitored cell in Delhi over the test year, with the 10th–90th percentile band.

Over all cells, the 10th–90th percentile bands contain 77– 79% of test measurements (Fig. 23). Their median width is 30–32 µg/m^3 for $PM_{2.5}$, 67–70 for PM_{10} and 13–14 for NO_2, and grows slowly with lead. The forecast AQI category is exactly right on 64–65% of test days and at most one category off on 98% (Fig. 24); persistence scores 56–59% (94–95%) and climatology 50% (92%).

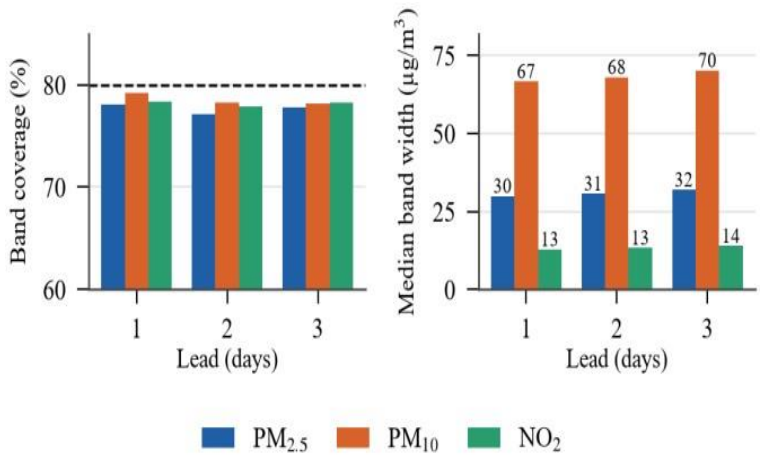


Fig. 23. Model B's 80% bands on the test year: share of measurements inside the band (left; dashed line at 80%) and median band width (right).

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

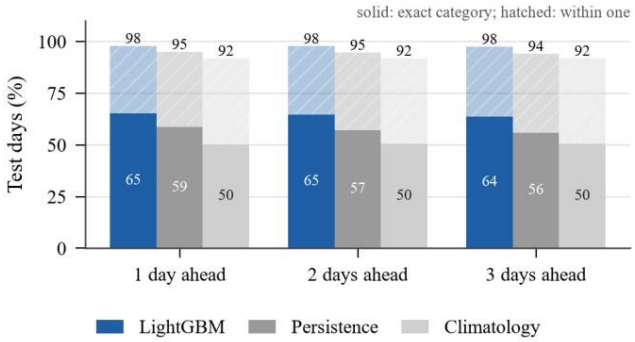


Fig. 24. Forecast AQI category against the measured category on 57,693– 58,564 test cell-days per lead: exactly right (solid) and at most one category off (hatched top).

K. LightGBM Against the LSTM

The chosen LSTM setting used a 14-day window, 64 hidden units and a learning rate of 10^{-3}, with a validation loss of 0.2226; the other three settings had losses of 0.2324–0.2341 (Fig. 25). On the same test days, the LSTM is close to LightGBM but slightly worse in every one of the nine cases (Fig. 26). The mean skill over the nine forecasts is +0.317 for the LSTM and +0.345 for LightGBM. On 8 CPU threads the LSTM took 13 minutes to select its epochs and 2 minutes to refit, while all 27 LightGBM models trained in 3 minutes together. The live system therefore uses LightGBM. Its forecasts rely most on the cell's own 7- and 30-day means and newest value, followed by the season, rain and the forecast lowest-hour boundary layer for the target day.

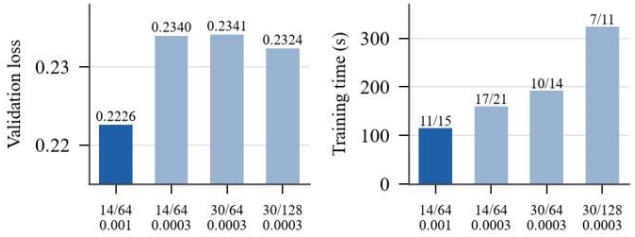


Fig. 25. The four LSTM settings tried (window in days / hidden units, learning rate): validation loss (left) and training time with best/last epoch (right). The chosen setting is dark.

