# How the Delhi air quality analysis works

The method behind the daily accountability cut

Prepared by the office of Ajay Maken

10 August 2026

## Contents

# 1 Contents

Every entry below is a link. Click a part to jump to it.

| Part | What it covers |
|---|---|
| I. What this document is | Who the paper is written for, the three things the site sets out to do, and why the weather has to be removed before any of them can be done honestly. |
| II. The data | Where every reading comes from, how incomplete days are excluded, and how the city average is weighted by population. |
| III. The measures | How an AQI figure is computed, the daily window used, and the limits it is measured against. |
| IV. De-weathering | The core method. What the model sees, why yesterday's pollution is deliberately withheld from it, how it is fitted and retrained, and what it costs to run. |

| Part | What it covers |
|---|---|
| V. The statistics | How this year is compared with last, how uncertainty is carried, and what "too close to call" means. |
| VI. The products | The daily report, the daily film, the weekly edition and the special investigations. |
| VII. The role of AI | What the AI does, the hard limits on what it cannot do, and the panel that reviews each draft. |
| VII-A. The long record | The satellite reconstruction of Delhi's air back to 1980, its independent cross-check, and what it cannot settle. |
| Disclaimers | The limits of the method, gathered in one place. |
| VIII. Literature | A published source for every method used here. |
| IX. Known limitations | What this method cannot settle, and what would count against our own conclusions. |
| Version history | What changed in each edition of this paper, and when. |

This paper is complete. Where a figure is an example rather than a fixed property of the method, it is dated and labelled as such.

# 2 Part I. What this document is

This site publishes a daily judgement about Delhi's air. This paper sets out how that judgement is made, in enough detail that a reader can check it rather than trust it.

It is written for three readers at once. A journalist needs to know what the numbers mean. A researcher needs the model specifications and the data sources. A citizen needs to know whether the air their household breathes is getting worse, and whether anyone is answerable for it. The paper is built for the most sceptical of the three, because a method that survives scrutiny serves the other two.

## 2.1 What the site is for

Delhi's air is discussed for four months of the year and forgotten for the other eight. Attention arrives with the smog in October and leaves once January has passed. The pollution keeps no such calendar, and neither does this site, which publishes every day of the year.

Three things follow from that, and the method exists to serve all three.

**It puts the measured record in front of people.** Every figure is computed from the hourly station readings compiled by the Centre for Research on Energy and Clean Air from the official CPCB and DPCC network, so a reader can see for themselves whether the air is improving.

**It carries what that pollution does to health**, drawn from the established epidemiological literature rather than from anything estimated here.

**It sets out what could be done.** A daily account that recorded only the damage would be a complaint rather than a case, so each finding is taken through to a remedy.

### 2.2 What has been done with it

None of this is published only to a website. On 5 August 2026 a submission built from this method went to Parliament and to the Ministry. It was sent to Shri Bhupender Yadav, Hon'ble Minister of Environment, Forest and Climate Change. It went to the Hon'ble Chairperson of the Department-related Parliamentary Standing Committee on Science and Technology, Environment, Forests and Climate Change, Rajya Sabha. And it went to Shri Jairam Ramesh, as a Member of that Committee.

The submission runs to 26 pages, one finding to a page, each carrying a dated source line, with a one-page summary for the Committee's secretariat to circulate. It asks the Committee to summon the Commission for Air Quality Management and the Ministry, and it offers to depose.

Both the submission and its covering letters are published on this site beside this paper, so the case put to Parliament and the method behind it can be read against each other.

### 2.3 Why the weather has to be removed

When Delhi's air improves, a government says its policies worked. When it worsens, the same government says the weather was against it. Neither claim can be checked from the raw pollution number, because that number contains both effects at once.

So the analysis separates them. Every day it asks a narrower question. Was Delhi's air worse than the same period last year, after the weather is accounted for? If it was, the weather is not the explanation.

## 3 Part II. The data

Every measurement used here is public. This office runs no sensors of its own and adjusts no readings. What it adds is the analysis, and any error in that is its own.

### 3.1 Ground measurements

Pollutant readings come from the Central Pollution Control Board's monitoring network, the same official data the government publishes. They reach this analysis through the Centre for Research on Energy and Clean Air, which compiles the raw hourly station data from the CPCB and the Delhi Pollution Control Committee into a single clean record.

