Baltic J. Modern Computing, Vol. 14 (2026), No. 3, 618-644
https://doi.org/10.22364/bjmc.2026.14.3.06

Benchmark-Driven Model Selection for Forecasting over Open Gridded Urban Air-Quality Fields

Alketa HYSO, Dezdemona GJYLAPI

Department of Computer Science, Faculty of Technical and Natural Sciences, University "Ismail Qemali", Albania;

alketa.hyso@univlora.edu.al, dezdemona.gjylapi@univlora.edu.al

ORCID 0009-0000-6256-9831, ORCID 0009-0005-5186-3636

Abstract. Urban air pollution remains a challenge in cities with limited monitoring infrastructure and limited access to proprietary forecasting systems. This study presents a reproducible, benchmark-driven framework for short-horizon forecasting over open gridded air-quality and meteorological fields, demonstrated across eight Albanian cities. Daily forecasts were produced for daily mean $PM_{2.5}$ and maximum European Air Quality Index at 1-, 2-, and 3-day horizons. Local univariate baselines were compared with pooled cross-city feature-based models under a common rolling-origin backtesting design. For European Air Quality Index, the pooled benchmark achieved lower cross-city mean MAE at all three horizons, although the uncertainty analysis indicated only limited inferential separation across horizons. $PM_{2.5}$ results were more mixed, with local baselines remaining competitive at some horizons. Supplementary comparisons showed that pooled training added value beyond city-specific fitting of the same advanced model. Overall, the results support explicit benchmark-based model selection rather than reliance on a single universally preferred forecasting model.

Keywords: air quality forecasting, $PM_{2.5}$, European Air Quality Index, machine learning, open gridded data, early warning, Albania.

1. Introduction

Urban air pollution remains a persistent challenge for environmental management and public health. Recent European evidence shows that exposure to multiple air pollutants frequently occurs in combination rather than isolation, making compound episodes a persistent feature of urban risk (Chen et al., 2024). The health burden also remains substantial at the pollutant-specific level: recent work has quantified a large mortality burden associated with ozone across Europe (Achebak et al., 2024), while epidemiological analysis continues to show that harmful cardiovascular effects can be detected even below current regulatory limits (Wang et al., 2024). In Albania, recent journal evidence on long-term biomonitoring likewise suggests that air-quality assessment remains an active and relevant environmental challenge, reinforcing the need for stronger local analytical tools (Lazo et al., 2024).

In this context, short-term air-quality forecasting is not merely a prediction task; it is part of operational risk management. Public authorities and municipal analysts need forecasts that are timely enough to support warning communication, transparent enough to justify interventions, and reproducible enough to be trusted over repeated updates. These requirements are especially important in settings where dense monitoring infrastructure, proprietary forecasting services, or long station histories are limited.

Recent research has expanded in both algorithmic and operational directions (Ke et al., 2022; Rahman et al., 2024). Recent reviews show that air-quality forecasting has increasingly moved beyond traditional statistical methods toward machine-learning, deep-learning, and hybrid spatiotemporal approaches (Mendez et al., 2023; Agbehadji and Obagbuwa, 2024; Houdou et al., 2024). At the same time, they point out continuing challenges related to consistent validation, cross-study comparability, interpretability, and reproducibility (Mendez et al., 2023; Agbehadji and Obagbuwa, 2024; Houdou et al., 2024). Methodological work further emphasizes leakage-aware evaluation, strong baselines, and horizon-wise assessment (Hewamalage et al., 2023; Januschowski et al., 2022; Makridakis et al., 2022; Strom and Gundersen, 2024), while operational studies stress the importance of connecting forecasting to deployable decision-support workflows (Ke et al., 2022; Rahman et al., 2024). However, there remains limited evidence on how these principles can be integrated into a practical early-warning workflow for a relatively small group of cities using openly accessible gridded environmental data.

To address this gap, the study develops a reproducible, benchmark-driven forecasting workflow for Albanian cities using open gridded environmental data. The framework compares local univariate baselines with pooled cross-city feature-based models for forecasting daily mean $PM_{2.5}$ and daily maximum European Air Quality Index (AQI) at 1-, 2-, and 3-day horizons. It then translates the benchmark results into an operational alerting layer through champion selection, alert scoring, and city-level attention ranking. The Albanian multi-city setting serves both as an application case and as a realistic open-data environment for evaluating model comparison and operational forecasting design under data-constrained conditions. The alerting layer is also evaluated retrospectively over the full rolling-origin backtest by comparing predicted and realized alert levels derived from held-out gridded AQI and $PM_{2.5}$ fields. The study therefore focuses on a practical question: which model family is most reliable for a given forecast target and horizon?

The remainder of the article is organized as follows. Section 2 reviews the relevant literature and positions the study within current research on machine-learning methods and operational air-quality forecasting. The subsequent sections describe the data and preprocessing pipeline, forecasting methodology, benchmark results, operational early-warning outputs, discussion, limitations, and conclusions.

2. Related work

Air-quality forecasting has developed rapidly in recent years. Review studies describe a broad shift from classical statistical models toward machine-learning, deep-learning, and hybrid spatiotemporal approaches, while also pointing to continuing challenges in validation consistency, interpretability, and reproducibility (Mendez et al., 2023; Agbehadji and Obagbuwa, 2024; Houdou et al., 2024).

Comparative studies show that air-quality forecasting performance varies across pollutants, forecast horizons, and study settings, with no single model consistently dominating across all conditions (Middya and Roy, 2022; Petrić et al., 2024). Application-oriented environmental modelling studies also emphasize that pollution dynamics depend strongly on local meteorological and geographic conditions (Lashko et al., 2025). Ensemble, machine-learning, and deep-learning methods do not follow a simple complexity ranking across pollutants or case studies (Petrić et al., 2024; Ozupak et al., 2025). At the same time, forecasting is increasingly being embedded in broader monitoring and decision-support contexts, including environmental-health risk mapping, virtual monitoring-station frameworks, and AQI classification workflows (Rajesh et al., 2025; Makhdoomi et al., 2025; Liu et al., 2024).

A substantial share of recent $PM_{2.5}$ and AQI research has focused on advanced predictive architectures. Recent examples include deep-learning ensembles, federated learning, Transformer-BiLSTM pipelines, optimization-enhanced models, attention-based recurrent networks, and hybrid regional forecasting systems (Gao et al., 2024; Hu et al., 2024; Cheng et al., 2025; Cengil, 2025; Meng et al., 2025; Hu et al., 2025). Together, these studies show how quickly the methodological toolkit for air-quality prediction is expanding.

Another strand of the literature emphasizes operational deployment. Ke et al. (2022) showed how automated machine-learning workflows can support routine air-quality forecasting, while Rahman et al. (2024) illustrated the shift toward service-oriented forecasting platforms with user-facing deployment logic. These contributions show that forecasting systems are increasingly being framed not only as predictive models, but also as practical decision-support tools.

The broader forecasting-methodology literature provides further guidance for model comparison. Hewamalage et al. (2023) discuss common pitfalls in time-series evaluation, including leakage and weak baselines. Januschowski et al. (2022) show that global tree-based models can remain highly competitive when temporal information is represented through structured tabular features, while findings from the M5 competition reinforce the value of strong benchmark design and global machine-learning approaches (Makridakis et al., 2022). Horizon-wise evaluation can also reveal performance patterns that aggregate accuracy measures may obscure (Strom and Gundersen, 2024).

Taken together, the literature documents rapid model development and growing operational interest, but offers more limited comparative evidence on reproducible benchmark design, structured pooled-versus-local learning, and target- and horizon-specific model selection in smaller multi-city settings based on openly accessible gridded modeled environmental data.

3. Data and preprocessing

3.1. Data sources and geographic scope

The study covers eight Albanian cities: Tirane, Durres, Elbasan, Shkoder, Vlore, Fier, Korce, and Berat. Hourly air-quality data were retrieved from the Open-Meteo Air Quality application programming interface **(**API) (Open-Meteo, 2026a), while hourly meteorological covariates were obtained from the Open-Meteo Historical Weather API

(Open-Meteo, 2026b). In the implementation used for this study, air-quality requests were issued to the https://air-quality-api.open-meteo.com/v1/air-quality endpoint with the cams_europe domain, and historical weather requests were issued to the https://archive-api.open-meteo.com/v1/archive endpoint. All requests used city coordinates in the World Geodetic System 1984 (WGS84) system and the Europe/Tirane timezone.

The air-quality dataset included hourly $PM_{2.5}$, PM_{10}, nitrogen dioxide (NO_2), ozone (O_3), sulphur dioxide (SO_2), carbon monoxide, the overall AQI, and pollutant-specific AQI sub-indices for $PM_{2.5}$, PM_{10}, NO_2, O_3, and SO_2 (Open-Meteo, 2026a). The meteorological dataset included hourly near-surface air temperature, relative humidity, precipitation, wind speed, wind direction, and mean sea-level pressure (Open-Meteo, 2026b). The analysis covered the period from 1 January 2024 to 28 April 2026.