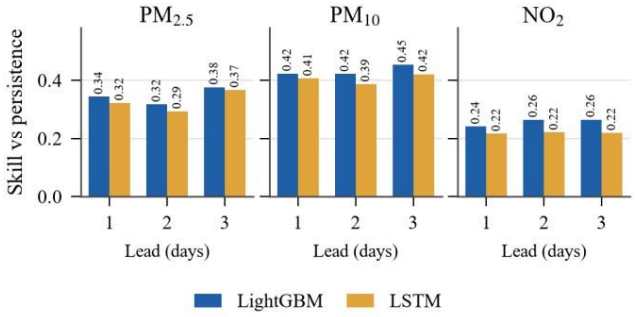


Fig. 26. Skill against persistence of LightGBM and the LSTM on the same test days.

VI. SYSTEM IMPLEMENTATION AND LIVE OPERATION

A. Daily Refresh and API

One command performs the whole update in five steps. It tops up OpenAQ's daily values for every sensor still reporting (about 25 minutes at the free API's rate limit) and cleans them with the rules of Fig. 1. It fetches the day's Sentinel-5P, fire and GFS inputs for all 30,113 cells from Earth Engine in about a minute. It runs Model A on every cell, and Model B at every cell whose monitor reported in the last 30 days (328 cells on 3 October 2026). It writes results for 594 cities: every town of at least 100,000 people plus every monitored city. Finally, it updates a scorecard that compares every refreshed day's output with the measurements that arrive later. Only measurements at least three days old are used, so a past day can be re-run exactly as it would have been; the live record was backfilled from 3 September 2026 this way.

A server written in TypeScript on the Bun runtime reads the refresh's Parquet and JSON files and serves them as a readonly JSON API (/api/status, /api/map, /api/city/{id}, /api/point and /api/models). It never runs a model, so every

request is a file read. The server has 10 automated tests on real excerpts of a refresh, and the Python pipeline has 63 unit tests covering the AQI formula, station cleaning, feature construction, both models and the refresh.

B. Web Application

The front end is written in React with TypeScript, Tailwind CSS and shadcn/ui. MapLibre GL draws India on a 3-D globe over Sentinel-2 cloudless imagery (EOxCloudless 2024 [41]), inside India's official outline [42] (Fig. 27). The 30,113 cells are painted as a single canvas image in the map's projection, because drawing them as separate polygons took about 10 s. City tags carry their AQI, filled where CPCB monitors exist and outlined where the value is only estimated. Clicking a city opens a panel (Fig. 28) with four tabs: the estimate with its range and the newest measurement; health guidance quoting only CPCB's health statements [8], the Health Ministry's NCDC handbook [43] and WHO [3]; the three-day forecast; and the TreeSHAP reasons.

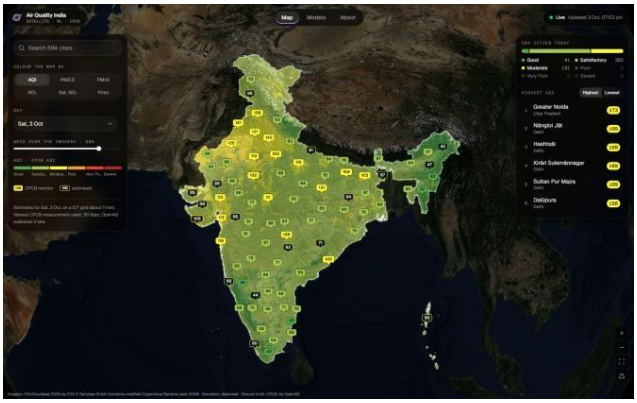


Fig. 27. The web application's India view for 3 October 2026. The 0.1° grid is coloured by estimated AQI; city tags are filled where CPCB monitors exist and outlined where the value is only estimated.



Fig. 28. City panel for Delhi, 3 October 2026: estimated AQI 143 (Moderate, PM_{10}-led), each pollutant with its 80% range, and the model against the newest CPCB measurement.

