

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

A QUALITY-CONTROLLED PM_{2.5} DATASET WITH FROZEN CROSS-CITY EVALUATION SPLITS FOR SIX CENTRAL ASIAN CITIES

Jaloliddin Musayev ^{1,*},
Asadbek Abdivayitov ²

¹ International House Tashkent Academic Lyceum, Tashkent, Uzbekistan

² First Specialized Boarding School, Karshi, Uzbekistan

* Corresponding author: jaloliddin2009applicant@gmail.com

ORCID iDs: Jaloliddin Musayev 0009-0003-0210-3687; Asadbek Abdivayitov
0009-0006-3484-3438.

Abstract

Central Asia is among the most polluted inhabited regions on earth and among the least monitored. Gridded products already assign PM_{2.5} across the region, but no open station-level benchmark exists to check them against, which this dataset provides.

It covers daily PM_{2.5} from 7 instruments in 6 cities, 2018-11-27 to 2024-12-31, screened by seven pre-registered quality rules and one added after inspection, with 5 reference monitors and 2 low-cost sensors labelled as such. Each station-day carries satellite, chemistry-transport and static predictors with measured latencies. The deposit holds the frozen splits, checksums and the v1.1.0 reference results. The observations are rebuilt by script from OpenAQ.

Splits were frozen and checksummed before the reported results existed. Leave-city-out over 6 folds withholds a city entirely, so error is measured where no monitor operates, against the same six baselines for every method. Over the six held-out cities the tuned retrospective-tier reference model reaches 28.01 $\mu\text{g}/\text{m}^3$ fold-mean RMSE and is not separable from pooled-debiased CAMS (29.77 $\mu\text{g}/\text{m}^3$, paired t $p = 0.1392$).

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

1. Introduction

Tursumbayeva et al. (2023) put annual PM_{2.5} in six Central Asian capitals at 4.3–12.6 times the WHO 2021 guideline, and attribute most of that burden to coal combustion, not transport, against official emissions inventories. Source-apportionment studies reach the same conclusion independently in Kazakhstan (Tursun et al., 2025) and Tajikistan (Papagiannis et al., 2024). The monitoring base beneath those numbers is thin and unevenly open. Turkmenistan operates no national network, Kazakhstan releases data only to users physically inside the country, and only Kyrgyzstan publishes in a fully open form (OpenAQ, 2025). Global gridded products (van Donkelaar et al., 2021) already assign PM_{2.5} values across Central Asia and the epidemiological literature consumes them, but nobody can check those values against ground observations, or score two methods on the same terms, because the region has no open station-level benchmark.

The objective of this work is to supply that benchmark. We fixed, before producing the reported results, which cities and days a method is evaluated on, what it is compared against, and how the comparison is scored. The evaluation withholds an entire city at a time, so a score answers the question the region actually faces: what is the PM_{2.5} concentration in a city where no monitor has ever operated? Methods are scored on those withheld cities against a mandatory ladder of constant and spatial-interpolation baselines, and against a bias-corrected CAMS forecast, the comparator on which the primary inference is run.

Why the protocol matters can be seen in the closest published analogue. Jin et al. (2022) estimate PM_{2.5} across Xinjiang, an arid, sparsely monitored region like this one, and report R^2 between 0.73 and 0.81 under 10-fold cross-validation with no spatial stratification. With 41 stations in 16 cities and 8-day averaging, that design places observations from the same station, often from the same window, on both sides of the split. The resulting figure describes interpolation among stations the model already holds. It says little about estimation where no monitor

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

exists, and a reported R^2 does not separate the two cases. Blocked validation is the established remedy for structured data (Roberts et al., 2016), for spatially derived predictors (Meyer et al., 2019) and for particulate data specifically (Alazmi and Rakha, 2022), and validation strategy is named among the field's systematically overlooked issues (Tang et al., 2024). No benchmark has enforced it for this region.

AQ-Bench (Betancourt et al., 2021) is the nearest precedent, with 5,577 stations worldwide split by spatial clustering at a 50 km threshold. It targets long-term ozone metrics from station metadata in a time-independent regression, and it excludes Central Asia. We adopt its spatial clustering rationale. Pollutant, target, temporal protocol and region differ. AirDelhi (Chauhan et al., 2023) is a second precedent for fine-grained particulate benchmarking, confined to a single city.

The benchmark covers 7 instruments across 6 cities, spanning 2018-11-27 to 2024-12-31. Splits were frozen and hashed before the reported results were produced, and a test checks the committed file against that digest on every run, so a split cannot change without a version raise. Leave-city-out withholds one city at a time over 6 folds. Blocked-temporal separates training from test in time, with a purge of 240 hours, wide enough that no training row's leave-city-out feature window reaches the block that follows. A variable unknowable at prediction time cannot enter a configuration presented as deployable, because every predictor carries a measured acquisition latency.

One fold is worth singling out. Khujand's two sensors begin on 2023-11-28 and 2023-12-06, after the training block closes, so the Khujand fold is zero-shot: the model that scores it has seen no Khujand label. Every fold refits on training and validation rows together, and the validation block ends on 2023-12-21, so at most the first weeks of Khujand's record enter the five other folds' pools. We report the primary inference both with and without Khujand (Statistical validation).

On these splits the tuned reference model scores $28.01 \mu\text{g}/\text{m}^3$ fold-mean RMSE against $29.77 \mu\text{g}/\text{m}^3$ for pooled-debiased CAMS, a difference six cities cannot

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

separate (paired t $p = 0.1392$), and any method producing a PM_{2.5} estimate can be scored on the same terms.

2. Materials and methods

Study region and period

The benchmark covers 6 Central Asian cities (Almaty, Ashgabat, Bishkek, Dushanbe, Khujand and Tashkent) over 2018-11-27 to 2024-12-31 (Figure 1). Cities were not selected: they are every city in the region with an openly published PM_{2.5} record meeting the inclusion rules below.

