Read the air record

Three quantities, with different meanings

How Hawa Ka Hisab measures and compares Delhi air

The record begins with measurements at monitoring stations. Code turns those readings into daily concentrations, an air-quality index and comparisons across periods. Models answer a different question: what concentrations they predict under a specified reference weather setting. None of these calculations directly measures a policy's effect on emissions.

This paper describes the current source implementation and the limits a reader needs to interpret it. A dated report can use a different period or method. Its own window, eligible stations, units, model version and uncertainty design govern its results. A change in code does not retrospectively certify every earlier report.

1. Three quantities that must remain distinct

Measured concentration is pollution recorded in the air at a monitor. Concentrations in micrograms per cubic metre, written as µg/m³, are not AQI points. **AQI** converts qualifying pollutant concentrations into sub-indices and uses the highest sub-index at a station. **Weather-normalised concentration**, also called weather-removed concentration, is a fitted model's prediction under a chosen weather reference.

A city average describes an eligible monitoring network. It is not every resident's exposure and does not mean all neighbourhoods have the same air. A worst-station value, a mean station AQI, a mean concentration and a population-weighted estimate answer different questions. Their labels and denominators must stay with their figures.

Quantity	What it describes
Measured concentration	Pollution recorded at an eligible monitor, in its stated unit.
AQI	An index calculated from qualifying pollutant concentrations.
Weather-normalised value	A model prediction under a specified reference weather setting.

Reading guide: data p. 2; clocks and AQI p. 3; display and population p. 4; models p. 5; reference weather p. 6; uncertainty p. 7; attribution p. 8; sources p. 9.

[1] Source: aqi_data.R, aqi_sidecar.R and daily weather source, checked 19 October 2026. These are three different quantities, not interchangeable measures.

Readings and missing data

Coverage and units belong with every figure

2. Readings, station identity and missing data

The principal hourly archive contains CPCB and DPCC station records compiled by the Centre for Research on Energy and Clean Air. Reports using another feed must identify that feed and its period. This office does not operate the monitoring instruments. The calculations and their errors remain this office's responsibility.

Records are matched by station identifier. A name change must not turn a monitor into a new station or join it to a different one. Timestamps are read as text and parsed as Indian Standard Time. Duplicate records do not create extra clock hours. Gas rolling means use a regular hourly grid so a gap remains a gap.

A daily pollutant value needs at least 16 distinct valid hours in its 24-hour AQI window. Carbon monoxide and ozone also require at least six available hours in a qualifying eight-hour block. The six-of-eight choice is this record's completeness convention; it should not be described as a separately established CPCB requirement. The eight-hour block must lie wholly within the documented-formula AQI window.

Value-range checks are analysis filters, not instrument detection limits. The weather-model code uses positive values below 1,000 for its targets other than carbon monoxide, whose upper analysis filter is 50,000. Carbon monoxide model targets are in μg/m³; NOx targets are in parts per billion. Units must be checked before pooling records. A rejected or missing reading is not a zero concentration. A missing eligible day is not a day of clean air.

The current weather producers prefer canonical units when those units occur for a species. Their fallback retains other units if no canonical-unit records occur. This fallback does not establish that a changed feed's units are comparable. Source and unit checks remain necessary for each accepted calculation.

[2] Source: aqi_data.R, lines 82-186; compute_dw_multipollutant_daily.R, lines 269-281, checked 19 October 2026. Range filters describe analysis choices, not sensor limits.

Daily clocks and AQI

A shared 4 PM label is not an identical clock mask

3. Daily windows and the AQI calculation

The current public daily AQI uses the 24-hour period from 4 PM IST on the preceding date to 4 PM on the report date. The AQI source includes the starting hour label and excludes the ending one. Its daily weather producer currently uses the opposite boundary convention: it excludes the starting label and includes the ending one. Both source predicates select 24 labels from complete hourly data, but only 23 labels are shared.

This boundary distinction is an open qualification issue for comparisons across those producers. The source audit establishes the different masks; it does not establish the numerical effect on a particular shipped report. A shared "4 PM" label alone does not prove identical clock slots. This paper does not silently change either producer.