The Open-Meteo documentation states that its European air-quality service is based on Copernicus Atmosphere Monitoring Service (CAMS) European air-quality products at approximately 0.1 degrees spatial resolution, while the historical weather service is based on reanalysis-driven meteorological products (Open-Meteo, 2026a, 2026b). Accordingly, the present framework should be interpreted as an open gridded-data workflow rather than a replacement for regulatory station measurements. No regulatory station measurements were used for model fitting or external evaluation in this study. AQI interpretation follows the European AQI categories implemented by Open-Meteo and is aligned with the broader AQI communication framework of the European Environment Agency (EEA) (European Environment Agency, 2026; Open-Meteo, 2026a).

3.2. Hourly harmonization and daily aggregation

Hourly air-quality and meteorological data were downloaded separately for each city and merged into a single hourly analytical table using city and timestamp as matching keys. The merged dataset was standardized programmatically: timestamps were converted to datetime format, calendar variables, including year, month, day, and hour, were derived, and pollutant, AQI, and meteorological fields were converted to numeric format, with invalid entries treated as missing values.

Because the forecasting task focused on short-horizon operational forecasting, the hourly data were aggregated to daily resolution. Daily means and maxima were computed for $PM_{2.5}$, particulate matter with an aerodynamic diameter of 10 μm or less (PM_{10}), nitrogen dioxide, ozone, sulphur dioxide, carbon monoxide, and AQI-related variables.

The retained meteorological summaries included daily mean, minimum, and maximum temperature, mean relative humidity, daily precipitation total, mean and maximum wind speed, mean sea-level pressure, and the number of hourly observations contributing to each city-day. In the downloaded snapshot used for this study, each retained city-day in the merged hourly analytical table contained 24 hourly timestamps, and none of the retained hourly air-quality or meteorological variables were missing after numeric coercion. Accordingly, no additional day-level completeness filtering, imputation, or gap-filling was applied before daily aggregation. Table 1 summarizes the aggregated dataset, while the overall workflow of the proposed methodology is summarized in Fig. 1.

Table 1. Summary of the daily aggregated dataset used in the forecasting experiments, including spatial coverage, study period, record counts, main variable groups, and forecast targets.

Item	Value
Geographic scope	8 Albanian cities
Study period	1 January 2024 to 28 April 2026
Source resolution	Hourly
Hourly records	163,008
Daily records	6,792 city-day observations
Daily analytical variables	49
Main variable groups	Pollutant summaries, AQI summaries, meteorological summaries, and risk indicators
Forecast targets	Daily mean $PM_{2.5}$ and daily maximum AQI

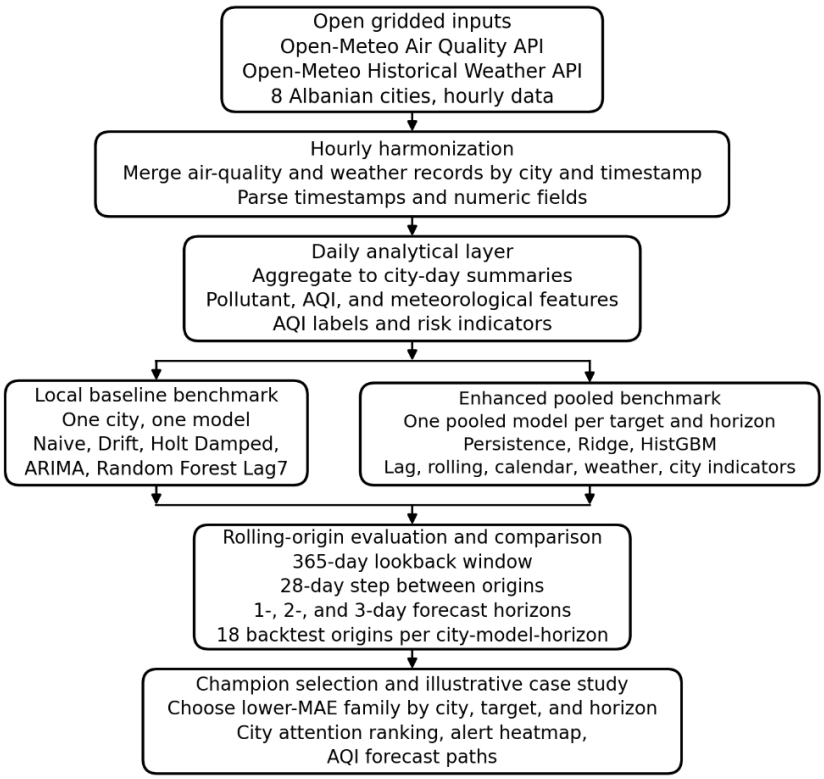

Figure 1. Schematic overview of the proposed forecasting methodology.

Daily AQI labels were assigned from the daily maximum AQI score using the European AQI communication bands as implemented by Open-Meteo and based on the European Environment Agency framework (European Environment Agency, 2026; Open-Meteo, 2026a): Good (0-20), Fair (21-40), Moderate (41-60), Poor (61-80), Very Poor (81-100), and Extremely Poor (>100). For operational analysis in this study, pollution episodes were defined as Poor-or-worse days, and high-risk days were defined as Very-Poor-or-worse days. A numeric AQI severity scale was used for ranking and event detection. Across the full 2024-2026 sample of 6,792 city-day observations, AQI labels were distributed as 24 Good, 4,599 Fair, 1,893 Moderate, 262 Poor, 10 Very Poor, and 4 Extremely Poor days. Under the study definitions, this corresponds to 276 pollution episodes and 14 high-risk days. For each day, the dominant pollutant signal was defined as the pollutant whose daily mean AQI sub-index was highest. In addition, the pipeline generated threshold-based screening flags for selected pollutants. The resulting dataset was used for forecasting, benchmark evaluation, and operational alert generation.

4. Forecasting methodology

4.1. Forecast targets and evaluation design

The forecasting layer predicts two daily targets:

- Daily mean $PM_{2.5}$ and
- Daily maximum AQI.

$PM_{2.5}$ was selected as the primary pollutant-specific target, whereas AQI was included because of its direct relevance to public-facing risk communication.

Forecasts were generated for 1-, 2-, and 3-day horizons using a rolling-origin backtesting design consistent with leakage-aware time-series evaluation practices (Hewamalage et al., 2023; Januschowski et al., 2022; Makridakis et al., 2022; Strom and Gundersen, 2024). Both benchmark families used a 365-day historical lookback, rolling forecast origins advancing in 28-day steps, and strictly chronological training and testing splits.

To make the temporal alignment explicit, let t denote the forecast issue date and h the forecast horizon (h = 1, 2, 3). All predictors were constructed using information available by the end of day t only, whereas the prediction target was the outcome on day t + h. In the pooled models, the current target value and the meteorological and cross-pollutant covariates refer to observed daily aggregates from day t, lagged target features use values from t - 1 backward, and rolling target summaries are computed over windows ending at t - 1. The workflow should therefore be interpreted as an end-of-day forecasting design: forecasts for days t + 1 to t + 3 are issued after the daily aggregates for day t have been completed, rather than as intraday nowcasts. No information from the forecast day or later dates was used in feature construction, model fitting, or evaluation. Operationally, this corresponds most naturally to a daily or overnight forecasting update in which authorities issue next-day to 3-day risk guidance after the current day has closed and its daily indicators have been consolidated. It is therefore realistic as a short-horizon planning and prioritization workflow, but it should not be interpreted as a same-day intraday warning design.

Each city-model-horizon combination was evaluated using 18 rolling forecast origins. Forecast accuracy was summarized primarily with mean absolute error (MAE), while mean absolute percentage error (MAPE) was retained as a secondary relative-error measure. In both benchmark families, MAPE was computed with a numerical safeguard that excluded only exact or near-zero denominators through |actual| < 1e-8; no additional lower-bound truncation or exclusion of low $PM_{2.5}$ values was applied. Because percentage-based errors can be sensitive to low target values, symmetric mean absolute percentage error (sMAPE) was also examined as a supplementary robustness metric.

4.2. Local baseline benchmark

The first benchmark family consisted of six local univariate forecasting methods fitted independently for each city-target series:

1. Naive persistence
2. Seasonal Naive (7-day seasonal period)
3. Drift
4. Holt damped trend
5. ARIMA
6. Random Forest with seven lagged target features

The Naive model repeated the most recent observation across all forecast horizons. The Seasonal Naive model repeated the observation from the same day of the previous week, providing an explicit weekly seasonal reference for the daily series. Drift extrapolated a linear trend estimated from the training window. Holt Damped used additive exponential smoothing with a damped trend component.

ARIMA considered the candidate orders (1,1,0), (1,1,1), and (0,1,1), selecting the converged specification with the lowest Akaike Information Criterion (AIC). If all candidate fits failed, the model reverted to the Naive forecast.

The Random Forest model used seven lagged target observations and a sequential time index as predictors, with 120 trees, minimum leaf size equal to 2, and random_state = 42. Multi-step forecasts were generated recursively.