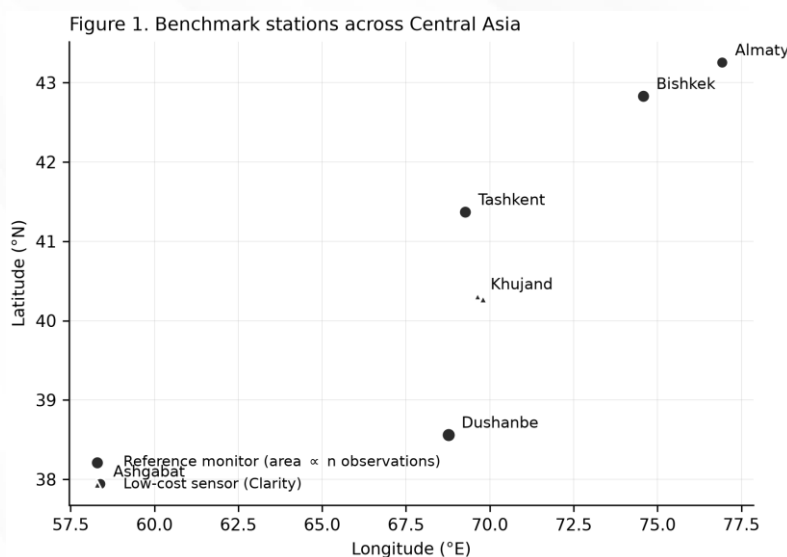


Figure 1. The 7 benchmark instruments across 6 Central Asian cities, marker area proportional to each station's observation count. Marker style distinguishes the 5 US diplomatic-post reference monitors from the 2 Clarity low-cost sensors at Khujand. Leave-city-out withholds every instrument in one city at a time, so the spacing between cities sets the extrapolation distance each fold demands.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Ground observations

Hourly PM_{2.5} was retrieved from the OpenAQ v3 archive for all stations within the region's bounding box. 317 candidate stations were assessed. Inclusion required, as pre-registered before the data were inspected:

- **Q7 span and completeness.** At least 2 years of record and 60% completeness within it. This rule excludes 306 stations, almost all low-cost units, and Astana at 42.8% completeness.
- **Q1 physical range.** Values outside $[0, 1000]$ $\mu\text{g}/\text{m}^3$ are masked.
- **Q2 flatlining.** ≥ 24 consecutive identical non-zero values are masked.
- **Q3 zero runs.** ≥ 6 consecutive zeros are masked.
- **Q4 unit sanity.** A station median outside $[1, 500]$ $\mu\text{g}/\text{m}^3$ rejects the series, catching mg/m^3 reported as $\mu\text{g}/\text{m}^3$ and AQI values reported as concentrations. Every decision taken under these rules records its effect on n and its direction of bias if wrong (data/DECISIONS.md).

Duplicate resolution (Q5). *Q5a*: one identifier at more than one coordinate rejects the series. *Q5b*: identifiers within 150 m are one instrument, detected by single-link clustering over the haversine distance, and merged. *Q5c*: any pair bit-identical on more than half of overlapping samples is one instrument regardless of separation. *Q5c* was added after the data were inspected, when *Q5b* passed a pair that proved to be one instrument (Results).

We merge by precedence and gap-fill, never averaging: the feed with more observations is primary, the other fills only the hours the primary lacks, and per-hour provenance is written to `panel_sources.parquet` by the panel rebuild. Three pairs merge under this rule (Bishkek, Ashgabat and Dushanbe), giving 7 instruments across 6 cities, each merged instrument carrying its city name as `station_id`.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Q6 timezone verification. Each station's diurnal composite is cross-correlated against a regional reference, and a station whose lag cannot be identified is flagged (Results).

Daily target. The prediction target is the local-calendar daily mean, requiring at least 18 hourly observations. Days are local-calendar because a UTC boundary falls mid-night in Central Asia, so an overnight peak would be divided between two calendar days and would lower both daily means.

Predictor sources

Each predictor carries a measured acquisition latency and a typed availability flag, so anything that cannot exist at prediction time is excluded from deployable configurations by an automated test. Three of five initial latency estimates proved wrong when measured, one by 774 days.

- **Satellite.** Sentinel-5P CO, NO₂, SO₂ and absorbing aerosol index (Veefkind et al., 2012); MODIS MAIAC AOD (Lyapustin et al., 2018). Missingness in these products is correlated with the target: SO₂ retrieves on 0.1% of December days against 91.0% in July. We attribute that floor mainly to solar geometry, with winter cloud and snow cover likely contributing, and it blinds the direct winter coal tracer through the coal season. A retrieval-validity feature (*_valid_px) was therefore computed for every product and carried, then excluded from Task N under Freeze 2 (Results). It remains available for Task F. Null rows are kept, because interpolating across a systematically missing extreme tail would invent the values that matter most.

- **Chemistry transport.** Copernicus CAMS global PM_{2.5} forecast, used both as a feature and, bias-corrected, as a baseline.

- **No meteorological predictor.** ERA5 single-level fields (Hersbach et al., 2020) were retrieved and latency-tested. At 163 h they cannot enter the deployable tier, and the retrieval was stopped before a multi-year archive existed, so they enter

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

neither tier: the reference model carries no boundary-layer height, wind, temperature or humidity term.

• **Static geography.** Elevation, terrain basin indices at multiple radii, VIIRS night-time lights, and population density.

Split construction

Splits were built by `src/ecopulse_ca/splits/builder.py`, hashed and committed, and a test checks the committed file against that digest on every run (Split integrity).

Temporal blocks. Train, a purge, validation, a second purge, test, and a reserved post-test block. The test block is calendar 2024, the last full year of reference coverage, since every StateAir feed stops on 2025-03-04, when that publication channel closed. Each purge is 240 hours, derived from the feature and horizon bounds, $\text{purge_hours} = \text{max_lag_hours} (168) + \text{max_horizon_hours} (72)$, so that no training row's feature window can reach the block that follows. The relation is stated in `config.purge_rule`.

`max_lag_hours` bounds the features admitted under leave-city-out, the protocol behind every headline number here. Task F additionally admits a station's own history, including a 30-day rolling mean whose 720 h window crosses the purge into the preceding block, as a forecaster at a monitored station genuinely would, so Task F figures carry a weaker separation guarantee than Task N figures and the two are never compared.