That compilation is what every number here is computed from. Without it this daily analysis would not be possible, and the debt is recorded plainly.

### 3.2 Weather

Weather comes from the Copernicus ERA5 reanalysis, obtained through Open-Meteo and Google Earth Engine. The variables used are the ones the model needs: wind speed and direction, temperature, boundary layer height and precipitation.

### 3.3 Satellite

Satellite products are European Space Agency Sentinel-5P, carrying TROPOMI, and NASA MODIS. Both are freely available through Google Earth Engine. They support wider evidence rather than the daily accountability claim.

### 3.4 When a day is thrown away

A station-day counts towards any average only if it has at least 16 valid hourly readings out of 24. Readings outside a plausible range are discarded before that count is taken.

This rule is applied before anything else. A thin network biases the city average in ways that would flatter or damn unfairly, so on days when too few stations reported, nothing is published at all.

### 3.5 Weighting the city average

A simple mean across stations would weight an industrial monitor with few residents around it the same as one in a dense colony. So the city figure is weighted by the population each station represents, using zones drawn around the station network with population derived from electoral rolls.

### 3.6 The network, in numbers

The size of the network is not fixed. Stations report or fail to report on any given day, which is why the denominator is stated on every figure rather than assumed.

On 9 August 2026, taken as a worked example, 39 stations passed the completeness rule for PM2.5 and 40 for PM10. Population weights were held for 44 stations. The year-on-year comparison for that day rested on 35 stations that reported in both years.

Those four numbers differ from one another, and the difference is the point. A figure quoted against "all stations" without saying which stations, on which day, under which rule, is not checkable. Every table on this site carries its own denominator.

## 4 Part III. The measures

### 4.1 The index

The Air Quality Index used here is the CPCB's own. Each pollutant is converted to a sub-index on the official scale and averaging window, and the worst sub-index becomes the day's AQI. The pollutant responsible is named, because an AQI figure without its driver tells a reader very little.

### 4.2 The daily window

The daily figure covers the 24 hours to 10 AM IST, not the calendar day.

The reason is practical, and the consequence is disclosed. A calendar day cannot be reported until it has ended, which would put every bulletin a day behind. The consequence is that the observed AQI here will not exactly equal the CPCB portal's published calendar-day value. A separate calendar-day cross-check is maintained against the portal for that reason.

Every report and every film states the window it used. A comparison a reader cannot inspect is one they should not trust.

### 4.3 What the figures are measured against

Two reference points are used, and they differ in kind. The World Health Organization guideline is a health-based value. The National Ambient Air Quality Standard is India's legal limit. A reading can comply with the second while being several times the first, so reports say which is meant.

## 5 Part IV. De-weathering

This is the core of the method and the part most worth attacking.

### 5.1 The problem

Pollution depends heavily on conditions nobody controls. Wind speed and direction, the height of the mixing layer, temperature, and rainfall all move the number. A still, cold night traps pollution near the ground. A windy afternoon disperses it. Neither is policy.

De-weathering estimates what the day's pollution would have been under normal weather. What remains is the part a government can be asked to answer for.

### 5.2 What the model sees

A random forest is trained for each of seven pollutants: PM2.5, PM10, CO, NO2, NOx, SO2 and ozone. It has fifteen predictors, in three families that the method treats differently.

| Family | Count | Predictors |
|---|---|---|
| Contemporaneous weather | 6 | wind speed, wind direction as sine and cosine, temperature, boundary layer height, precipitation |
| Lagged weather | 5 | trailing 6h and 24h rainfall, 6h change in boundary layer height, 6h-lagged wind speed, 24h mean wind speed |
| Identity and time | 4 | station, hour, weekday, decimal-date trend |

*Source: the model formula in* `dw_pm25_pipeline/train_dw_pm25_z2.R`*, read from the file at build time rather than transcribed.*

The third family is the methodological heart. During normalisation the weather is substituted and the identity and time terms are held fixed. So the procedure removes weather without removing the secular trend it is trying to measure.