In practical terms, this baseline benchmark followed a one-city, one-model design: each model was trained only on the historical target series of a single city and did not borrow information from other cities. The local baseline benchmark was specified as a fixed benchmark set rather than as a fully tuned model collection. Naive, Seasonal Naive, and Drift have no tunable hyperparameters in this implementation. Holt Damped estimated its internal smoothing parameters separately within each rolling training window through the standard fitting routine, but no hyperparameter search was performed on the rolling-origin backtest. ARIMA order selection was restricted to the pre-specified candidate set and used training-window AIC only. Likewise, the Random Forest baseline retained fixed settings throughout the main benchmark and was not tuned on the rolling-origin evaluation period.

This design kept the baseline comparison reproducible, leakage-safe, and interpretable, while still combining transparent reference methods, classical statistical forecasting models, and a simple non-linear machine-learning comparator. Such a benchmark structure is consistent with recent recommendations that forecasting systems should be evaluated against strong and interpretable baselines rather than assuming that greater model complexity will automatically improve forecasting performance (Hewamalage et al., 2023; Januschowski et al., 2022; Makridakis et al., 2022).

4.3. Enhanced pooled benchmark

The second benchmark family used a pooled cross-city feature-based forecasting design with structured temporal and environmental predictors. Instead of fitting separate models for each city, forecasting was reformulated as a supervised learning problem for each target and horizon, following recent global tabular forecasting approaches (Januschowski et al., 2022; Makridakis et al., 2022).

Three models were evaluated:

- Persistence Current, which uses the current target value as the forecast.
- Ridge Global Features
- HistGradientBoosting (HistGBM) Global Features

Feature construction included:

- current target value on the issue date t,
- lagged target observations at t - 1, t - 2, t - 3, t - 7, t - 14, t - 21, and t - 28 days,
- rolling target means, standard deviations, and maxima over 3-, 7-, 14-, and 28-day windows ending at t - 1,
- calendar variables known at the issue date (day of week, month, day of year, and weekend indicator),
- sinusoidal seasonal encodings,
- issue-day meteorological covariates from day t (daily mean and maximum temperature, mean relative humidity, daily precipitation total, mean and maximum wind speed, and mean sea-level pressure),
- issue-day cross-pollutant context variables from day t,
- one-hot encoded city indicators.

$PM_{2.5}$ models received issue-day AQI, PM_{10}, NO_2, and O_3 context variables, whereas the AQI models received $PM_{2.5}$, PM_{10}, NO_2, and O_3 predictors.

Direct forecasting was applied by defining separate supervised targets for the 1-, 2-, and 3-day horizons (Strom and Gundersen, 2024). For each horizon h, each supervised row was anchored on an issue date t and mapped to the target value on day t + h. At each forecast origin, models were trained only on rows whose target dates had already been observed by that origin and were then evaluated on rows with feature date equal to the current issue date. Training windows again covered the previous 365 days, rolling forecast origins advanced every 28 days, and a minimum of 400 pooled training rows was required before model fitting.

The enhanced pooled benchmark was likewise implemented with fixed hyperparameters throughout the main rolling-origin study. Ridge Global Features used alpha = 2.0, while HistGBM Global Features used learning_rate = 0.05, max_depth = 4, max_iter = 120, min_samples_leaf = 20, and random_state = 42 across all targets and forecast horizons. No hyperparameter optimization was performed on the 18 rolling forecast origins for these main enhanced models. This was a deliberate design choice intended to keep the benchmark leakage-safe, reproducible, and focused on model-family comparison under a common feature design rather than on model-specific best-case tuning.

4.4. Supplementary uncertainty and stability analysis

To assess the stability of the benchmark comparisons, supplementary paired bootstrap analyses were conducted for the globally selected best enhanced model and the

corresponding best local baseline at each target-horizon combination. Two resampling views were used. First, a city-level paired bootstrap treated each city's mean error as one paired unit (n = 8 cities). Second, a clustered bootstrap resampled cities while retaining all aligned forecast cases within each selected city, thereby preserving within-city dependence across rolling origins (n = 144 forecast cases per target-horizon comparison). For both analyses, 95% confidence intervals were computed for the difference in mean MAE, defined as baseline minus enhanced MAE so that positive values indicate lower error for the enhanced benchmark.

The six target-horizon benchmark contrasts were treated as the primary pre-specified comparisons. No formal family-wise correction was used in the main benchmark tables, because these contrasts were interpreted primarily as descriptive forecast comparisons rather than as a family of simultaneous confirmatory hypothesis tests. However, as a conservative sensitivity check, Bonferroni-adjusted confidence intervals were also examined across these six primary contrasts. The feature-group ablation and XGBoost analyses were treated as supplementary exploratory comparisons and were not included in this primary family-wise adjustment.

4.5. Feature-group ablation analysis

To better interpret the sources of improvement in the enhanced pooled benchmark, a feature-group ablation analysis was conducted for the HistGBM Global Features model. The analysis retained the same rolling-origin evaluation design, training window, forecast horizons, and pooled training setup used in the main enhanced benchmark, while selectively removing predefined groups of predictors from the feature set.

Three feature groups were examined:

- meteorological covariates,
- cross-pollutant context variables,
- one-hot encoded city indicators.

For each forecast target and horizon, the full enhanced feature set was compared against reduced variants in which one feature group at a time was removed. Performance was again summarized using cross-city mean MAE, allowing the contribution of each feature group to be assessed relative to the full enhanced benchmark.

4.6. Supplementary city-specific advanced-model comparison

To assess whether the gains of the enhanced pooled benchmark were due mainly to the structured feature design or also to pooled cross-city training, a supplementary city-specific comparison was conducted using the HistGBM Global Features model. For each forecast target and horizon, the same temporal, meteorological, and cross-pollutant predictors used in the enhanced pooled benchmark were retained, but the model was trained separately for each city using only that city's historical observations. Because each model was fitted within a single city, one-hot encoded city indicators were not included.

The comparison retained the same rolling-origin evaluation design as the main enhanced benchmark, including the 365-day training window, 28-day origin stride, and 1-, 2-, and 3-day forecast horizons. Performance was evaluated against the pooled cross-city HistGBM reference using cross-city mean MAE and city-level win counts. Under the fixed 365-day rolling training design, city-specific training windows retained between 334 and

365 usable supervised rows after lag and rolling-feature construction and horizon-specific target alignment. This variation does not reflect missing daily observations. Instead, it arises because the earliest part of each 365-day window cannot contribute usable supervised rows once the 28-day lag structure and the 1- to 3-day forecast-horizon shift are applied.

4.7. Supplementary XGBoost comparative experiment

To examine whether a different tree-based global learner could challenge the existing enhanced benchmark without changing the feature design, two supplementary XGBoost analyses were conducted. The first was a direct comparison between XGBoost Global Features and HistGBM Global Features under the same pooled feature design, the same 365-day rolling training window, the same 28-day origin stride, and the same 1-, 2-, and 3-day forecast horizons as the main enhanced benchmark. In this direct comparison, the default XGBoost configuration used n_estimators = 300, learning_rate = 0.05, max_depth = 4, min_child_weight = 5.0, subsample = 0.9, colsample_bytree = 0.9, and tree_method = hist.

The second analysis was a separate time-split XGBoost tuning experiment designed to avoid optimistic hyperparameter selection. For this analysis, the 18 rolling forecast origins were divided chronologically into 12 development origins (30 December 2024 to 3 November 2025) and 6 later evaluation origins (1 December 2025 to 20 April 2026). Candidate XGBoost configurations were compared only on the development origins, and the best-performing configuration for each target-horizon pair was then evaluated only on the held-out origins. This tuned evaluation was assessed against both the default XGBoost specification and the development-selected enhanced incumbent model.

4.8. Champion selection and operational alerts

The operational layer first identified the best-performing model within each benchmark family for every city, target, and forecast horizon, and then selected the lower-error family winner using backtest MAE. The operational layer was designed around explicit champion selection rather than forecast averaging, so that deployment decisions remained directly traceable to benchmarked model-family comparisons.

Predicted daily maximum AQI served as the primary alert signal, while predicted $PM_{2.5}$ acted as a supporting escalation factor relative to the city-specific historical 90th and 95th percentiles. AQI categories were mapped into the operational alert levels Stable, Watch, Warning, High, Severe, and Critical. The present operational alert layer is deterministic: each alert level is derived from point forecasts of daily maximum AQI and daily mean $PM_{2.5}$, and no forecast-specific prediction intervals or probabilistic exceedance estimates are propagated into the alert rules.

Operationally:

- AQI values classified as Good, Fair, Moderate, Poor, Very Poor, and Extremely Poor mapped by default to Stable, Watch, Warning, High, Severe, and Critical alerts, respectively.
- $PM_{2.5}$ values at or above the city-specific 90th percentile escalated alerts to at least High.
- $PM_{2.5}$ values at or above the 95th percentile escalated alerts to at least Severe.
- Very Poor AQI combined with extreme $PM_{2.5}$ triggered Critical alerts.

These escalation thresholds were defined for operational benchmarking purposes within the present study and should not be interpreted as official public-health alert standards. The complete operational mapping is summarized in Table 2.

For each city, the highest alert reached within the 1- to 3-day forecast window was retained for operational ranking, heatmap generation, and forecast visualization. In this way, the workflow linked benchmarked forecasting performance to a transparent operational alerting layer rather than treating prediction as an end in itself.