Leave-city-out (6 folds). Each fold withholds one city entirely, every station in it, for the whole record. This is the protocol the benchmark is built around, because it is the only one whose error is measured where no monitor exists.

Spatial blocking is contested, and the objection applies to a different estimand. Wadoux et al. (2021) show that spatially blocked cross-validation is pessimistically biased for *map accuracy* over a fixed population, where design-based validation on a probability sample is the correct estimator. The quantity

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

here is the error incurred in a city that contributed no training row, for which withholding the city is the estimand itself. Meyer and Pebesma (2021) frame the complementary question, whether a held-out city falls inside the model's area of applicability.

Leave-station-out (2 folds). Only Khujand holds more than one instrument, so both folds are its low-cost pair, and the builder records the other five cities as ineligible in `leave_station_out.json`. The Dushanbe merge removed the only reference-grade city with two devices. This protocol therefore says nothing about generalisation across the reference network and supports no headline claim, and leave-one-out protocols carry a documented failure mode of their own (Austin et al., 2025). The folds are retained for methods that target low-cost sensor transfer. Reference implementation

Three feature tiers are defined by availability at prediction time (`static_only`, `deployable` and `retrospective`), so that a retrospective number can never be mistaken for a deployment claim.

We define two tasks separately and never pool them. **Task N** (nowcasting at unmonitored sites) withholds whole cities and admits no local history. **Task F** (forecasting at monitored stations) permits the station's own lagged observations. As implemented here it predicts the next-day daily mean from lags of 1, 2 and 7 days plus 7- and 30-day rolling means, and is therefore single-horizon.

The reference model is LightGBM. We select hyperparameters by grid search over 16 combinations on the validation block only, refit the selected configuration on train plus validation with the purge block withheld, and score it once on the test block. Every configuration runs 5 seeds (0, 1, 2, 3, 4) and is reported as mean $\pm$ standard deviation.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Baselines. Baselines are task-specific. Task N is scored against six rungs, three constants (the training-pool mean and the global mean and median), nearest monitor, inverse-distance weighting and ordinary kriging, tabled at daily resolution in Section 4 (t3_06, which also carries the oracle_city_constant diagnostic, flagged legal = False and not a rung). Task F has its own hourly ladder of four history rungs, persistence, diurnal persistence, a same-hour 7-day mean and climatology (t3_01), at 24, 48 and 72 h, which the daily, single-horizon Task F reference model is not scored against. CAMS is tabled separately in three variants, raw, locally debiased and pooled-debiased (t4_01). The pooled-debiased variant, whose bias correction is fitted on the training block only with the held-out city excluded from its own correction, is the comparator for the primary statistical test and is not a rung of either ladder.

Statistical methods

The primary metric is the root mean squared error of the daily mean, in $\mu\text{g}/\text{m}^3$, computed per held-out city and averaged over folds. Model-versus-baseline comparison uses the Diebold–Mariano test on squared-error differentials with Newey–West HAC variance and the Harvey–Leybourne–Newbold correction. Because the loss differential is serially correlated within station and station-days cluster within cities, the primary analysis aggregates to one value per city and tests those 6 observations, with sensitivity analyses across three HAC truncation lags (all deposited in t6_06_significance.csv, the longest tabled), a cluster bootstrap and an exact sign-flip permutation test whose attainable floor is stated. Per-fold tests are Holm-corrected (Holm, 1979) and read as descriptive, since both comparators are estimated models (Diebold, 2015).

Software and generative-AI assistance

Analysis code is Python 3.12, with exact dependencies and versions in Code availability. Generative-AI assistance is declared in full under *Use of generative AI*.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

3. The released benchmark

Coverage of the 7 instruments over the record is shown in Figure 2. The benchmark is deposited as a single versioned archive, `eco-pulse-ca v1.1.0.splits.sha256` carries the digest of `splits.json`, the file to cite. The three split sidecars are projections the builder writes from the same payload and are not separately digested.

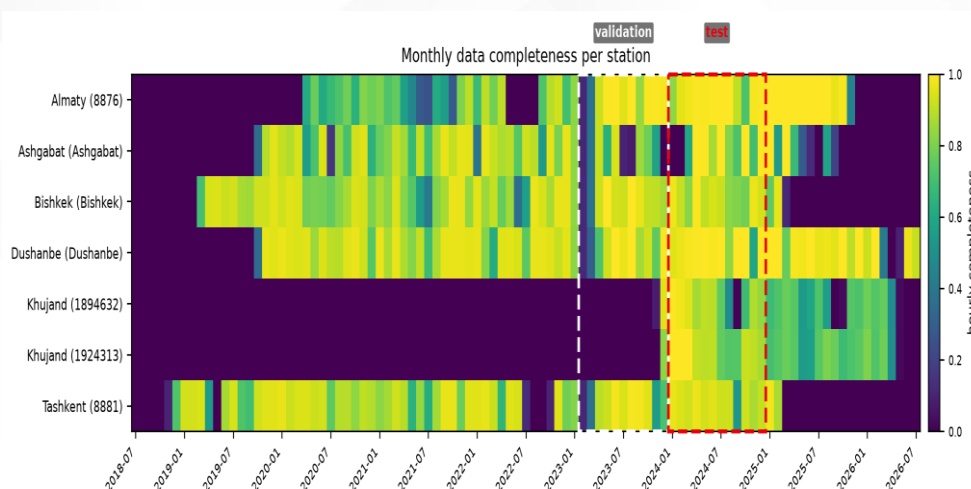


Figure 2. Monthly hourly-completeness for each of the 7 benchmark stations. Dashed boxes mark the validation (2023-01-11 to 2023-12-21) and test (2024-01-01 to 2024-12-31) blocks. Both Khujand sensors begin on 2023-11-28 and 2023-12-06, after the training block closes, and the record ends unevenly, with 2 of 7 stations, Bishkek and Tashkent (8881), stopping at the StateAir closure on 2025-03-04 while the remainder continue through a longer-lived feed. All of that lies beyond the test block and none of it is used.