### 5.3 The choice a referee will look for

Only lagged **weather** enters the model. Lagged **pollutant** values are deliberately excluded, and the trainer records this in its own header.

The reason matters. Feeding yesterday's pollution into a model meant to isolate emissions would launder the answer through the target variable. The model would appear to explain more, and the explanation would be worthless.

### 5.4 How the model is fitted

Each pollutant is tuned independently. The grid is mtry of 3, 4 or 5, by leaf size of 5 or 10, at 500 trees. Selection is on mean held-out R squared across five contiguous chronological folds, expanding-window style. That is 210 tuning fits per full retrain across the seven pollutants.

The kept model is a single forest of 800 trees per pollutant. The random seed is 42.

### 5.5 Retraining

A full retrain runs weekly, under a scheduled task on Sunday at about 03:30 IST.

One caveat is printed rather than smoothed over. The trainer's own gate also permits a retrain on any day from October to February. That clause has never been exercised in the available log, because no such month has occurred while the log existed. So the observed cadence is weekly, and the in-season clause is a permission this paper has not yet seen used.

### 5.6 What it costs to run

Measured, not estimated. Nine completed full seven-pollutant retrains are recorded, at a mean wall clock of 136 minutes, range 130 to 160, on one workstation using all cores but two. The consistency across nine runs is itself worth reporting.

*Source:* `dw_pm25_pipeline/_weekly_retrain.log`*, parsed at build time.*

### 5.7 A second model family, and what it is for

A boosted regression tree model is trained on the same data, as a deliberately separate family: 5,000 trees, interaction depth 6, shrinkage 0.01, bagging fraction 0.5, minimum 10 observations per node.

**Production uses the random forest for all seven pollutants.** The boosted trees exist to be checked against it. This paper does not claim two independent model families in the published product, because that is not yet true.

On 9 August 2026 the two were compared on 1.49 million rows, 20 per cent held out:

| Pollutant | Boosted trees | Forest | Difference |
|---|---|---|---|
| CO | 0.5572 | 0.4509 | +0.106 |
| NO2 | 0.6004 | 0.5218 | +0.079 |
| Ozone | 0.6364 | 0.5833 | +0.053 |
| NOx | 0.6369 | 0.6020 | +0.035 |
| SO2 | 0.5397 | 0.5950 | −0.055 |

| Pollutant | Boosted trees | Forest | Difference |
|---|---|---|---|
| PM10 | 0.6589 | 0.7473 | −0.088 |
| PM2.5 | 0.6988 | 0.7882 | −0.089 |

Held-out and training R squared agree closely, so the gains are not overfitting. The same split had been found independently two months earlier, on less data. The gains are largest where the forest is weakest, which is where the published product is least certain.

Whether to adopt the better model for those four pollutants is under evaluation, and no published number has been changed while that is decided.

# 6 Part V. The statistics

## 6.1 Comparing this year with last

A single day is a noisy thing to compare against, and choosing which day to compare against invites the fair criticism that a flattering one was picked.

So the baseline is not a single day. It is the average of a fifteen day window: the same calendar date last year, plus or minus seven days. Only stations that reported in both years are used, so the comparison is between like and like rather than between two different networks.

## 6.2 Carrying uncertainty

Every comparison is an estimate from a sample of stations. So every headline figure is published with an interval around it, computed by resampling the stations.

## 6.3 Too close to call

Where the interval around a year-on-year difference contains zero, the site says the result is too close to call. It does not report the point estimate as a finding.

This rule costs us stories. It is also the reason the site can be believed on the days when a finding does survive it.

## 6.4 Which finding leads

The lead is chosen by a fixed rule set, not by a person and not by a language model.

The rules, in the order they apply. Only findings describing worsening, weather corrected change or exceedance may lead. A de-weathered or citywide finding outranks a single-station or observed one. A lead of the same kind is not repeated within a set number of days, unless the strongest available finding is of that kind anyway, in which case correctness beats novelty.

Supporting points follow in priority order. A health characterisation and the policy asks are always carried last.

# 7 Part VI. The products

Four things are published from the same analysis. They are built from the same computed figures, so they cannot disagree with one another.

**The daily report.** A single PDF covering the day's air, the station detail and the attribution question. Its figures are copied from the computed sources rather than recalculated, which is what stops the report and the film telling different stories.