Table 2. Study-defined operational mapping used to convert predicted daily maximum AQI and city-relative $PM_{2.5}$ escalation signals into final alert levels.

Final alert level	**Default AQI trigger**	**$PM_{2.5}$ escalation rule**	**Operational interpretation**
Stable	Good AQI	$PM_{2.5}$ below the city-specific 90th percentile	Low short-term concern
Watch	Fair AQI	$PM_{2.5}$ below the city-specific 90th percentile	Conditions to monitor
Warning	Moderate AQI	$PM_{2.5}$ below the city-specific 90th percentile	Elevated short-term concern
High	Poor AQI	Default AQI-driven High alert	High AQI-driven concern
High	Fair or Moderate AQI	$PM_{2.5}$ at or above the city-specific 90th percentile and below 95th percentile	$PM_{2.5}$-driven escalation to High
Severe	Very Poor AQI	Default AQI-driven Severe alert	Very high AQI-driven concern
Severe	Poor AQI	$PM_{2.5}$ at or above the city-specific 95th percentile	$PM_{2.5}$-driven escalation to Severe
Critical	Extremely Poor AQI	Any $PM_{2.5}$ level	Highest AQI-driven alert
Critical	Very Poor AQI	$PM_{2.5}$ at or above the city-specific 95th percentile	Joint AQI and extreme $PM_{2.5}$ escalation

Note: The final alert level is defined as the maximum of the AQI-default alert and any $PM_{2.5}$-based escalation implied by the same forecast instance.

4.9. Retrospective validation of the operational alert layer

Because the alerting layer was intended as the operational bridge between benchmark results and city-level prioritization, it was also evaluated retrospectively over the full rolling-origin backtest. First, the benchmark-defined champion map was fixed by selecting the lower-MAE source and model for each city, forecast target, and forecast horizon from

the main benchmark summaries. Second, these fixed champion selections were reapplied to the corresponding historical out-of-sample rolling forecasts. Third, the predicted daily maximum AQI and predicted daily mean $PM_{2.5}$ values were converted into final alert levels using the same operational rule set described in Section 4.8, including the city-specific $PM_{2.5}$ p90 and p95 escalation thresholds. Fourth, realized alert levels were derived from the held-out target-day AQI and $PM_{2.5}$ values from the same gridded daily fields. Predicted and realized alerts were then compared using exact alert-level agreement, within-one-level agreement, mean absolute alert-score gap, a multiclass confusion matrix, and binary precision-recall summaries for High-or-worse and Severe-or-worse alerts. This retrospective step should be interpreted as an evaluation of the alerting workflow over open gridded proxy fields rather than against regulatory station observations.

Because the champion map was derived from the same rolling-origin backtest period to which it was retrospectively reapplied, the resulting alert metrics represent an internal retrospective evaluation rather than an independent external validation.

5. Results

5.1. Baseline benchmark results

The local baseline benchmark provided the reference layer for evaluating the enhanced pooled benchmark. It served not only as a performance baseline, but also as a practical point of comparison for assessing whether the more complex pooled feature-based benchmark provided meaningful gains beyond established city-specific methods.

Table 3 summarizes the best-performing local baseline model for each target and forecast horizon. For daily maximum AQI, the strongest baseline model varied across horizons: Naive performed best at the 1-day horizon, Holt Damped at 2 days, and ARIMA at 3 days. This pattern suggests that AQI forecasting within the local baseline family did not favour a single method uniformly, and that the most suitable local benchmark depended on forecast horizon.

Table 3. Best-performing local baseline model for each target and forecast horizon.

Forecast target	Forecast horizon (days)	Best baseline model	Mean MAE	Mean MAPE
Daily max AQI	1	Naive	3.12	7.99
Daily max AQI	2	Holt Damped	4.15	10.94
Daily max AQI	3	ARIMA	4.49	12.38
Daily mean $PM_{2.5}$	1	Random Forest Lag7	2.04	18.36
Daily mean $PM_{2.5}$	2	ARIMA	3.00	30.78
Daily mean $PM_{2.5}$	3	ARIMA	2.78	31.54

For daily mean $PM_{2.5}$, the strongest baseline pattern was more mixed. Random Forest with seven lagged target features achieved the lowest cross-city mean MAE at the 1-day horizon, whereas ARIMA remained the strongest baseline at the 2-day and 3-day horizons. The added Seasonal Naive benchmark did not outperform the strongest local baselines at any target-horizon combination.

Because MAPE can be sensitive to low denominators, supplementary sMAPE summaries were also examined for the benchmark-winning baseline and enhanced models. The sMAPE results preserved the same broad interpretation as the MAE/MAPE comparisons: AQI results remained consistently favourable to the enhanced pooled benchmark, while $PM_{2.5}$ remained mixed across horizons, with the baseline retaining the advantage at the 2-day horizon and the 3-day comparison remaining very close.

Taken together, these baseline results establish a strong local reference point for the enhanced pooled benchmark. They also reinforce the broader expectation that forecast performance should be evaluated separately by target and horizon rather than assuming that one local model family will dominate under all short-term forecasting conditions.

5.2. Enhanced benchmark results

The enhanced pooled benchmark produced the clearest point-estimate gains for AQI forecasting. HistGBM Global Features achieved the lowest cross-city mean MAE among the enhanced models at all AQI forecast horizons.

Fig. 2 compares the best local baseline models with the best enhanced pooled models across targets and forecast horizons, while Table 4 reports the corresponding numerical values.

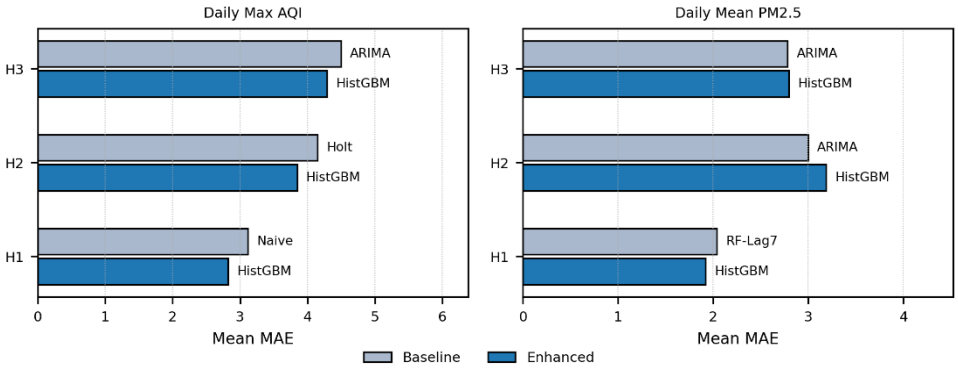

Figure 2. Mean absolute error comparison between the best local baseline models and the best enhanced pooled models for daily maximum AQI and daily mean $PM_{2.5}$ across the 1-, 2-, and 3-day forecast horizons.

Viewed together, Fig. 2 and Table 4 show that the enhanced pooled benchmark altered the benchmark ranking more clearly for AQI than for $PM_{2.5}$.

Table 4. Comparison between the best-performing enhanced model and the best-performing local baseline model for each forecast target and horizon, summarized by mean MAE and mean MAPE across cities.

Forecast target	Forecast horizon (days)	Best enhanced model	Enhanced mean MAE	Best baseline model	Baseline mean MAE	MAE improvement	MAPE improvement
Daily max AQI	1	HistGBM Global	**2.82**	Naive	3.12	0.29	0.71
Daily max AQI	2	HistGBM Global	**3.84**	Holt Damped	4.15	0.30	0.99
Daily max AQI	3	HistGBM Global	**4.29**	ARIMA	4.49	0.20	0.77
Daily mean $PM_{2.5}$	1	HistGBM Global	**1.92**	Random Forest Lag7	2.04	0.11	0.66
Daily mean $PM_{2.5}$	2	HistGBM Global	3.19	ARIMA	**3.00**	-0.19	-0.88
Daily mean $PM_{2.5}$	3	HistGBM Global	2.79	ARIMA	**2.78**	-0.01	0.72

For AQI, the same enhanced model remained best at all three horizons and outperformed three different baseline references, which suggests a more consistent directional pattern in the point-estimate comparisons than was observed for $PM_{2.5}$. At the same time, the absolute AQI errors increased with forecast horizon for both benchmark families, so the result should be read as a persistent relative point-estimate advantage under progressively harder forecast conditions rather than as evidence of strong inferential separation at all horizons. For $PM_{2.5}$, the contrast was structurally less stable: the enhanced benchmark offered only a small gain at 1 day, while the baseline family remained stronger or nearly equivalent thereafter. In practical terms, Fig. 2 and Table 4 therefore suggest that pooled feature-based learning was more consistently useful for cross-city AQI forecasting than for replacing strong local $PM_{2.5}$ baselines across the whole short-horizon range.

Because the observed gains were relatively modest, Tables 5 and 6 provide supplementary stability evidence showing the uncertainty around the average MAE differences and the frequency with which the enhanced benchmark outperformed the best local baseline across cities.

Table 5. Stability evidence for daily maximum AQI, reported as baseline minus enhanced MAE, paired city-level 95% confidence intervals, and city-level win counts. Positive MAE differences favor the enhanced benchmark.