The frozen benchmark definition

These five files, under `benchmark/splits/`, are the benchmark (Table 1), and nothing else in the archive is needed to evaluate a method on it.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Table 1. The frozen benchmark definition.

File	Format	Contents
splits.json	JSON	Complete benchmark definition: version, stations, temporal blocks, both fold sets, protocol configuration.
temporal_blocks.json	JSON	Train, purge, validation, purge, test and reserved blocks with UTC bounds.
leave_city_out.json	JSON	The 6 spatial folds, each naming held-out city, held-out stations and training stations.
leave_station_out.json	JSON	The 2 within-city folds, plus the five cities named ineligible.
splits.sha256	text	SHA-256 of splits.json, with the freeze timestamp.

splits.json holds a stations table (7 records: station_id, city, latitude, longitude, n_observations), the six temporal_blocks in UTC, the 6 leave_city_out and 2 leave_station_out folds with the ineligible-city list, and a config block (max_lag_hours, max_horizon_hours, purge_hours, test_year, seeds, purge_rule). Station identifiers are OpenAQ location_id values as strings, except that a merged pair carries its city name (Materials and methods).

Reference results

Table 2 lists the reference results, provided so that a new method can be placed on the same axes without re-running the reference implementation. All are CSV with a header row, under paper/tables/.

Table 2. Deposited reference result files.

File	Rows	Contents
t3_01_task_f_baselines_hourly.csv	60	Task F baseline ladder, hourly, by station and horizon.
t3_02_task_n_baselines_hourly.csv	28	Task N baseline ladder, hourly, by station, with exceedance metrics.
t3_06_task_n_baselines_daily.csv	49	Task N daily ladder on the learned models' evaluation rows: the table to compare a daily model against.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

File	Rows	Contents
t4_01_cams_baseline_variants.csv	21	CAMS raw, locally debiased and pooled-debiased, per station.
t5_01_loco_untuned.csv	108	Untuned gradient boosting, leave-city-out, per fold and seed.
t5_02_loco_tuned.csv	123	Tuned gradient boosting, all three tiers, per fold and seed, with selected hyperparameters; also carries the Task F rows.
t5_07_missingness_test.csv	60	Freeze-2 ablation on the test block, with and without retrieval-count features, per fold and seed.
t6_01_predictions_task_n.csv	2,214	Row-level test-block predictions: station_id, date, pm25, lgbm, lgbm_seed0-lgbm_seed4, pooled (debiased CAMS), fold.
t6_02_dm_lgbm_vs_cams.csv	7	Diebold-Mariano per fold and pooled.
t6_06_significance.csv	7	Primary and sensitivity inference with unit, statistic, p and CI.
t6_07_per_fold_holm.csv	6	Per-fold p -values with Holm step-down correction.
t7_06_leave_khujand_out.csv	2	Primary inference over the five reference-grade cities.

The v1.1.0 archive was built on 2026-08-14 and predates two of the tables above. t5_07_missingness_test.csv and t7_06_leave_khujand_out.csv were produced afterwards and are held in the GitHub repository; a later deposit will carry them.

Scoring a method

Read splits.json, which is self-contained. Fit only on the rows a fold permits: under leave-city-out no observation from the held-out city may enter training, including indirectly through a neighbour feature or a bias correction estimated on that city. Predict on the test block (2024-01-01 to 2024-12-31), score the local-calendar daily mean over at least 18 hourly observations, and compare against t3_06_task_n_baselines_daily.csv, which is scored on the same rows at the same resolution. State the task and the tier, and never pool tasks or quote a retrospective number as a deployment claim. Report per city with dispersion over at least the 5

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

seeds used here (0, 1, 2, 3, 4), state the unit of analysis for any significance claim, since station-days are not independent, and give the truncation lag of any Diebold–Mariano test with a sensitivity sweep. The reference implementation is held to the same conditions.

t6_01_predictions_task_n.csv is the most reusable record: it permits any alternative loss, significance procedure or aggregation to be applied to the reference implementation without retraining it, and its per-seed columns make the ensemble decomposable. The archive also holds the per-fold, per-band and per-season error tables (t7_01–t7_03), discussed in Section 5, and the ranking-robustness table (t7_05), cited in Section 4. Neither the hourly panel nor the predictor matrix is deposited: both are rebuilt by the pipeline from the sources named under Predictor sources (Data availability). panel_sources.parquet, the per-hour provenance of each merge, is likewise written by the rebuild and is not in the archive.

4. Results

Observation quality control

The reference-grade observations originate with the US EPA AirNow programme, whose *AirNow Data Exchange Guidelines* (August 2025) state that its observational data "are not fully verified or validated" and "should be considered preliminary". Fully validated data in EPA's Air Quality System were not used, so every result here inherits that status. Seven rules (Q1–Q7) were declared before the data were inspected (Section 2; data/DECISIONS.md), each recording its effect on n and its direction of bias if wrong. Duplicate identity and timezone verification each changed the benchmark, and Q5c was added after inspection.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Duplicate identity. The US embassy monitors are published twice, by StateAir and by AirNow, under separate identifiers. Where the two records agree on position (Bishkek, 57 m; Ashgabat, 40 m) a 150 m co-location rule catches them, and where they disagree it does not. Dushanbe's two records sit 6.06 km apart and are one instrument: over the 33,462 hours in which both report, 94.0% of readings are bit-identical and 99.9% of the remainder match exactly at a five-hour offset, the same measurements timestamped in local time (Dushanbe is UTC+5), so 99.99% of overlapping hours are the same reading. The comparable Khujand pair, 14.4 km apart, is bit-identical on 0.3% of hours.

We therefore added a value-identity rule (Q5c): flag any pair bit-identical on more than half of overlapping samples, regardless of separation. Independent instruments would not agree to floating point, and the threshold sits midway in a measured $36\times$ gap between coincidence (2.6% for unrelated pairs) and duplication.