C. Live Scorecard

The system was run daily, with backfill, from 3 September to 15 October 2026. Its scorecard compares each day's output with the CPCB measurements that arrived later (Fig. 29). Live estimates at monitored cells have an MAE of 10.0 $\mu g/m^3$ for $PM_{2.5}$ over 40 checked days and 11,358 cell-days, lower than in cross-validation, but their R^2 is also lower (0.43). This is expected for six weeks of the monsoon-to-autumn transition, which has less variation to explain than eight years of all seasons; these cells were in the final model's training data, but these days were not. Live forecast R^2 is 0.44–0.60. The daily errors (Fig. 30) stay between 7.6 and 12.9 $\mu g/m^3$ for $PM_{2.5}$ and NO_2. PM_{10}'s error rises from 15–20 $\mu g/m^3$ in the first half of September to 24–31 $\mu g/m^3$ in October, as concentrations climb after the monsoon (Fig. 4).

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

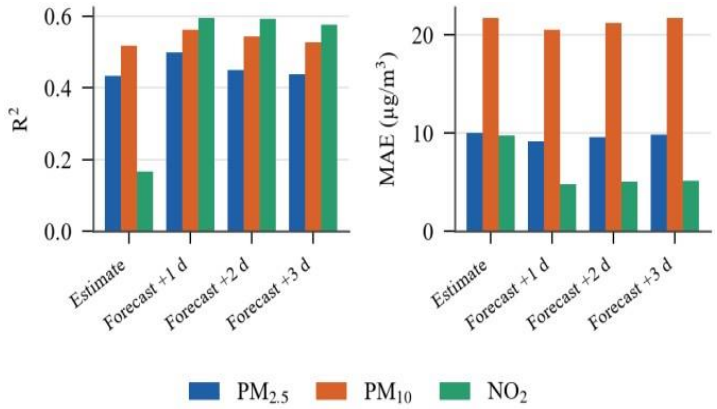


Fig. 29. Live scorecard, 3 September – 15 October 2026: R^2 and MAE of the estimates (40 checked days) and of the forecasts (37–39 days) at monitored cells.

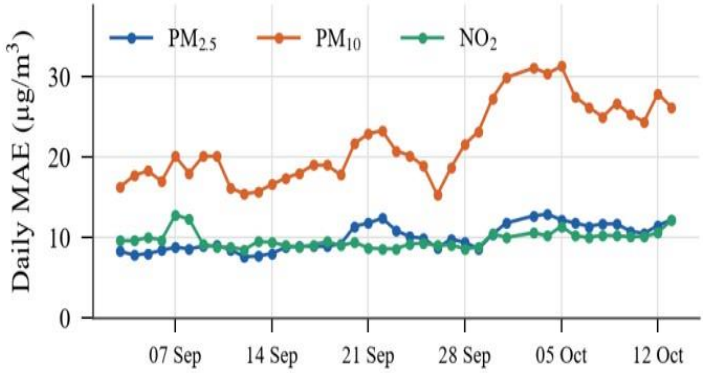


Fig. 30. Daily MAE of Model A's live estimates at monitored cells, 3 September – 13 October 2026.

D. India on 15 October 2026

Fig. 31 shows the AQI category estimated for every cell on the newest live day, 15 October 2026. 22,182 of the 30,113 cells are Satisfactory, 6,161 Moderate and 1,770 Good, with a median AQI of 82; PM_{10} sets the AQI in 28,540 cells. Moderate cells lie mostly in the north-west and centre of the country, and Good cells mostly in the north-east and along the south-west coast. Of the 594 mapped cities and towns (Fig. 32), 95 of the 278 with a CPCB monitor and 121 of the 316 with only an estimate are Moderate; none is Poor or worse on this day.

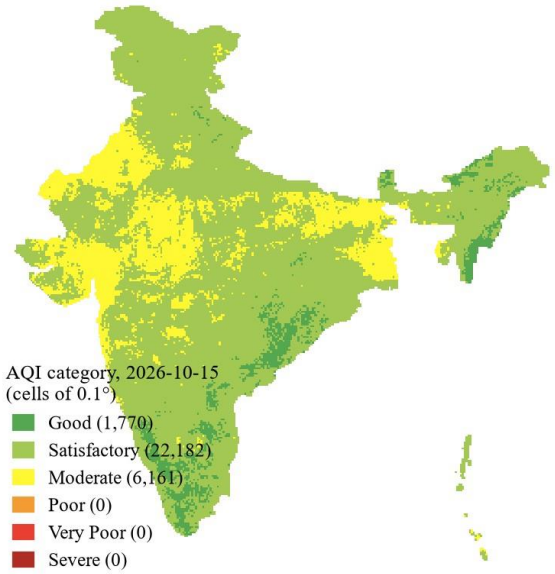


Fig. 31. Estimated CPCB AQI category for all 30,113 cells of India, 15 October 2026.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