H	Comparator	Delta B-E	95% CI	Wins E:B
1	Naive	0.29	[-0.07, 0.61]	7:1
2	Holt	0.30	[-0.32, 0.96]	4:4
3	ARIMA	0.20	[0.02, 0.38]	6:2

Table 6. Stability evidence for daily mean $PM_{2.5}$, reported as baseline minus enhanced MAE, paired city-level 95% confidence intervals, and city-level win counts. Positive MAE differences favour the enhanced benchmark.

H	Comparator	Delta B-E	95% CI	Wins E:B
1	RF-Lag7	0.11	[-0.04, 0.27]	6:2
2	ARIMA	-0.19	[-0.34, -0.01]	1:7
3	ARIMA	-0.01	[-0.24, 0.19]	6:2

Note: Delta B-E denotes baseline minus enhanced MAE, so positive values indicate lower error for the enhanced benchmark. Wins E:B summarizes how many of the eight city-level comparisons were won by the enhanced and baseline models, respectively.

The additional stability analysis indicates that the AQI point-estimate advantage should be interpreted cautiously. The unadjusted 95% city-level confidence intervals showed the clearest support for AQI at the 3-day horizon, whereas the 1-day and 2-day AQI intervals crossed zero. For $PM_{2.5}$, the uncertainty results remained mixed: the 2-day comparison favored the local baseline, whereas the 1-day and 3-day comparisons were not clearly separated. A conservative Bonferroni sensitivity check was also examined across the six primary target-horizon benchmark contrasts (Supplementary Table S1). Under that adjustment, none of the six comparisons remained clearly separated from zero. Taken together, these results suggest that Table 4 should be read primarily as point-estimate benchmark evidence, whereas Tables 5 and 6 provide the uncertainty context needed for cautious directional interpretation rather than strong confirmatory claims

To assess whether the $PM_{2.5}$ percentage errors were disproportionately influenced by low-concentration days, Supplementary Table S2 reports sMAPE alongside MAE and MAPE for the benchmark-winning baseline and enhanced models across the six primary target-horizon comparisons. For $PM_{2.5}$, sMAPE was somewhat lower than MAPE at the longer horizons, as expected, but it preserved the same broad benchmark interpretation. At the 1-day horizon, the enhanced model remained better under sMAPE (17.40% versus 18.15% for the best baseline). At the 2-day horizon, the baseline retained the advantage (27.67% versus 29.25%), and at the 3-day horizon the comparison remained very close (27.77% versus 28.13%). In the held-out $PM_{2.5}$ benchmark cases, the minimum actual daily mean $PM_{2.5}$ value was 2.81 $\mu g/m^3$, indicating that extreme near-zero denominator inflation was limited in practice.

In practical terms, these error differences were generally modest and are discussed further in relation to operational significance in Section 6.

5.3. Feature-group ablation results

To better interpret the enhanced pooled benchmark, a descriptive feature-group ablation analysis was performed for the HistGBM Global Features model. Across targets and horizons, meteorological covariates provided the most consistent contribution, as their removal increased cross-city mean MAE in all six target-horizon combinations. The largest deterioration was observed for 3-day-ahead AQI forecasting, suggesting that

weather information was particularly important for preserving medium-short-horizon AQI accuracy within the pooled feature design. Table 7 summarizes the cross-city mean MAE values for the full feature set and for the reduced feature-group variants.

Table 7. Feature-group ablation results for the HistGBM enhanced pooled benchmark, reported as cross-city mean MAE by target and forecast horizon.

Forecast target	Horizon	Full features	No weather	No cross-pollutant context	No city indicators
Daily max AQI	1	2.82	2.90	**3.01**	2.81
Daily max AQI	2	3.84	**4.00**	3.83	3.77
Daily max AQI	3	4.29	**4.67**	4.27	4.37
Daily mean $PM_{2.5}$	1	1.92	1.96	**1.98**	1.90
Daily mean $PM_{2.5}$	2	3.19	**3.40**	3.09	3.12
Daily mean $PM_{2.5}$	3	2.79	**2.93**	2.77	2.80

Cross-pollutant context variables were most helpful at shorter horizons, particularly for 1-day-ahead AQI and $PM_{2.5}$ forecasting, although their contribution became weaker and less consistent at longer horizons. By contrast, city indicators showed a smaller and more mixed effect. In some cases, removing city indicators or cross-pollutant context produced very small improvements in mean MAE, indicating that these feature groups did not contribute uniformly across all target-horizon settings. These cases should be interpreted cautiously, however, because the ablation results are descriptive and do not provide a causal attribution of feature importance. Overall, the results suggest that the strongest and most stable contribution within the enhanced pooled benchmark came from meteorological context, while the role of cross-pollutant and city-specific predictors was more target- and horizon-dependent.

5.4. Supplementary city-specific advanced-model comparison

A supplementary comparison was conducted to test whether the strongest enhanced model could achieve lower forecast error when trained separately for each city rather than across all cities jointly. The results in Table 8 did not support that expectation. For daily maximum AQI, the pooled cross-city HistGBM model outperformed the city-specific variant at all three horizons in the cross-city average summary, reducing mean MAE from 3.33 to 2.82 at 1 day, from 4.34 to 3.84 at 2 days, and from 4.91 to 4.29 at 3 days. The pooled model also won 8 of 8 city-level comparisons at the 1-day horizon, 7 of 8 at the 2-day horizon, and 6 of 8 at the 3-day horizon.

Table 8. Comparison between the city-specific and pooled cross-city HistGBM feature-based models, summarized by cross-city mean MAE and city-level win counts.

Forecast target	Forecast horizon (days)	City-specific mean MAE	Pooled mean MAE	Local minus pooled MAE	City-specific wins	Pooled wins
Daily max AQI	1	3.33	**2.82**	0.50	0	8
Daily max AQI	2	4.34	**3.84**	0.50	1	7
Daily max AQI	3	4.91	**4.29**	0.62	2	6
Daily mean $PM_{2.5}$	1	2.23	**1.92**	0.31	1	7
Daily mean $PM_{2.5}$	2	3.23	**3.19**	0.04	3	5
Daily mean $PM_{2.5}$	3	2.91	**2.79**	0.12	2	6

For daily mean $PM_{2.5}$, the pooled cross-city model again remained stronger overall, although the margin was smaller. Mean MAE decreased from 2.23 to 1.92 at 1 day, from 3.23 to 3.19 at 2 days, and from 2.91 to 2.79 at 3 days. The pooled model won 7 of 8 city-level comparisons at the 1-day horizon, 5 of 8 at the 2-day horizon, and 6 of 8 at the 3-day horizon. Across all 48 city-target-horizon comparisons, the city-specific advanced model won 9 cases, compared with 39 for the pooled cross-city version. These results suggest that the gains of the enhanced benchmark cannot be explained only by the use of a stronger model class and richer local features; pooled cross-city training itself appears to contribute additional predictive value.

5.5. Supplementary XGBoost Comparison

Two supplementary XGBoost analyses were conducted to provide additional context for the enhanced benchmark. The first was a direct comparison between XGBoost Global Features and HistGBM Global Features under the same pooled feature design and the full 18-origin rolling evaluation. The second was a separate time-split tuning experiment, in which XGBoost hyperparameters were selected on earlier development origins and then evaluated only on later held-out origins.

In Tables 9 and 10, the reported “XGB wins” and “HistGBM wins” refer to city-level MAE win counts out of the eight cities for each target-horizon comparison. Under the full 18-origin rolling evaluation, XGBoost performed clearly better than HistGBM for $PM_{2.5}$ at the 1-day horizon, reducing mean MAE from 1.92 to 1.78 and winning all 8 city-level comparisons. It also performed slightly better for AQI at the 2-day and 3-day horizons, although these gains were modest. However, XGBoost did not outperform the existing enhanced benchmark uniformly across all target-horizon combinations, so it should be interpreted as a competitive challenger rather than as a consistently superior replacement.

Table 9. Comparison between XGBoost Global Features and HistGBM Global Features for daily maximum AQI under the same pooled feature design and 18-origin rolling evaluation. Negative XGB-HistGBM values indicate lower cross-city mean MAE for XGBoost, while "XGB wins" and "HistGBM wins" denote city-level MAE win counts out of the eight cities.

Horizon (days)	XGB MAE	HistGBM MAE	XGB-HistGBM	XGB wins	HistGBM wins
1	2.84	2.82	0.02	4	4
2	3.82	3.84	-0.02	6	2
3	4.28	4.29	-0.01	3	5

Table 10. Comparison between XGBoost Global Features and HistGBM Global Features for daily mean $PM_{2.5}$ under the same pooled feature design and 18-origin rolling evaluation. Negative XGB-HistGBM values indicate lower cross-city mean MAE for XGBoost, while "XGB wins" and "HistGBM wins" denote city-level MAE win counts out of the eight cities.

Horizon (days)	XGB MAE	HistGBM MAE	XGB-HistGBM	XGB wins	HistGBM wins
1	1.78	1.92	-0.15	8	0
2	3.20	3.19	0.01	4	4
3	2.79	2.79	-0.01	3	5

The direct comparison between the two tree-based pooled models was therefore mixed rather than one-sided. XGBoost achieved lower cross-city mean MAE in four of the six target-horizon combinations, with the clearest gain observed for $PM_{2.5}$ at the 1-day horizon. HistGBM remained slightly stronger for AQI at the 1-day horizon and $PM_{2.5}$ at the 2-day horizon, while the remaining differences were small.