The Dushanbe records were merged under the precedence-and-gap-fill rule in Section 2, never averaging, because averaging two copies of one measurement fabricates a third value where they differ. The merge is index-aligned and does not realign the hours the secondary feed stamps in local time, so a gap filled from that portion lands five hours late: of 2,781 hours filled from the secondary feed, 56 are such displaced copies, 0.11% of the merged record, the last on 2023-10-31, so none falls in the test block. Merging changed every fold's RMSE. The decision, its effect on n and its direction of bias are logged as D-012 (data/DECISIONS.md). The benchmark holds 7 instruments across 6 cities, and the version was raised to 1.1.0.

Timezone correctness. Each station's diurnal composite is cross-correlated against a regional reference (Section 2, Q6). An initial implementation rejected both Khujand sensors for an apparent 12-hour shift, an artefact: the reference self-correlates at $r = +0.71$ under a 12-hour rotation, because Central Asian urban

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

PM2.5 is bimodal with peaks roughly half a day apart. The check now flags a station whose lag cannot be identified. That change from rejection to flagging is logged as D-006 and recorded as a post hoc revision in D-007 (data/DECISIONS.md). Only Khujand holds more than one instrument, so a constant, lifelong offset at a single-instrument city is undetectable by any check in this suite.

Instrument grade

5 of 7 instruments are US diplomatic-post monitors published by AirNow or StateAir (is_monitor = true in the OpenAQ census): BAM/FEM-class beta-attenuation instruments under the programme's documented QA regime, and the region's only consistent multi-country reference. The 2 Khujand instruments are Clarity Node-S low-cost optical sensors (is_monitor = false), which use light scattering, not gravimetric-equivalent measurement, and carry the humidity- and composition-dependent uncertainty documented for that class (Zheng et al., 2018). Their readings enter as published, because Khujand has no co-located reference against which a local correction could be fitted. The feed carries the value OpenAQ publishes for each sensor's pm25 parameter. Whether that is Clarity's raw optical mass or its vendor-calibrated output is not recorded in the metadata we retrieved, and no humidity correction was applied here.

Khujand is also the zero-shot fold. Its sensors begin on 2023-11-28 and 2023-12-06, after the training block closes, so the city contributes no row to the training block and none to its own fold. The 36 Khujand station-days inside the validation block, which ends on 2023-12-21, do enter the other five folds' tuning and refit pools. Inside the benchmark window the two sensors cover 1.09 and 1.07 years, so both meet the pre-registered 2-year span rule only because we count their observations past the window's end. We kept them for coverage of a sixth city.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Split integrity

A test checks the committed splits.json against splits.sha256 on every run, by canonical form and by raw bytes, so a changed split forces a version raise, a reason in data/DECISIONS.md and regeneration of every published number. The purge is verified arithmetically at both boundaries, $\text{purge_hours} = 240 = \text{max_lag_hours} (168) + \text{max_horizon_hours} (72)$. Assertions check that no fold trains on its held-out city or station, that every evaluated row falls inside the test block and that each fold evaluates only its own city, and a further test parses the tuning function and fails if it references the test block.

Reference implementation behaviour

These results characterise the task and are not a claim of method performance. Table 3 scores the ladder at a single temporal resolution on the frozen test block, over 5 seeds.

Table 3. Leave-city-out daily RMSE, reference model and baselines.

Model (leave-city-out, daily)	RMSE $\mu\text{g}/\text{m}^3$
nearest_monitor	33.50
training_pool_mean	32.75
train_global_mean	32.70
train_global_median	30.99
ordinary_kriging	29.75
idw_k5_p2	29.44
LightGBM, retrospective (log target)	28.01

The tuned reference model leads every admissible baseline on the fold mean, $28.01 \pm 0.35 \mu\text{g}/\text{m}^3$ against $29.44 \mu\text{g}/\text{m}^3$ for the strongest admissible rung (idw_k5_p2), and explains little within-city variation. The ordering is not robust across cities: the paired difference over 6 cities is not separable ($p = 0.586$), per-fold differences span -10.73 to $+5.10 \mu\text{g}/\text{m}^3$, and removing one city reverses the

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

lead in one of 6 subsets. Seed noise is not the cause: the margin is $4.1 \times$ the seed standard deviation (`t7_05_ranking_robustness.csv`). A constant equal to the held-out city's own test-block mean scores $28.12 \mu\text{g}/\text{m}^3$ but uses test labels, so it is a diagnostic floor and not an admissible rung. Mean per-fold R^2 is -0.04 , spanning -0.55 to 0.52 with 3 of 6 folds negative, against a pooled R^2 of 0.13 over the global mean, so the pooled figure alone would substantially overstate what the model does.

Exceedance F1 has a high floor: 4 of 6 cities clear the WHO 24-hour guideline on most test days, from 88% at Dushanbe down to 24% at Bishkek, so a classifier that always predicts "exceeds" scores $F1 = 0.741$ at a base rate of 61.8%. Peirce skill score is reported alongside because it is zero for that classifier by construction.

Error structure across cities

Figure 3 plots each held-out city's RMSE and mean bias against its observed mean concentration.

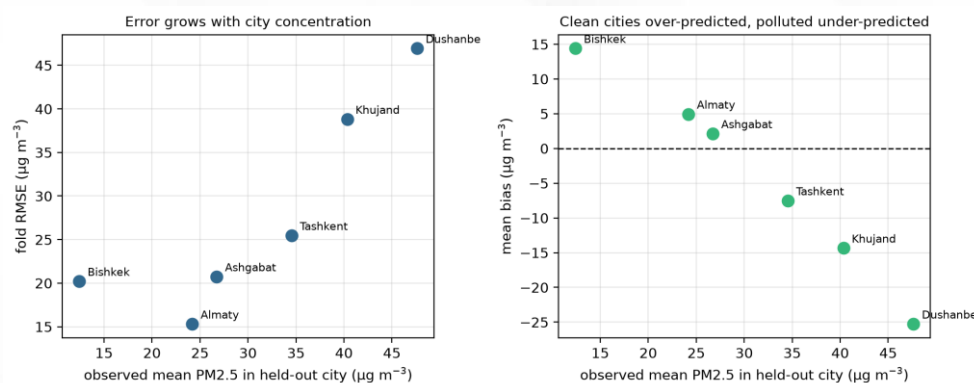


Figure 3. Leave-city-out error against the held-out city's mean PM2.5. Left: fold RMSE rises with city concentration (Spearman $\rho = 0.94$). Right: mean bias falls monotonically from $14.4 \mu\text{g}/\text{m}^3$ in Bishkek, the cleanest city, to $-25.3 \mu\text{g}/\text{m}^3$ in Dushanbe, the most polluted ($\rho = -1.00$).