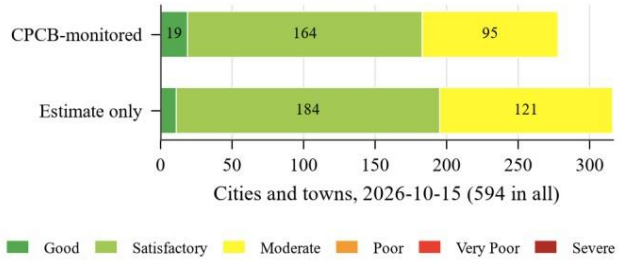


Fig. 32. AQI category of the 594 mapped cities and towns on 15 October 2026, split by whether a CPCB monitor exists.

E. Case Study: Delhi, 3 October 2026

Averaged over Delhi's 15 cells, Sentinel-5P recorded NO_2 of 89.4 µmol/m^2, CO of 33.7 mmol/m^2 and an aerosol index of 0.66; GFS gave a boundary layer of 637 m (94 m at its lowest hour), 30.5 °C, 53% humidity, 2.1 m/s wind and no cloud; and the nearest usable monitor was 10 km away. Model A estimated $PM_{2.5}$ 53.6 (34.0–84.2), PM_{10} 164.1 (104.9–256.5) and NO_2 29.4 (14.1–60.2) µg/m^3. By (2), the sub-indices are 143 for PM_{10}, 90 for $PM_{2.5}$ and 37 for NO_2, so the AQI is 143: Moderate, led by PM_{10} (Fig. 28). The largest contributions to the PM_{10} estimate were the interpolated neighbour value (+22.7 µg/m^3), latitude (+16.7), the lowest-hour boundary layer (−11.2) and Sentinel-5P CO (−10.5). The newest CPCB measurement, for 1 October, gave an AQI of 173, against the model's 139 for that day. Model B forecast an AQI of 124, 126 and 132 for 4–6 October, all Moderate.

VII. DISCUSSION AND LIMITATIONS

What the satellite adds. The title of this work promises satellite assistance, and we found its marginal value to be small. Removing Sentinel-5P and fire inputs costs at most 0.02 R^2, even though TreeSHAP ranks Sentinel-5P CO and NO_2 among the top inputs. Weather, season, land use and position are strongly correlated with the same signal (Fig. 8), so the model can substitute them, and cloud and failed retrievals leave the four trace gases unobserved on 23–31% of cell-days. On the live map the satellite still supplies an independent observation of the air.

Limitations. NO_2 remains hard away from monitors (held-out R^2 0.23). Forecasts are made only at monitored cells, and their bands are too narrow at the most polluted cells. Each estimate represents an 11-km cell, not a street. Ozone, SO_2 and CO are not estimated, so the AQI is computed from three pollutants and may understate it where ozone dominates, and the upper concentrations of CPCB's Severe band are a convention. The three-day delay in OpenAQ's data costs 0.02–0.07 R^2 in forecasting. Earth Engine's free Community tier (150 EECU-hours per month) was exhausted by the historical extraction, so daily operation runs in its restricted mode. The live scorecard covers six weeks so far, all before the winter peak.

VIII. CONCLUSION AND FUTURE WORK

We built and evaluated an end-to-end system that estimates daily $PM_{2.5}$, PM_{10} and NO_2 for every 11-km cell of India and forecasts them up to three days ahead. Every reported number comes from tests that reproduce the live system's conditions: whole cities held out, a year held out, three-day-old measurements and the GFS forecasts that were actually issued. A gradient-boosted hybrid of satellite, weather, land and interpolated monitor inputs explains 63% of $PM_{2.5}$ and PM_{10} variation in cities it never saw, with calibrated intervals, and the forecaster beats persistence by a skill of up to +0.45 and an LSTM by a small but consistent margin. The system runs from one command a day and presents its results, with explanations and health guidance, on a web globe for 594 cities. Future work will add daily VIIRS aerosol retrievals, count only fires upwind of each cell, average satellite columns over several days to reduce cloud gaps, use CPCB's real-time feed to cut the measurement delay, and extend the system to ozone, SO_2 and CO so that the AQI uses every pollutant CPCB defines.