The separate time-split tuning experiment led to a similarly selective conclusion. On the held-out evaluation origins, tuned XGBoost improved AQI forecasting at the 2-day horizon, reducing mean MAE from 2.53 under the default XGBoost specification to 2.43 and outperforming the held-out HistGBM incumbent (2.68). For AQI at the 1-day horizon and $PM_{2.5}$ at the 1-day and 3-day horizons, the selected configuration remained the default XGBoost setting. By contrast, tuning worsened AQI at the 3-day horizon and $PM_{2.5}$ at the 2-day horizon relative to default XGBoost. Taken together, these two supplementary analyses suggest that XGBoost is a credible challenger, especially for $PM_{2.5}$ at the 1-day horizon and AQI at the 2-day horizon, but they do not support replacing the existing enhanced benchmark uniformly.

5.6. Champion selection and retrospective validation of the operational alert layer

The champion-selection results confirmed that no single forecasting family was consistently preferred across all cities, targets, and forecast horizons. For each city-target-horizon combination, the best-performing model within the local and enhanced benchmark families was first identified, after which the lower-MAE family winner was selected for operational use. For daily maximum AQI, the enhanced pooled benchmark supplied the selected champion in 14 of the 24 city-horizon combinations. For daily mean $PM_{2.5}$, the distribution was more horizon-dependent: enhanced models were selected more frequently at the 1-day and 3-day horizons, whereas local baselines were more competitive at the 2-day horizon. Across all 48 city-target-horizon combinations, the enhanced pooled benchmark supplied 28 champions, compared with 20 supplied by the local baseline benchmark.

These results produced a target-, city-, and horizon-specific deployment map rather than a single universally preferred forecasting model. The fixed champion selections were subsequently used to generate the historical out-of-sample AQI and $PM_{2.5}$ forecasts included in the retrospective alert evaluation.

The operational alert layer was retrospectively evaluated across all rolling-origin forecast cases generated by the fixed benchmark-defined champion selections. In total, the evaluation covered 432 city-origin-horizon cases, corresponding to 8 cities, 18 rolling origins, and 3 forecast horizons. Across these cases, the final alert level matched exactly in 73.1% of cases and was within one alert level in 92.4% of cases. The mean absolute alert-score gap was 0.35 alert levels overall. These results indicate that the alerting layer was able to reproduce much of the broad operational risk structure implied by the held-out gridded AQI and $PM_{2.5}$ fields, although not with uniform accuracy across all alert categories.

Performance remained reasonably stable across forecast horizons, although the 2-day horizon was weaker than the 1-day and 3-day horizons. Exact alert-level agreement was 75.0% at the 1-day horizon, 66.7% at the 2-day horizon, and 77.8% at the 3-day horizon, while within-one-level agreement was 92.4%, 86.8%, and 97.9%, respectively. The corresponding mean absolute alert-score gaps were 0.33, 0.49, and 0.25 alert levels. Table 11 summarizes these overall and horizon-specific validation metrics

Table 11. Retrospective validation summary of the operational alert layer over the full rolling-origin backtest.

Evaluation scope	N	Exact match	Within-one-level	Mean alert gap	Precision (High+)	Recall (High+)	Precision (Severe+)	Recall (Severe+)
Overall	432	0.731	0.924	0.354	0.545	0.197	1.000	0.059
1-day	144	0.75	0.924	0.326	0.600	0.273	1.000	0.143
2-day	144	0.667	0.868	0.486	0.500	0.167	NA	0.000
3-day	144	0.778	0.979	0.250	0.500	0.111	NA	0.000

Note: High+ denotes High-or-worse alerts, Severe+ denotes Severe-or-worse alerts, and Mean alert gap denotes the mean absolute difference between predicted and realized alert scores.

The confusion matrix helps clarify where agreement and disagreement were concentrated across alert categories. Agreement was strongest for the more common Watch and Warning states: 235 of 274 observed Watch cases (86%) and 76 of 97 observed Warning cases (78%) were recovered exactly. By contrast, higher-risk cases were more frequently downgraded. Among 44 observed High cases, only 4 (9%) were classified as High, while most were classified as Warning (22; 50%) or Watch (18; 41%). Among 17 observed Severe cases, only 1 (6%) was classified as Severe, while 7 (41%) were classified as High, 5 (29%) as Warning, and 4 (24%) as Watch. Consistent with this pattern, the binary event metrics indicate moderate precision but low recall for elevated-risk alerts. For High-or-worse alerts, precision was 54.5% and recall was 19.7% overall. For Severe-or-worse alerts, precision reached 100.0%, but recall was only 5.9%, reflecting both the rarity of these episodes and the tendency of the workflow to under-escalate the highest-risk cases. Fig. 3 reports the corresponding confusion matrix across final alert levels.

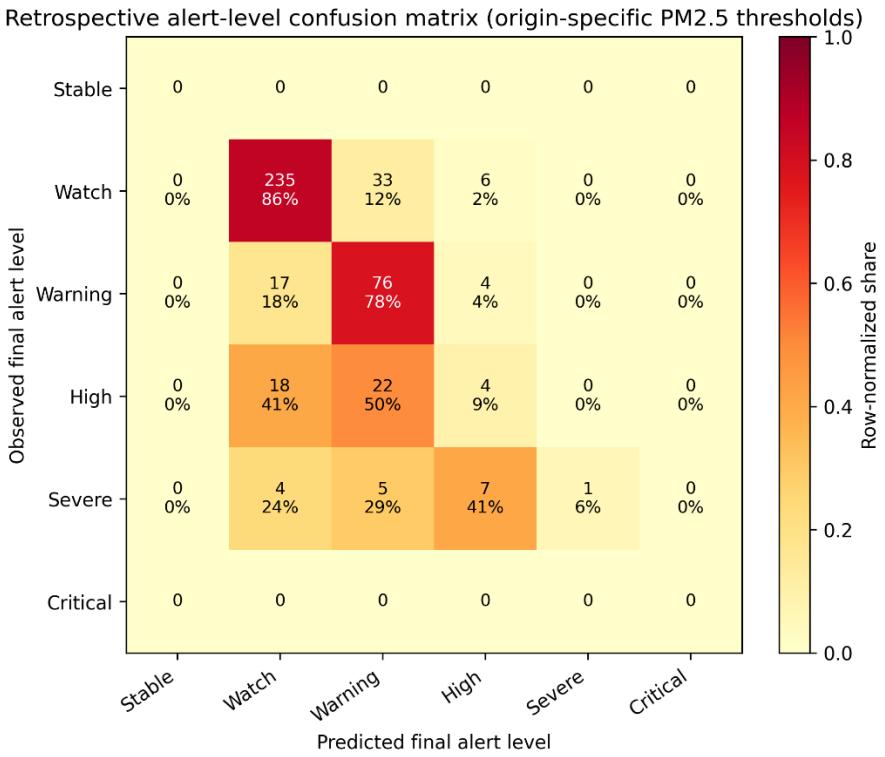

Figure 3. Retrospective confusion matrix for final operational alert levels across all rolling-origin forecast cases; cell labels report case counts and row-normalized shares.

Overall, the combined results show that benchmark-driven champion selection can be translated into a transparent operational alerting workflow. However, the limited recall for higher-risk episodes indicates that the resulting alerts are more appropriately interpreted as a city-level prioritization and triage mechanism over open gridded proxy fields than as a validated high-sensitivity public warning system.

5.7. Illustrative operational case study

The operational workflow was further illustrated through a fixed forecast snapshot issued on 28 April 2026 for the 29 April-1 May 2026 period. For each city, the workflow applied the benchmark-defined champion selections, generated AQI and $PM_{2.5}$ forecasts, converted these predictions into final alert levels, and retained the highest-risk forecast instance within the 1- to 3-day forecast window. Table 12 reports the resulting city-level outputs. In this snapshot, seven cities reached the Warning level, whereas Korce remained at Watch. Tirane ranked highest because it reached the largest predicted AQI value within the forecast window, followed by Vlore, Fier, and Berat. For the retained peak-risk rows, AQI provided the main driver of the final ranking, while $PM_{2.5}$ mainly acted as a supporting escalation signal within the study-defined operational rules.

Table 12. City attention ranking produced by the operational early-warning layer.