The two panels are not equally strong evidence. RMSE scales with the variability of whatever is predicted, so the left panel is partly a scale effect: dividing each fold's RMSE by that city's observed standard deviation leaves no monotone

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

relation with concentration ($\rho = -0.03$). The bias panel is monotone across every fold ($\rho = -1.00$) because the predictions are flat: predicted city means span only $6.77 \mu\text{g}/\text{m}^3$, from 22.37 to 29.14, while observed city means span $35.23 \mu\text{g}/\text{m}^3$, from 12.42 to 47.64. A predictor that flat is biased upward in clean cities and downward in polluted ones by arithmetic, and over six folds that arithmetic will tend to produce a perfect rank correlation. The finding is therefore the flatness: transferred to a city it has never seen, the model returns something close to a regional level and does not move it with the city. The same pattern holds within the concentration range, with bias $10.1 \mu\text{g}/\text{m}^3$ on days below the WHO 24-hour guideline and $-90.4 \mu\text{g}/\text{m}^3$ above six times it, where RMSE reaches $100.9 \mu\text{g}/\text{m}^3$ on the 6.6% of rows in that band, and winter (DJF) RMSE is $51.0 \mu\text{g}/\text{m}^3$ against $16.0 \mu\text{g}/\text{m}^3$ in summer.

One candidate cause is internal to the model. It fits $\log_{1p}$ of the daily mean and inverts with $\exp_{m1}$ and no retransformation correction, so what it returns is closer to a conditional median than a conditional mean. For a right-skewed target that alone biases predictions low, and most where the tail is heaviest, in polluted cities, on extreme days and in winter, which is where the negative bias here is largest. Refitting with a smearing correction after the test block has been seen would make the test set a selection criterion, which the frozen protocol forbids, so the correction is deferred to a later version and this remains a candidate contributor, not an established cause. It cannot by itself explain the positive bias in the cleanest cities.

Statistical validation

The estimand is the reduction in squared error at a city with no local training labels, so the unit of generalisation is the city, and aggregating to one value per city gives 6 observations. The quantity tested is the loss differential Δ in $(\mu\text{g}/\text{m}^3)^2$, the model's squared error minus that of pooled-debiased CAMS, which scores $29.77 \mu\text{g}/\text{m}^3$ fold-mean RMSE against the model's 28.01. Negative values favour the model, and Table 4 reports the primary and sensitivity tests.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Table 4. Primary and sensitivity inference against pooled-debiased CAMS.

	Test	Unit	n	Δ (95% CI)	p
Primary	paired t on city means	city	6	-96.2 (-237.0, +44.5)	0.1392
Primary	exact sign-flip permutation	city	6	-96.2	0.1250
Sensitivity	station-day, independence assumed	station-day	2214	-101.7 (-133.3, -70.2)	2.6e-10
Sensitivity	station-day, Newey–West HAC (lag 60 d)	station-day	2214	-101.7 (-171.7, -31.7)	0.0044
Sensitivity	cluster bootstrap over cities	city	6	-96.2 (-196.5, -2.9)	0.0428

The interval, $(-237.0, +44.5)$ ($\mu\text{g}/\text{m}^3$)², spans zero and is wide enough to contain both a substantial improvement over CAMS and a moderate degradation, so with 6 cities the study establishes neither that the model is better nor that it is not, and $p = 0.1392$ alone would invite a reading of no effect that the interval does not support. The station-day rows show what an independence assumption would return and are not the inference we rely on, because the loss differential is serially correlated within station (first-order autocorrelation 0.25). With 6 clusters the cluster-robust variance estimator is downward-biased, as it is below roughly 30–50 clusters (Cameron and Miller, 2015), so an exact permutation test is reported alongside. Its smallest attainable two-sided p -value is 0.03125, a floor imposed by 6 cities and not evidence. Per-fold Diebold–Mariano tests are Holm-corrected (Holm, 1979) for 6 comparisons, and 3 of 6 survive at $\alpha = 0.05$.

Khujand carries 26.7% of the pooled rows as the only two-station city, and it is also the only city with low-cost labels. `t7_06_leave_khujand_out.csv` recomputes the primary inference without it, and the null persists, with paired t $p = 0.1392$ over all 6 cities against 0.2165 over the five reference-grade ones.

The evidence does not support a claim that the reference model outperforms bias-corrected CAMS under the unit of generalisation this benchmark is built around.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Frozen configurations and their test-block outcome

The test block was scored under two frozen configurations, both selected on the validation block without reference to test performance.

Freeze 1: target transform. Daily PM_{2.5} here has skew 2.79 and excess kurtosis 13.5, so squared error on the raw scale is dominated by a few extreme days. Four formulations (raw, log_{1p}, and residual-against-interpolator variants of each) were compared across three model families on validation only, log_{1p} won for every family, and once frozen and scored on test the fold-mean RMSE fell from 30.24 to 28.05 $\mu\text{g}/\text{m}^3$.

Freeze 2: feature exclusion. A validation ablation indicated that satellite retrieval-count features harmed leave-city-out generalisation, which is plausible since retrieval success depends on local surface brightness, snow cover and solar geometry. Frozen, then scored once on test, the validation gain of 1.75 $\mu\text{g}/\text{m}^3$ shrank to 0.25 $\mu\text{g}/\text{m}^3$ on the fold mean and the sign varies by city (t5_07_missingness_test.csv, at the frozen hyperparameters). We report that non-replication as found, because reverting after seeing a test result would make the test set a selection criterion, as with the smearing correction above, and the non-transfer of validation-based feature selection over six cities is itself information for reusers. No city, station, date or evaluation period was excluded at any point on the basis of its effect on a score.