ACKNOWLEDGMENT

The authors thank the Central Pollution Control Board for its monitoring data; OpenAQ and Project Parisar for making that data openly available; ESA/Copernicus for Sentinel-5P and Sentinel-2; NASA for MODIS and FIRMS; ECMWF for ERA5; NOAA for GFS; USGS, CIESIN and GeoNames for land, population and place data; datameet for India's outline; EOX IT Services for the EOxCloudless imagery; and Google for Earth Engine.

REFERENCES

[1] A. Pandey *et al.*, "Health and economic impact of air pollution in the states of India: The Global Burden of Disease Study 2019," *Lancet Planet. Health*, vol. 5, no. 1, pp. e25–e38, 2021, doi: 10.1016/S25425196(20)30298-9.

[2] J. de Bont *et al.*, "Ambient air pollution and daily mortality in ten cities of India: A causal modelling study," *Lancet Planet. Health*, vol. 8, no. 7, pp. e433–e440, 2024, doi: 10.1016/S2542-5196(24)00114-1.

[3] World Health Organization, *WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva, Switzerland: WHO, 2021.

[4] M. Brauer *et al.*, "Examination of monitoring approaches for ambient air pollution: A case study for India," *Atmos. Environ.*, vol. 216, Art. no. 116940, 2019, doi: 10.1016/j.atmosenv.2019.116940.

[5] OpenAQ, "OpenAQ: Open air quality data platform, API v3." [Online]. Available: https://openaq.org

[6] Bar & Bench, "Brahmapuram fire: Kerala High Court takes suo motu cognizance of waste dump blaze, toxic smog in Kochi," Mar. 2023. [Online]. Available: https://www.barandbench.com

[7] S. Das et al., "Air quality degradation from open waste burning: A case study of Brahmapuram, Kochi, India," *Aerosol Sci. Eng.*, 2026, doi: 10.1007/s41810-026-00415-7.

[8] Central Pollution Control Board, *National Air Quality Index*. New Delhi, India: Ministry of Environment, Forest and Climate Change, Oct. 2014.

[9] S. Mandal *et al.*, "Nationwide estimation of daily ambient $PM_{2.5}$ from 2008 to 2020 at 1 km^2 in India using an ensemble approach," *PNAS Nexus*, vol. 3, no. 3, Art. no. pgae088, 2024.

[10] S. Wang *et al.*, "Reconstructing long-term (1980–2022) daily ground particulate matter concentrations in India (LongPMInd)," *Earth Syst. Sci. Data*, vol. 16, pp. 3565–3577, 2024, doi: 10.5194/essd-16-35652024.

[11] A. Kawano *et al.*, "Improved daily $PM_{2.5}$ estimates in India reveal inequalities in recent enhancement of air quality," *Sci. Adv.*, vol. 11, no. 4, Art. no. eadq1071, 2025, doi: 10.1126/sciadv.adq1071.

[12] J. Kang *et al.*, "Improving estimation of surface $PM_{2.5}$ by including satellite observations of gases, aerosols, and radiation in tandem," *Environ. Res. Lett.*, vol. 20, no. 12, Art. no. 124019, 2025, doi: 10.1088/1748-9326/ae1e17.

[13] D. Griffin *et al.*, "Development and validation of satellite-derived surface NO_2 estimates using machine learning versus traditional approaches in North America," *Atmos. Meas. Tech.*, vol. 18, pp. 7497– 7511, 2025.

[14] L. Mamić, M. Gašparović, and G. Kaplan, "Developing $PM_{2.5}$ and PM_{10} prediction models on a national and regional scale using open-source remote sensing data," *Environ. Monit. Assess.*, vol. 195, no. 6, Art. no. 644, 2023.

[15] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 3146–3154.