Historical daily reports used a 10 AM window before the October change. Historical rewrites must print their actual window. Calendar-year studies and special comparisons have separate inclusion rules. Their dates must not be inferred from the live daily convention.

The documented-formula calculation first aggregates concentrations. It takes 24-hour means for PM2.5, PM10, nitrogen dioxide and sulphur dioxide, and the highest qualifying contained eight-hour mean for carbon monoxide and ozone. It then converts each concentration with CPCB's published breakpoint formula. A station AQI needs at least three qualifying pollutants, including PM2.5 or PM10. NOx is a model diagnostic, not an extra pollutant added to this public daily AQI roster.

The city AQI in the daily sidecar is the arithmetic mean of eligible station AQIs. Matched-station changes and population-weighted summaries are separate outputs. An index ceiling of 500 does not establish that an instrument cannot measure a higher concentration. Averaging concentration before conversion can differ from averaging hourly indices; these operations must not be treated as interchangeable.

Primary calculation sources: CPCB's calculation note and CPCB's AQI calculator. The record's AQI 500 investigation explains the index and concentration distinction.

Clock label	Daily AQI	Daily weather
Starting 4 PM	Included	Excluded
23 intervening labels	Included	Included
Ending 4 PM	Excluded	Included

Both predicates select 24 labels on complete data. They share 23; this diagram does not estimate the effect on reported values.

[3] Source: aqi_data.R, lines 162-178; daily weather source, lines 316-332, checked 19 October 2026. CPCB calculation note and calculator checked 18 October 2026.

Display and population

Selection rules and proximity estimates

4. What is displayed in the daily report

The daily report's weather-removed figures appear only when the city's weather-removed year-on-year change is positive. On such a day, a pollutant's weather-removed cell also needs that pollutant's change to be positive. This is a display rule, not an uncertainty test. An omitted cell is not a zero estimate, and the displayed cells are not a complete distribution of changes.

Annual summaries and special investigations state their own rules. A special study must retain material contrary results and uncertainty needed to assess its conclusion. A shorter public report can condense diagnostics, but it must not remove a material limitation. The daily film is also a selected account: Ajay Maken chooses its narratives and approves the final film. A narrative omitted from that film can still belong in the written record.

5. Population near monitors

The population total, the division of that total and its spatial placement are separate inputs. The current surface uses an official annual Delhi population projection, 2025 electoral shares to allocate the total among constituencies, and WorldPop 2026 R2025A v1 to distribute each constituency's allocation within it. The raster supplies relative spatial weights; its raw sum does not replace the adopted official city total.

Population near a monitor means residents within the stated proximity threshold. It does not mean the monitor represents only those residents, nor that everyone in that area received the same dose. Overlapping areas must not count the same residents twice. The current population headline uses monitors in the Poor or worse band, with effective AQI above 200. A day with no eligible Poor monitor is not relabelled with Moderate monitors to obtain a nonzero headline.

Population estimates are rounded down for the stated display unit. Missing required Poor-or-worse fields block that display rather than selecting another category. Historical estimates retain their dated population basis; they should not be presented as current-production estimates without recalculation. The separate population methodology and definitions gives the adopted sources and allocation formulas.

[4] Source: Daily report display source, lines 467-496, checked 19 October 2026. Population methodology published 18 October 2026; adopted official annual total, 2025 shares and WorldPop 2026 R2025A v1.

Models and validation

Different score designs answer different questions

6. Fitting the weather models

Separate random forests are fitted for PM2.5, PM10, nitrogen dioxide, NOx, sulphur dioxide, carbon monoxide and ozone. The current formula has 15 predictors: wind speed; sine and cosine of wind direction; temperature; boundary-layer height; precipitation; five lagged or accumulated weather measures; station identity; hour; weekday; and a time trend. Lagged pollutant concentrations are excluded from this formula. Their exclusion does not make the remaining signal a direct emissions measurement.

The current forest tuning grid tries three candidate-variable counts and two minimum-node sizes. Tuning forests have 500 trees; the final fitted forest has 800. The source divides chronologically sorted rows into five blocks and evaluates later blocks using earlier rows. This permits up to four forward validation folds for each configuration, subject to minimum-data and station-level checks. Rows sharing one timestamp can fall on either side of a row-count boundary.