Reproducibility

The reference implementation is deterministic across all 6×5 fold-seed pairs (t5_02, t6_01), the ensemble satisfies the convexity bound to its members in every fold, and a test re-extracts every manuscript figure from the CSVs and fails on drift. Regeneration is described under Code Availability.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

5. Discussion

For the one model class evaluated here, over 6 cities and one test year, the error in transferring a PM_{2.5} model to an unmonitored Central Asian city sits mainly in the city's overall level, with the day-to-day pattern a smaller part. Whether that reflects the region and the network or this model cannot be separated by this design, and the predictor family most likely to carry a city's level, meteorology, is absent from every tier. Per-fold, per-band and per-season figures are in t7_01–t7_03.

Two comparisons follow from that. Jin et al. (2022) report R^2 between 0.73 and 0.81 over Xinjiang under a split that does not separate cities, and AQ-Bench splits spatially but targets a time-independent metric; neither design would have exposed a level error of this kind, because neither withholds a city and then asks for its concentration. A reuser should therefore read a high pooled R^2 on this benchmark as a claim about between-city variation until the per-city figures say otherwise.

Six cities set the floor on every inference reported here, and seven further constraints qualify it.

- With 6 clusters no procedure can attain a two-sided p below 0.03125 without distributional assumptions the data do not support.
- Two of 7 instruments are low-cost and form the zero-shot Khujand fold, whose results are not comparable in kind to the other five.
- Only Khujand holds two distinct instruments, so a constant lifelong timing offset at any other city would be invisible to every check in the suite.
- The Dushanbe merge does not realign the hours its secondary feed stamps in local time: 56 filled hours (0.11% of that record) are five-hour-displaced copies, none in the test block.
- The evaluated record ends with the 2024 test block, and every StateAir feed stopped on 2025-03-04. As of 2026-08-14 the same instruments are still republished through AirNow at Ashgabat (to 2025-09-24), Almaty (to 2025-11-

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaaoa.com/index.php/11>

14) and Dushanbe (still reporting), while Bishkek, Tashkent and Astana ceased on 2025-03-04, so no result here speaks to current conditions, and extending the record forward would change the benchmark and requires a version bump.

- Kazakhstan contributes one city: Astana failed the completeness rule at 42.8% against a required 60%.
- Task F is single-horizon, predicting next-day daily means only, so it is not scored against the hourly ladder in $t3_01$ at any of its 24, 48 or 72 h horizons.
- The ground-truth panel is not redistributed, and regenerating results from raw observations requires rebuilding it (Data availability).

6. Conclusions

Central Asia has had no open station-level benchmark against which PM_{2.5} estimates could be checked, or two methods compared on the same terms. This release supplies one: daily observations from 7 instruments in 6 cities over 2018-11-27 to 2024-12-31, evaluation splits frozen and checksummed before the reported results existed, and a baseline ladder every submission is scored against. Under the protocol the benchmark is built around, which withholds a city entirely, a tuned gradient-boosting reference model reaches 28.01 $\mu\text{g}/\text{m}^3$ fold-mean RMSE against 29.77 $\mu\text{g}/\text{m}^3$ for pooled-debiased CAMS. 6 cities cannot separate those (paired $t p = 0.1392$), the interval spans both a useful improvement and a moderate degradation, and mean per-fold R^2 is -0.04. What the benchmark shows most clearly is where the error sits. It is in the held-out city's overall level, with its day-to-day pattern a smaller part.

Whether that reflects the region, the network, or this model class cannot be separated by 6 cities and one test year. Two additions would settle it on these same splits: meteorology, the predictor family most likely to carry a city's level and absent from every tier here, and cities outside the diplomatic-post network, which would show whether the flatness is a property of the region or of the training sample.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

Data availability

The frozen benchmark (split definitions, checksums and reference result tables) is deposited in Zenodo at <https://doi.org/10.5281/zenodo.21930669> under a CC BY 4.0 licence, as version 1.1.0 of eco-pulse-ca, together with the full pipeline code and no hourly PM2.5 observations. The identifier is the version DOI and resolves to this release. Results reported against this benchmark should cite it with the splits.sha256 checksum, so that a score is attributable to one frozen split. Archive contents are described in Section 3.

Ground PM2.5 observations are accessed through the OpenAQ archive (openaq.org) and originate with the US Department of State AirNow and StateAir programmes and with Clarity.

Licence status, verified 2026-08-14. Per-location licence records from OpenAQ's /v3/locations endpoint are tabulated in data/MANIFEST.md. Six of the ten source feeds carry an explicit licence permitting redistribution, the four AirNow feeds as *US Public Domain* and the two Clarity feeds as *CC0 1.0*, together 63.6% of contributing observation-hours. The four StateAir feeds carry no licence record, an absence that is systematic across all 33 StateAir providers and their 36 locations and that tracks the provider label. We treat the null as unassigned metadata while recording that no licence has been issued, and because a single deposit licence would assert a permission the evidence does not support for every feed, the observations are left at source. The Department of State's archived *Data Use Statement*, recovered from a 2014-05-12 Internet Archive capture and quoted in full in data/MANIFEST.md, states that "[a]ir quality data should not be altered in any way and should be disseminated as received", which is a further reason not to redistribute a merged, daily-aggregated panel. We attribute the observations to the Department and carry its caveat that they "are not fully verified or validated".

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaaoa.com/index.php/11>

Retrieval. All ten feeds are in OpenAQ's open-data archive on Amazon S3 (`s3://openaq-data-archive/records/csv.gz/locationid={id}/`), downloadable anonymously and verified for each on 2026-08-14, and through the v3 API with a free key, the route the pipeline uses. README.md gives the commands that rebuild the panel.

Code availability

All code used to build the benchmark, run the reference implementation, produce every table and figure, and render this manuscript is openly available at <https://github.com/jalencik/eco-pulse-benchmark>, archived at the deposit above, under an MIT licence. The archive snapshots the code at version 1.1.0; the version that produced this manuscript is the repository head. The environment is Python 3.12, with `pyproject.toml` declaring compatible ranges and `requirements-lock.txt` recording the exact versions the deposited tables were produced with.