City	Final alert	Peak date	Horizon (days)	Peak AQI	Peak $PM_{2.5}$ (µg/m³)	AQI model	$PM_{2.5}$ model
Tirane	Warning	01-05-26	3	47.98	10.81	Holt Damped	HistGBM Global
Vlore	Warning	30-04-26	2	47.04	12.02	Drift	Naive
Fier	Warning	30-04-26	2	47.00	12.36	Naive	ARIMA
Berat	Warning	29-04-26	1	47.00	11.55	Naive	HistGBM Global
Durres	Warning	29-04-26	1	45.77	14.65	HistGBM Global	Ridge Global
Shkoder	Warning	29-04-26	1	43.26	11.36	HistGBM Global	HistGBM Global
Elbasan	Warning	29-04-26	1	42.89	11.65	HistGBM Global	HistGBM Global
Korce	Watch	29-04-26	1	40.00	10.71	Holt Damped	HistGBM Global

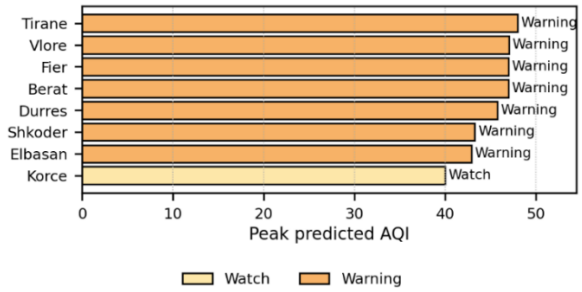

Figure 4. Operational city attention ranking based on the highest alert reached within the 1- to 3-day forecast window; bar length indicates peak predicted AQI, and bar color indicates final alert level.

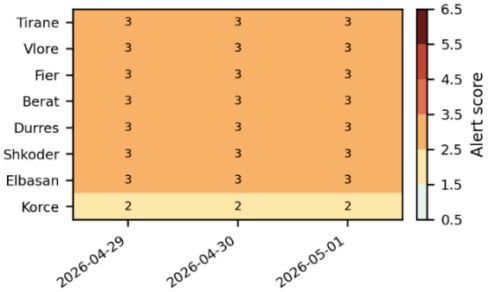

Figure 5. Heatmap of alert scores across cities and forecast dates for the fixed case-study forecast; darker colors indicate higher alert levels in the operational alerting layer.

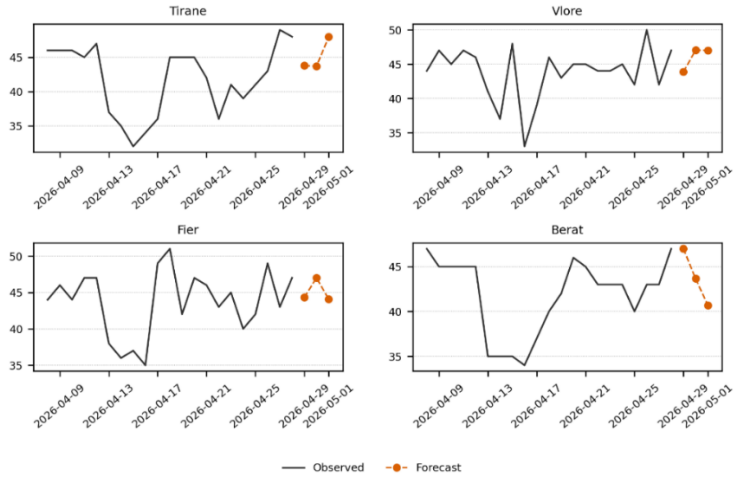

Figure 6. Observed and forecast daily AQI trajectories for the four highest-priority cities in the operational ranking.

Fig. 4–6 visualize the operational ranking, temporal alert distribution, and forecast trajectories for the highest-priority cities.

Together, these outputs illustrate how benchmarked forecasting models can be translated into interpretable operational early-warning information for short-term urban air-quality monitoring.

6. Discussion

The principal finding of this study is not that one forecasting model was universally best, but that short-horizon air-quality forecasting in open-data urban settings is more appropriately treated as a benchmark-driven model-selection problem. The enhanced pooled feature-based benchmark performed more strongly for daily maximum AQI than for daily mean $PM_{2.5}$. In the cross-city average summary, AQI showed lower mean MAE under the enhanced benchmark at all three forecast horizons, whereas $PM_{2.5}$ results remained mixed and clearly horizon-dependent. However, the accompanying uncertainty analysis indicated that the AQI advantage was more limited inferentially than the point estimates alone might suggest, with the clearest unadjusted support appearing at the 3-day horizon. This pattern supports evaluating forecast performance separately by target and by horizon rather than expecting one model family to dominate across all forecasting tasks. At the same time, the absolute size of these gains should be kept in perspective. For AQI, the mean MAE reductions were only about 0.20-0.30 points, and for $PM_{2.5}$ the only favorable enhanced-model contrast was 0.1145 $\mu g/m^3$ at the 1-day horizon. These are modest differences relative to the wider AQI communication bands and are therefore more meaningful for benchmark-guided model choice, near-threshold cases, and city prioritization than as evidence of large routine changes in operational alert decisions.

The supplementary analyses help clarify why the two targets behaved differently. Within the enhanced pooled benchmark, meteorological covariates made the most consistent contribution, while cross-pollutant context was more useful at shorter horizons and city indicators had a smaller and less consistent role. The pooled cross-city HistGBM model also remained stronger than the same model trained separately for each city, suggesting that the gains of the enhanced benchmark were not due only to richer local predictors. Taken together, these findings indicate that the stronger results arose from the combination of structured temporal features, meteorological context, and pooled training rather than from any single component alone. At the same time, these gains should not be interpreted as a clean causal demonstration of cross-city transfer by itself, because the pooled benchmark also differed from the local baseline family in predictor structure and model class.

The $PM_{2.5}$ results are especially important for interpretation. Although the enhanced benchmark improved the 1-day horizon, the local baseline family remained competitive at longer horizons, and the 3-day comparison was nearly neutral in MAE terms. This reinforces a practical point: more flexible or more complex models should not be adopted simply because they appear stronger in principle. The supplementary XGBoost comparisons support the same conclusion from another angle. They show that alternative tree-based global learners can be competitive for specific target-horizon combinations, but they do not justify replacing the benchmark logic with a single new preferred model.

The practical value of the framework therefore lies less in identifying one winning algorithm and more in linking benchmark evidence to deployment decisions. Model-family wins varied across targets and horizons rather than converging on a single dominant family, and the champion-selection layer translated that heterogeneity into an operationally usable deployment map.

An alternative operational design would be to combine forecasts across models rather than selecting a single champion for each city-target-horizon case. Such ensemble or model-averaging approaches can improve robustness in settings where no single model family dominates consistently. The present study did not evaluate forecast combinations, because its primary aim was to preserve transparent benchmark-to-deployment logic through explicit model-family selection. However, ensemble-based alerting remains a plausible extension for future work, particularly if the goal shifts from interpretable winner selection toward greater stability under model uncertainty.

The retrospective alert evaluation further showed that the benchmark-to-alert translation can recover much of the broad low-to-moderate risk structure present in the held-out gridded forecast cases, even though the alert layer was less sensitive for rarer higher-risk episodes and tended to under-escalate the most severe cases. Accordingly, the framework is best interpreted as a transparent prioritization and triage workflow over open gridded proxy fields rather than as a validated high-sensitivity public warning service. A further qualification is that the present alerting layer is based on deterministic point forecasts only and does not yet propagate forecast-specific uncertainty into the operational rules. The confidence intervals reported earlier in the benchmark section quantify uncertainty in average model-comparison differences, not prediction intervals for individual future AQI or $PM_{2.5}$ values. As a result, the current alerts should be interpreted as deterministic prioritization outputs rather than as probabilistic statements of exceedance risk. In future operational versions, prediction intervals or probabilistic forecasting approaches could support uncertainty-aware alerting, especially for near-threshold cases where escalation decisions may be sensitive to relatively small forecast differences. Even with that qualification, the study moves beyond pure forecast comparison by showing how reproducible benchmarking, explicit model-family choice, and downstream alert logic can be integrated within a transparent workflow for smaller open-data urban settings.

7. Limitations and future work

Several limitations should be acknowledged. First, the air-quality and meteorological inputs used in this study are open gridded products rather than regulatory ground-station measurements (Open-Meteo, 2026a, 2026b). The framework should therefore be interpreted as a reproducible forecasting workflow over modeled environmental fields rather than as direct observational validation of urban air quality. Second, the analysis covers only eight cities and a relatively short study period, which limits the number and diversity of rare or extreme pollution episodes available for model training and evaluation. Third, although the alerting layer was evaluated retrospectively over the rolling-origin backtest, it has not yet been externally validated against regulatory monitoring observations or prospectively tested as a public warning service.

These limitations suggest several directions for future work. A first priority is external validation against regulatory monitoring stations, where such observations are available,

together with prospective testing in a live forecasting setting. A second priority is to extend the framework through richer exogenous inputs and stronger spatial representations, including satellite products, low-cost sensor streams, traffic or emissions proxies, and hybrid regional models (Agbehadji and Obagbuwa, 2024; Li et al., 2025; Hu et al., 2025). Operational reliability should also be strengthened by incorporating prediction intervals, probabilistic exceedance estimates, or other uncertainty-aware forecasting outputs directly into the alert logic, particularly for near-threshold cases. Further work should additionally examine online performance monitoring, forecast combinations, and more explicit links between predicted risk levels and environmental-health response actions (Rajesh et al., 2025; Ke et al., 2022; Rahman et al., 2024; Strom and Gundersen, 2024).