The source's stable-selection rule can retain the previous parameter choice unless a challenger improves the current cross-validation score beyond the stated margin. A source configuration is not proof that every historical fit used it. Model versions must retain their training period, fitted settings and dated assessment.

The all-seven boosted-tree trainer defaults to 5,000 trees, depth six, shrinkage 0.01, bag fraction 0.5 and minimum node size ten. These can be overridden by its execution environment. A separate smoke-test trainer has different defaults and must not be substituted in the paper. The all-seven trainer uses a random 80/20 row split. That test can share stations and nearby times between training and testing; it is not an independent future-period test.

Forest out-of-bag scores, forest forward-validation scores and boosted-tree random-held-out scores are different assessments. A higher value under one design does not establish superiority under another. Similar training and held-out scores alone do not rule out temporal dependence or prove the validity of a retrospective policy comparison. The dated model-performance archive publishes these designs and scores separately. Its screening thresholds are not causal or deployment certificates.

Assessment	What is held out	Interpretation limit
Forest OOB	Rows not sampled into a tree	Other nearby times and stations remain in training.
Forest forward validation	Later row blocks	Up to four folds; tied timestamps can cross a row boundary.
Boosted-tree test	Random 20% of rows	Not an independent future-period forecast test.

Scores from these different designs cannot rank models as if they were one common test.

[5] Source: train_dw_pm25_z2.R, lines 145-154 and 371-421; train_dw_brt_all7.R, lines 64-68 and 240-269, checked 19 October 2026. Dated model archive: 18 October 2026.

Reference weather

What the daily source does and does not establish

7. What daily weather normalisation actually does

The standard daily implementation is configured for 64 prediction repetitions. It substitutes reference weather while holding station identity, hour, weekday and trend at the observed row's values. Its climatology is keyed by day of year and hour, with a circular seasonal smoothing window. Wind direction is treated circularly. Lagged-weather predictors are supplied at their climatological values.

The current trainer's climatology function writes mean fields, not the spread fields used by the daily sampler. When spread fields are absent, the daily normalisation function sets their standard deviations to zero. On that path, repeating a normal draw 64 times repeats the same weather values. Therefore a configuration of "64 draws" does not, by itself, establish 64 distinct weather realisations or adequate weather-reference uncertainty. A specific model package needs its own metadata check before making that claim.

If nonzero spread fields are supplied, the daily source draws several weather variables separately around their means, with wind handled in vector space and lagged weather held at its reference values. This is not the same method as resampling complete, jointly observed 24-hour weather trajectories. It does not automatically preserve every relationship among weather variables and their lags.

For carbon monoxide and ozone, the daily source averages hourly predictions across repetitions and then takes the largest qualifying eight-hour mean. Taking each draw's maximum first and then averaging is a different operation. With variable draws the two can differ. A study using the latter method must name it and must not claim to reproduce the former merely because both use 64 predictions.

Special retrospective calculations can specify a different, frozen weather reference and aggregation order. They require their own matched observations, model-byte provenance, coverage and uncertainty checks. No unfinished special calculation is certified by this paper, and no unavailable historical model bytes are claimed to have been reproduced.

[6] Source: Current forest trainer, climatology function, and daily weather source, lines 357-468 and 638-699, checked 19 October 2026. Source functions, not a live model-package schema audit.

Comparisons and uncertainty

A range covers its resampling design

8. Comparisons and uncertainty

The daily year-on-year baseline uses the corresponding date in the previous year with a seven-day band on either side. Matching by station identifier reduces changes caused solely by a different station roster. It does not repair missing months, different valid clock hours or changes in the underlying instruments. Every comparison needs its actual eligible stations, days and weighting.

The daily city bootstrap resamples station-level summaries. For percentage changes it resamples paired numerator and denominator station values and recalculates the ratio. The standard implementation uses 1,000 replicates and reports percentile ranges. This is conditional on the supplied station values and resampling design. It does not refit the pollution model, independently resample a future period or automatically account for spatial dependence among stations.

The daily station band is calculated from the hourly spread of the model predictions divided by the square root of the available hour count, with a normal multiplier. It is not a full model predictive interval. Consecutive hours can be dependent. For a gas maximum, a band derived from the hourly mean's variability is not a validated uncertainty interval for the maximum.