[16] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: A survey," *Artif. Intell. Rev.*, vol. 56, pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.

[17] J. Bernacki and R. Scherer, "A comprehensive review of data-driven techniques for air pollution concentration forecasting," *Sensors*, vol. 25, no. 19, Art. no. 6044, 2025, doi: 10.3390/s25196044.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[19] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

[20] J. Linnenbrink, C. Milà, M. Ludwig, and H. Meyer, "kNNDM CV: k-fold nearest-neighbour distance matching cross-validation for map accuracy estimation," *Geosci. Model Dev.*, vol. 17, pp. 5897–5912, 2024, doi: 10.5194/gmd-17-5897-2024.

[21] Project Parisar, "India Air Quality Index (AQI) Dataset 2010–2024," Kaggle, CC BY-SA 4.0. [Online]. Available: https://www.kaggle.com/datasets/omsandeeppatil/indian-aqi-stations

[22] GeoNames, "GeoNames geographical database," CC BY 4.0. [Online]. Available: https://www.geonames.org

[23] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017, doi: 10.1016/j.rse.2017.06.031.

[24] J. P. Veefkind *et al.*, "TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications," *Remote Sens. Environ.*, vol. 120, pp. 70–83, 2012, doi: 10.1016/j.rse.2011.09.027.

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

[26] H. Hersbach *et al.*, "The ERA5 global reanalysis," *Q. J. R. Meteorol. Soc.*, vol. 146, no. 730, pp. 1999–2049, 2020, doi: 10.1002/qj.3803.

[27] National Centers for Environmental Prediction, NWS, NOAA, "NCEP GFS 0.25 degree global forecast grids historical archive," NCAR Research Data Archive, 2015, doi: 10.5065/D65D8PWK.

[28] J. J. Danielson and D. B. Gesch, "Global multi-resolution terrain elevation data 2010 (GMTED2010)," U.S. Geological Survey, OpenFile Rep. 2011-1073, 2011, doi: 10.3133/ofr20111073.

[29] Center for International Earth Science Information Network (CIESIN), Columbia University, "Gridded Population of the World, version 4 (GPWv4): Population density, revision 11," NASA SEDAC, 2018, doi: 10.7927/H49C6VHW.

[30] M. Friedl and D. Sulla-Menashe, "MODIS/Terra+Aqua land cover type yearly L3 global 500 m SIN grid V061 (MCD12Q1)," NASA EOSDIS Land Processes DAAC, 2022, doi: 10.5067/MODIS/MCD12Q1.061.

[31] A. Lyapustin, Y. Wang, S. Korkin, and D. Huang, "MODIS Collection 6 MAIAC algorithm," *Atmos. Meas. Tech.*, vol. 11, no. 10, pp. 5741– 5765, 2018, doi: 10.5194/amt-11-5741-2018.

[32] L. Gudmundsson, J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen, "Technical note: Downscaling RCM precipitation to the station scale using statistical transformations – A comparison of methods," *Hydrol. Earth Syst. Sci.*, vol. 16, no. 9, pp. 3383–3390, 2012, doi: 10.5194/hess16-3383-2012.

[33] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proc. 23rd ACM Nat. Conf.*, 1968, pp. 517–524, doi: 10.1145/800186.810616.

[34] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[35] V. Vovk, "Cross-conformal predictors," *Ann. Math. Artif. Intell.*, vol. 74, no. 1–2, pp. 9–28, 2015, doi: 10.1007/s10472-013-9368-4.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[37] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[38] A. H. Murphy, "Skill scores based on the mean square error and their relationships to the correlation coefficient," *Mon. Weather Rev.*, vol. 116, no. 12, pp. 2417–2424, 1988.

[39] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978, doi: 10.2307/1913643.

[40] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

[41] EOX IT Services GmbH, "EOxCloudless 2024" (contains modified Copernicus Sentinel data 2024), CC BY-NC-SA 4.0. [Online]. Available: https://cloudless.eox.at

[42] datameet, "Maps of India: Administrative boundaries," CC0. [Online]. Available: https://github.com/datameet/maps

[43] National Centre for Disease Control, Ministry of Health and Family Welfare, *Air Pollution and its Impact on Health: Handbook for ASHA*. New Delhi, India: NCDC, 2024.