**The daily film.** A short bulletin in Hindi, narrated, covering the same findings as the written report. A standing rule requires the narration to carry every negative finding the report prints, and an automatic check stops the day's run if one is missing.

**The weekly edition.** A longer piece taking the week together, where a single day is too noisy to support a conclusion.

**Special investigations.** Published when the data warrants a closer look than the daily cycle allows, such as the examination of the government's 2025 improvement claim.

One rule binds all four. The figures are computed once and copied into each product. Nothing recalculates a number for its own use, which is what makes it impossible for the report and the film to disagree about what the day showed.

# 8 Part VII. The role of AI

## 8.1 What the AI does, and what it cannot do

No language model computes a number, and none can place a value into a report. Every figure on this site is produced by deterministic R and Python code run over public data.

Which findings lead each day is decided by a published rule set, not by a person and not by a model. That rule set is set out in Part V.

A language model does write the day's Hindi wording, so that the bulletin reads freshly rather than as one template with the numbers swapped. It is held inside a hard boundary. Every number and every technical term in the computed line must survive the rewrite unchanged, and the length must stay within fixed limits. A line that fails any of those checks is thrown away, and the computed wording is used instead.

The review panel reads the draft and the data behind it. It can object. It cannot edit, and no agent writes to the code that produces the numbers. The panel does not run on a schedule. It runs only when Ajay Maken asks for it, and the software refuses to start it any other way.

## 8.2 The panel

Each day's draft can be reviewed by an AI-generated and trained panel covering atmospheric chemistry, boundary layer meteorology, exposure science, environmental statistics, and narrative accuracy, together with an adversarial reviewer whose only task is to attack the draft the way a hostile spokesperson would.

Points that survive review are applied before publication. Points that do not are recorded and discarded.

### 8.3 Why this is disclosed at all

The findings on this site are uncomfortable for more than one party, including parties this office has belonged to. That is only defensible if the method is stricter than the conclusion, and a method nobody can inspect is not stricter than anything.

## 9 Part VII-A. The long record, 1980 to 2022

The daily analysis described so far begins where Delhi's monitoring network begins, and that is the 2010s. The years in which the city's air was actually ruined were never measured at the time. A separate exercise reconstructs them, and because it answers a different question by a different method it is set out here on its own.

### 9.1 The peer-reviewed foundation

The starting point is a published dataset rather than an in-house one, so the long-run headline does not rest on choices made in this office. LongPMInd (Wei and colleagues, *Earth System Science Data*, 2024) reconstructs daily PM2.5 for all of India on a roughly 10 km grid from 1980 to 2022. It is built with LightGBM, a gradient-boosted decision tree method in which the computer grows thousands of small yes-or-no rule trees, each correcting the last. It is trained on Central Pollution Control Board ground measurements, using satellite haze, NASA's MERRA-2 atmospheric composition and the European ERA5 weather reanalysis as inputs.

Its accuracy on the most demanding test available, predicting whole years it never saw, is about $R^2 = 0.66$, where $R^2$ is the share of real variation captured and 1.0 would be perfect. The full paper is reproduced as Annexure A of the long-record report, which is lawful because it is published under the Creative Commons Attribution 4.0 Licence.

### 9.2 An independent cross-check

Trusting a single product would be a weakness, so a second reconstruction was built independently, using Google Earth Engine and a random forest trained on Delhi's own ground stations. Tested the same demanding way it reaches the same 0.66. Two reconstructions, built by different teams with different methods, agree on both the level and the shape of Delhi's history.

### 9.3 De-weathering, and the CNG test

The de-weathering principle is the same one Part IV sets out for the daily cut, applied at monthly resolution: what would each month's pollution have been under average weather for that time of year, and what is left once the weather-explained part is removed.

The 2002 to 2003 CNG transition is tested by interrupted time series, which marks the policy date and asks whether the trend shifts or bends immediately afterwards, beyond the path it was already following. It is applied both to PM2.5 and to black carbon, the sooty and largely vehicle-derived particle that the switch was most expected to cut.