One command runs the whole chain, make reproduce, and it is deterministic, reproducing all 31 result tables byte-identically under SHA-256 across consecutive runs. Verifying the frozen splits needs neither credentials nor a rebuild (`sha256sum -c splits.sha256`) and the test suite runs offline, while regenerating the reference results requires the ground-truth panel; the figures that derive from it are listed in `paper/scripts/extract_numbers.py`, which carries them forward on a clone without the panel and says so. The quality-control rules, split builder, Diebold–Mariano implementation with Harvey–Leybourne–Newbold correction, and the primary and sensitivity inference are original to this work, and 592 automated tests enforce split immutability, absence of leakage, table provenance and manuscript-number consistency.

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

1. References

2. Asmaa Alazmi and Hesham Rakha (2022). Assessing and Validating the Ability of Machine Learning to Handle Unrefined Particle Air Pollution Mobile Monitoring Data Randomly, Spatially, and Spatiotemporally. International Journal of Environmental Research and Public Health. <https://doi.org/10.3390/ijerph191610098>
3. George I. Austin, Itsik Pe'er and Tal Korem (2025). Distributional bias compromises leave-one-out cross-validation. Science Advances. <https://doi.org/10.1126/sciadv.adx6976>
4. Clara Betancourt et al. (2021). AQ-Bench: a benchmark dataset for machine learning on global air quality metrics. Earth system science data. <https://doi.org/10.5194/essd-13-3013-2021>
- A. Colin Cameron and Douglas L. Miller (2015). A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources. <https://doi.org/10.3368/jhr.50.2.317>
5. Sachin Chauhan et al. (2023). AirDelhi: Fine-Grained Spatio-Temporal Particulate Matter Dataset From Delhi For ML based Modeling. Advances in Neural Information Processing Systems 36. <https://doi.org/10.52202/075280-3298>
6. Francis X. Diebold (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. Journal of Business and Economic Statistics. <https://doi.org/10.1080/07350015.2014.983236>
7. Aaron van Donkelaar et al. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. Environmental Science & Technology. <https://doi.org/10.1021/acs.est.1c05309>
8. Hans Hersbach et al. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society. <https://doi.org/10.1002/qj.3803>

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

9. Sture Holm (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. <https://doi.org/10.2307/4615733>
10. Xiaoye Jin et al. (2022). Machine learning driven by environmental covariates to estimate high-resolution PM2.5 in data-poor regions. PeerJ. <https://doi.org/10.7717/peerj.13203>
11. Alexei Lyapustin et al. (2018). MODIS Collection 6 MAIAC algorithm. Atmospheric measurement techniques. <https://doi.org/10.5194/amt-11-5741-2018>
12. Hanna Meyer et al. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling. <https://doi.org/10.1016/j.ecolmodel.2019.108815>
13. Hanna Meyer and Edzer Pebesma (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution. <https://doi.org/10.1111/2041-210x.13650>
14. OpenAQ Inc. (2025). OpenAQ air quality data platform, API v3. Accessed 2026-07-29. <https://openaq.org>
15. Stefanos Papagiannis et al. (2024). Air quality challenges in Central Asian urban areas: a PM2.5 source apportionment analysis in Dushanbe, Tajikistan. Environmental Science and Pollution Research. <https://doi.org/10.1007/s11356-024-33833-6>
16. David R. Roberts et al. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. <https://doi.org/10.1111/ecog.02881>
17. Dié Tang, Yu Zhan and Fumo Yang (2024). A review of machine learning for modeling air quality: Overlooked but important issues. Atmospheric Research. <https://doi.org/10.1016/j.atmosres.2024.107261>

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

18. Madina Tursumbayeva et al. (2023). Cities of Central Asia: New hotspots of air pollution in the world. Atmospheric Environment. <https://doi.org/10.1016/j.atmosenv.2023.119901>
19. Kazbek Tursun et al. (2025). Dominant sources of PM_{2.5} in Kazakhstan's urban cities: A PMF and HYSPLIT-based study for air quality management in Central Asia. Urban Climate. <https://doi.org/10.1016/j.uclim.2025.102706>
20. Pepijn Veefkind et al. (2012). 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 Sensing of Environment. <https://doi.org/10.1016/j.rse.2011.09.027>
21. Alexandre M.J.-C. Wadoux et al. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling. <https://doi.org/10.1016/j.ecolmodel.2021.109692>
22. Tongshu Zheng et al. (2018). Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments. Atmospheric measurement techniques. <https://doi.org/10.5194/amt-11-4823-2018>

Author contributions

Jaloliddin Musayev: Conceptualisation; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing — original draft; Writing — review and editing; Visualisation; Project administration.

Asadbek Abdivayitov: Data curation; Investigation.

All authors read and approved the submitted manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank the OpenAQ project for maintaining open access to the underlying observations, and the US Department of State AirNow and StateAir programmes,

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 9, September 2026



This article/work is licensed under CC by 4.0 Attribution

<https://eurekaopenaccess.com/index.php/11>

whose diplomatic-post monitors constitute the reference-grade portion of this record.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. All computation was performed on the authors' personal hardware, and every data source used is publicly accessible at no cost.

Use of generative AI

During the preparation of this work the corresponding author used Anthropic Claude (Claude Code) to assist with software implementation, data-pipeline construction, statistical tooling, and drafting and editing of the manuscript text. The methodological decisions the dataset rests on (pre-registration of the quality rules, freezing and checksumming the splits before the reported results were produced, the leave-city-out protocol, the treatment of informative missingness, and every retraction recorded in data/DECISIONS.md) were made by the authors, who verified every reported figure against the regenerated result tables, reviewed and edited all output, and take full responsibility for the content. Generative AI is not listed as an author and cannot be held accountable for the work. No reported figure is typed by hand (Results, Reproducibility), and prose-level editing performed with AI assistance is recorded in the repository's commit history. The benchmark version described here is **1.1.0**.