8. Conclusion

This study developed a reproducible benchmark-driven workflow for short-horizon urban air-quality forecasting over open gridded environmental fields in eight Albanian cities. Rather than identifying one universally best forecasting model, the study showed that model choice was target- and horizon-dependent. The enhanced pooled feature-based benchmark performed more strongly for daily maximum AQI, whereas daily mean $PM_{2.5}$ remained more mixed and continued to justify comparison against strong local baselines. Taken together with the uncertainty analysis, these results support cautious benchmark-based model selection rather than reliance on a single universally preferred forecasting family.

The supplementary analyses further showed that the stronger enhanced results were associated with the combined contribution of structured temporal predictors, meteorological covariates, and pooled cross-city training, rather than with any single model component in isolation. At the same time, the city-specific and XGBoost comparisons reinforced the broader benchmark conclusion that no single forecasting family should be assumed to dominate across all targets and horizons.

The retrospective alert evaluation extended the study beyond forecast comparison alone by testing how benchmark-selected forecasts translated into operational alert levels over the rolling-origin backtest. These results support the workflow as a transparent city-level prioritization and triage mechanism over open gridded proxy fields, while also showing that sensitivity remained limited for rarer higher-risk episodes.

Accordingly, the contribution of the study lies in integrating reproducible benchmarking, explicit model-family selection, and downstream alert logic within a practical open-data forecasting workflow, while external validation against regulatory observations remains an important next step.

Code and Data Availability

Code and materials associated with this study are publicly available at https://github.com/AlketaHyso/Forecast-Urban-Air-Quality. The archived release corresponding to the reported workflow is Version 1.1.1, preserved on Zenodo at https://doi.org/10.5281/zenodo.21327666. Additional reproducibility details, including the API query structure, preprocessing logic, benchmark configuration, and software environment, are reported in Supplementary Note S1.

References

Achebak, H., Garatachea, R., Pay, M. T., Jorba, O., Guevara, M., Pérez García-Pando, C., Ballester, J. (2024). Geographic sources of ozone air pollution and mortality burden in Europe, *Nature Medicine* 30, 1732–1738. https://doi.org/10.1038/s41591-024-02976-x

Agbehadji, I. E., Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction, *Atmosphere* 15(11), 1352. https://doi.org/10.3390/atmos15111352

Cengil, E. (2025). The power of machine learning methods and PSO in air quality prediction, *Applied Sciences* 15(5), 2546. https://doi.org/10.3390/app15052546

Chen, Z.-Y., Petetin, H., Méndez Turrubiates, R. F., Achebak, H., Pérez García-Pando, C., Ballester, J. (2024). Population exposure to multiple air pollutants and its compound episodes in Europe, *Nature Communications* 15, 2094. https://doi.org/10.1038/s41467-024-46103-3

Cheng, P., Wei, C., Zhang, J., Wang, H. (2025). Air quality index forecasting based on quadratic decomposition and Transformer-BiLSTM: A case study of Beijing, *Atmosphere* 16(12), 1334. https://doi.org/10.3390/atmos16121334

Diachenko, G., Laktionov, I., Vizniuk, A., Gorev, V., Kashtan, V., Khabarlak, K., Shedlovska, Y. (2024). An improved approach to prediction of maize disease occurrence based on weather monitoring and machine learning: Case of the forest-steppe and northern steppe of Ukraine, *Baltic Journal of Modern Computing* 12(4), 387–414. https://doi.org/10.22364/bjmc.2024.12.4.03

European Environment Agency. (2026). *European air quality index*. https://airindex.eea.europa.eu/AQI/

Gao, Z., Mo, X., Li, H. (2024). Prediction of PM2.5 concentration based on deep learning, multi-objective optimization, and ensemble forecast, *Sustainability* 16(11), 4643. https://doi.org/10.3390/su16114643

Hewamalage, H., Ackermann, K., Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices, *Data Mining and Knowledge Discovery* 37, 788–832. https://doi.org/10.1007/s10618-022-00894-5

Hu, M., Lu, X., Chen, Y., Li, Z., Wang, Y., Fung, J. C. H. (2025). AirQFormer: Improving regional air quality forecast with a hybrid deep learning model, *Sustainable Cities and Society* 119, 106113. https://doi.org/10.1016/j.scs.2024.106113

Hu, Y., Cao, N., Guo, W., Chen, M., Rong, Y., Lu, H. (2024). FedDeep: A federated deep learning network for edge assisted multi-urban PM2.5 forecasting, *Applied Sciences* 14(5), 1979. https://doi.org/10.3390/app14051979

Houdou, A., El Badisy, I., Khomsi, K., Haziti, M. E., El Merabet, Y. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review, *Aerosol and Air Quality Research* 24, 230151. https://doi.org/10.4209/aaqr.230151

Januschowski, T., Wang, Y., Torkkola, K., Erkkila, T., Hasson, H., Gasthaus, J. (2022). Forecasting with trees, *International Journal of Forecasting* 38(4), 1473–1481. https://doi.org/10.1016/j.ijforecast.2021.10.004

Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., Zhang, H. (2022). Development and application of an automated air quality forecasting system based on machine learning, *Science of the Total Environment* 806, 151204. https://doi.org/10.1016/j.scitotenv.2021.151204

Lashko, Y., Sukach, S., Laktionov, I., Chencheva, O., Rieznik, D., Korostova, O. (2025). Predictive mathematical and computer model for determining harmful effects of dust pollution on the environment and workers, *Baltic Journal of Modern Computing* 13(2), 436–452. https://doi.org/10.22364/bjmc.2025.13.2.08

Lazo, P., Kane, S. S., Qarri, F., Allajbeu, S., Bekteshi, L. (2024). 15 years of moss biomonitoring for air quality assessment in Albania, *Aerosol and Air Quality Research* 24, 240011. https://doi.org/10.4209/aaqr.240011

Liu, Q., Cui, B., Liu, Z. (2024). Air quality class prediction using machine learning methods based on monitoring data and secondary modeling, *Atmosphere* 15(5), 553. https://doi.org/10.3390/atmos15050553

Makhdoomi, A., Sarkhosh, M., Ziaei, S. (2025). PM2.5 concentration prediction using machine learning algorithms: An approach to virtual monitoring stations, *Scientific Reports* 15, 8076. https://doi.org/10.1038/s41598-025-92019-3

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2022). The M5 competition: Conclusions, *International Journal of Forecasting* 38(4), 1576–1582. https://doi.org/10.1016/j.ijforecast.2022.04.006

Mendez, M., Merayo, M. G., Nunez, M. (2023). Machine learning algorithms to forecast air quality: A survey, *Artificial Intelligence Review* 56, 10031–10066. https://doi.org/10.1007/s10462-023-10424-4

Meng, X., Xie, C., Tang, X., Pan, Y. (2025). Prediction of particulate matter 2.5 concentration based on attention mechanism and convolutional BiLSTM network, *Discover Applied Sciences* 7, 1372. https://doi.org/10.1007/s42452-025-07891-5

Middya, A. I., Roy, S. (2022). Pollutant specific optimal deep learning and statistical model building for air quality forecasting, Environmental Pollution 301, 118972. https://doi.org/10.1016/j.envpol.2022.118972

Open-Meteo. (2026a). *Air quality API*. https://open-meteo.com/en/docs/air-quality-api

Open-Meteo. (2026b). *Historical weather API*. https://open-meteo.com/en/docs/historical-weather-api

Ozupak, Y., Alpsalaz, F., Aslan, E. (2025). Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction, *Water, Air, & Soil Pollution* 236, 464. https://doi.org/10.1007/s11270-025-08122-8

Petrić, V., Hussain, H., Časni, K., Vuckovic, M., Schopper, A., Ujević Andrijić, Ž., Kecorius, S., Madueno, L., Kern, R., Lovrić, M. (2024). Ensemble machine learning, deep learning, and time series forecasting: Improving prediction accuracy for hourly concentrations of ambient air pollutants, *Aerosol and Air Quality Research* 24, 230317. https://doi.org/10.4209/aaqr.230317

Rahman, M. M., Nayeem, M. E. H., Ahmed, M. S., Tanha, K. A., Sakib, M. S. A., Uddin, K. M. M., Babu, H. M. H. (2024). AirNet: Predictive machine learning model for air quality forecasting using web interface, *Environmental Systems Research* 13, 44. https://doi.org/10.1186/s40068-024-00378-z

Rajesh, M., Ganesh Babu, R., Moorthy, U., Easwaramoorthy, S. V. (2025). Machine learning-driven framework for real-time air quality assessment and predictive environmental health risk mapping, *Scientific Reports* 15, 28801. https://doi.org/10.1038/s41598-025-14214-6

Strom, E., Gundersen, O. E. (2024). Performance metrics for multi-step forecasting measuring win-loss, seasonal variance and forecast stability: An empirical study, *Applied Intelligence* 54, 10490–10515. https://doi.org/10.1007/s10489-024-05715-4

Wang, Y., Yazdi, M. D., Wei, Y., Schwartz, J. D. (2024). Air pollution below US regulatory standards and cardiovascular diseases using a double negative control approach, *Nature Communications* 15, 8451. https://doi.org/10.1038/s41467-024-52117-8

Received May 26, 2026, revised July 13, 2026, accepted August 17, 2026