### 9.4 What it found, and what it cannot settle

Delhi's fine-particle pollution sat near 53 micrograms per cubic metre through the 1980s, which was already more than three times the WHO limit, then climbed steeply through the 2000s to a plateau

near 86 where it has remained. The de-weathered line sits almost on top of the reconstructed one, which is the visual proof that the rise came from emissions rather than from luck with the weather.

Three limits are stated in the report and belong here too. LongPMInd reports a 10 km area-average, which is naturally lower than a busy monitored junction, so the exact level shifts with the spatial question asked while the trend and the verdict do not. The reconstruction sees particles and not gases, so it cannot resolve the gas-specific gains in carbon monoxide, benzene and sulphur dioxide that were the CNG switch's real achievement, and those need separate ground data. And a reconstruction is a validated estimate rather than a measurement: it is the best available answer for the decades before the monitors, presented as an estimate and not as a reading.

# 10 Disclaimers

These are the limits of the method, gathered in one place rather than argued through the paper.

- The gas measurements, nitrogen dioxide in particular, are not treated as settled. Peer reviewed international work has documented unit inconsistencies in this data.
- Ozone is not used as an accountability claim. Ozone chemistry can move a number for reasons that have nothing to do with policy.
- Nothing is published on days when too few monitoring stations reported.
- Where a year-on-year difference sits inside its margin of error, the site says it is too close to call rather than claiming a result.

# 11 Part VIII. Literature

Nothing in this method is novel. That is deliberate. Every technique used here is established in the peer reviewed literature, and the sources are listed so a reader can go to them rather than take this paper's word.

## 11.1 De-weathering

The approach of removing meteorological influence from air quality series using machine learning, then resampling weather from a climatological distribution, follows Grange and colleagues.

- Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E. and Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics*, 18, 6223 to 6239.
- Grange, S. K. and Carslaw, D. C. (2019). Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment*, 653, 578 to 588.

The second of those is the closest published parallel to what this site does daily: it uses normalisation to ask whether a policy intervention left a real signal.

## 11.2 The learners

- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5 to 32.
- Wright, M. N. and Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data. *Journal of Statistical Software*, 77.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189 to 1232.

### 11.3 Air quality analysis in R

- Carslaw, D. C. and Ropkins, K. (2012). openair: an R package for air quality data analysis. *Environmental Modelling and Software*, 27 to 28, 52 to 61.

### 11.4 Uncertainty and multiple comparisons

- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1 to 26.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289 to 300.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703 to 708.

### 11.5 Standards, index construction and data

- Central Pollution Control Board (2014). *National Air Quality Index*. Report of the expert group constituted by CPCB, which defines the sub-index breakpoints and the averaging windows used here.
- World Health Organization (2021). *WHO global air quality guidelines: particulate matter, ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.*
- Hersbach, H. and colleagues (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999 to 2049.

### 11.6 A note on how to read this list

These are the sources the method rests on, cited as they are commonly known. A reader checking this paper should go to the original in each case rather than rely on the description here, which is the same standard this paper asks to be held to.

## 12 Part IX. Known limitations

- The de-weathering model explains a majority of variance for particulates and less for the gases. Every estimate carries that uncertainty and the intervals are published with it.
- The nitrogen dioxide unit question is unresolved and is disclosed on every report.
- The in-season retraining clause has not been exercised, so this paper describes a cadence it has observed and a permission it has not.
- The boosted-tree comparison is a single held-out evaluation repeated twice. It is evidence, not proof, and it has not yet changed anything published.

### 12.1 What would change our conclusions

If de-weathered year-on-year differences moved inside their intervals, the site would report no demonstrable change, and has done so before. If the nitrogen dioxide unit question were resolved against the current reading, the gas findings would need restating. Both are stated here so that a reader knows what would count against us.

## 13 Version history

- **Edition 1.0**, 10 August 2026. First complete edition. All nine parts written, including the literature and the network inventory.
- **Draft 0.2**, 10 August 2026. All parts drafted except the literature. Contents table made clickable, with a line describing each part.
- **Draft 0.1**, 10 August 2026. Parts I, IV, VII and IX drafted from verified sources.

This paper is versioned, and a change to the method changes this document. Where a figure here is later found to be wrong, the correction is recorded in this history rather than made silently.