Calendar-block resampling used in a special investigation answers a different uncertainty question from a daily station bootstrap. Its block length, matched panel and calendar treatment belong with its reported ranges. A range crossing zero permits no change under that resampling design. If a conclusion changes across reasonable design sensitivities, that instability must be disclosed. None of these intervals has automatic unconditional coverage for all measurement, missing-data, model and reference-weather uncertainty.

Range	Variation represented	Important omitted uncertainty
City station bootstrap	Resampled station summaries and paired ratios	No model refit; spatial dependence is not automatically covered.
Daily station band	Hourly spread of fitted predictions	Not a full predictive interval or calibrated gas-maximum interval.
Special-study blocks	A stated calendar-block design	Depends on panel, block length and model/reference assumptions.

Each band represents its own source of variation. Combining labels does not make the intervals cover every uncertainty.

[7] Source: Daily weather source, lines 477-504 and 735-757, checked 19 October 2026. Its 1,000 station replicates and hourly station band have different interpretations.

Attribution and corrections

A model difference does not establish a cause

9. What the evidence can establish

A weather-normalised difference is descriptive and conditional on its model and reference weather. Models fitted using years included in a retrospective comparison have seen those target years. Their fitted-period results are not independent forecasts. Changes can reflect emissions, chemistry, measurement practices, station coverage, model specification and aspects of weather not adequately represented by the chosen reference.

Attributing a change to one intervention requires a stated causal design and credible assumptions, including an appropriate counterfactual and checks for competing changes. A remaining model residual, a long run of positive signs or a smooth historical chart does not provide that design on its own. Ozone requires particular care because its chemistry does not give a simple one-to-one emissions interpretation.

The long-record report separates earlier model-based estimates from the later observing record. Its observing-system change around 2000 remains a limitation. The corrected report covers 1980 to July 2026; the earlier withdrawn report must not be treated as current evidence.

10. Code, language models and corrections

R and Python code produce the numerical outputs. Language models can assist in writing code and explaining findings. Their generated prose is not a measurement, and a numerical statement is not validated merely because software printed it. Source hashes, immutable model versions, dated output checks and independent arithmetic and coverage checks provide separate evidence.

Language assistance and a human film approval do not replace scientific qualification. Automatic expert panels are not part of the daily run; any requested expert review has a separately defined scope. Public reports need their own content checks, visual inspection of every final page and verification of the exact published bytes.

A correction should identify the affected report, the changed method or value, and its effect on interpretation. Original correspondence and previously sent reports are preserved. A corrected website entry is not authority to replace an as-sent document, republish a held report or send messages to another recipient.

[8] Source: Implementation and authority record checked 19 October 2026. Corrected long record checked 18 October 2026, covering 1980 to July 2026 and retaining the observing-system seam.

Sources and method record

Checked source descriptions, with retained limits

11. Source and version record

These calculation descriptions were checked on 19 October 2026 against the current daily AQI aggregation, sidecar and breakpoint source; the random-forest and all-seven boosted-tree trainers; both daily weather producers; and the current population surface. The dated model and population records are linked below. Small source fixtures demonstrate the identified algorithms, not the validity of a fitted pollution model.

The source-conditional weather description and unmeasured boundary-hour impact remain explicit limitations. No live model package was opened to inspect its climatology in this source audit. No numerical correction to shipped daily figures is claimed. This paper does not replace the dated methods, coverage and validation required for each report.

The dated implementation record lists source fingerprints, the checked algorithm descriptions and their limits. Fingerprints identify the source checked; they are not a complete public raw-data and executable-code archive. This paper does not certify unfinished historical recalculation or special-study results.

Public source links

CPCB calculation note

CPCB AQI calculator

Dated model-performance records

Population allocation and proximity definitions

Corrected 1980 to July 2026 long record

Index ceiling and concentration investigation

Responsibility for this explanation and its errors rests with the publisher. Dated report methods and the retained limits control the interpretation of each result.

[9] Source: Source implementation record dated 19 October 2026; cited public model and population records dated 18 October 2026. Primary calculation documents linked on this page.