EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(10), 118–133
https://doi.org/10.12912/27197050/231162
ISSN 2719–7050, License CC-BY 4.0

Received: 2026.07.27
Accepted: 2026.08.31
Published: 2026.09.10

Explainable multi-horizon fine particulate matter forecasting with spatial transfer and local calibration in Quito, Ecuador

David Elías Dáger López[1]

[1] Faculty of Sciences and Engineering, Universidad Estatal de Milagro, Ciudadela Universitaria Dr. Rómulo Minchala Murillo, km 1.5 vía Milagro–Virgen de Fátima, Postal Code 091050, Milagro, Guayas, Ecuador
E- mail: ddagerl@unemi.edu.ec

ABSTRACT

Accurate short-term particulate-matter forecasting is needed for proactive exposure management, but reported performance often relies on random data splits or station-specific models that do not reveal temporal leakage, spatial distribution shift, or forecast uncertainty. This study developed an explainable and uncertainty-aware framework for hourly fine particulate matter (PM2.5) forecasting across five stations of the Metropolitan Air Quality Monitoring Network of Quito, Ecuador. The harmonized 2019–2025 panel contained 306,840 station-hour records. Models were trained on 2019–2022, selected using 2023 only, and evaluated on an untouched 2024–2025 test set for 1-, 6-, 12-, and 24-h horizons. CatBoost, selected before test evaluation, obtained RMSE values of 6.48, 7.78, 8.06, and 8.04 μg m^{-3}, respectively, and significantly outperformed persistence, seasonal persistence, and climatology. For observations above 25 μg m^{-3}, RMSE increased to 11.64–16.18 μg m^{-3} and 88.2–98.3% of cases were underpredicted, revealing a systematic upper-tail limitation. Split-conformal intervals calibrated on 2023 achieved 89.11–90.43% coverage at the 90% nominal level and 95.15–95.51% at the 95% level. Under strict leave-one-station-out transfer, CatBoost reduced RMSE relative to persistence by 15.48%, 31.36%, 34.36%, and 19.13%, but failed to outperform persistence for Tumbaco at 1 h. Conformal coverage at unseen stations fell to 87.12% at the 90% level for 1-h forecasts, diagnosing spatial shift. SHAP and ablation analyses showed that recent PM2.5 history was the dominant information source, with meteorology providing its largest incremental benefit at 1 h. A smoothed hourly-bias calibration using 90 local days reduced aggregate zero-shot RMSE by 7.65%, 6.32%, 5.05%, and 2.94%; in Tumbaco, reductions reached 29.99%, 20.35%, 16.75%, and 12.31%. The framework offers a reproducible basis for transferable forecasting and local calibration, but the pronounced underestimation of high concentrations precludes interpreting it as a stand-alone warning system without additional extreme-event safeguards.

Keywords: fine particulate matter, air-quality forecasting, explainable machine learning, conformal prediction, spatial transfer, local calibration, tropical Andes.

INTRODUCTION

Fine particulate matter with an aerodynamic diameter below 2.5 μm (PM2.5) is a major environmental-health concern because it penetrates deeply into the respiratory system and is associated with cardiovascular, respiratory, and premature-mortality risks. The 2021 World Health Organization (WHO) air-quality guidelines recommend an annual PM2.5 concentration of 5 μg m^{-3} and a 24-h concentration of 15 μg m^{-3}, emphasizing that these health-based reference values are not equivalent to national regulatory limits (World Health Organization, 2021). Forecasts that anticipate elevated concentrations can complement monitoring by supporting alerts, short-term emission-control actions, and risk communication.

Hourly PM2.5 dynamics are difficult to forecast because they arise from a nonlinear combination of local emissions, atmospheric stability, wind, precipitation, radiation, chemical transformations, and temporal persistence.

Artificial-intelligence approaches have therefore become prominent in air-pollution forecasting (Chadalavada et al., 2025). Reviews of PM2.5 prediction show rapid growth in recurrent networks, convolutional architectures, transformers, decomposition hybrids, and ensemble methods, but also identify recurring weaknesses: inconsistent experimental protocols, limited interpretability, data leakage, and insufficient external validation (Chadalavada et al., 2025; Zhou et al., 2024). Recent studies have achieved strong short-term performance using ensembles and spatiotemporal fusion (Feng et al., 2024; Gutiérrez-Avila et al., 2022), yet numerical comparisons across cities remain difficult because pollutant distributions, forecast horizons, temporal splits, and evaluation rules differ substantially.

Urban air quality can respond rapidly to changes in anthropogenic activity. During COVID-19 lockdown conditions in Karachi, Ali et al. (2024) reported a 53.55% decrease in mean PM2.5, illustrating how abrupt reductions in mobility, commercial activity, and industrial activity can materially alter urban particulate-matter levels.

These limitations are particularly relevant for tropical Andean cities. Quito lies in a narrow high-altitude basin at approximately 2,850 m above sea level, where complex topography, intense solar radiation, localized emissions, and heterogeneous urban form shape pollutant dispersion (Valencia et al., 2020). Previous research in Quito has investigated urban-background modelling and remote-sensing-based particulate-matter estimation (Alvarez-Mendoza et al., 2019; Valencia et al., 2020). The official Metropolitan Air Quality Monitoring Network (Red Metropolitana de Monitoreo Atmosférico de Quito, REMMAQ) provides a valuable long-term hourly record (Secretaría de Ambiente del Distrito Metropolitano de Quito, n.d.), but a forecasting model intended for decision support must still demonstrate that it generalizes to future periods and across stations with distinct concentration regimes.

Spatial transfer is a central but underexamined problem. Models trained at data-rich sites are increasingly transferred to data-limited monitoring stations (Jairi et al., 2024; Yao et al., 2024). Nevertheless, transfer accuracy can deteriorate when the target station differs in baseline concentration, diurnal cycle, meteorology, or source composition. A model may therefore appear accurate in a pooled random split while failing at an unseen station. Strict leave-one-station-out (LOSO) evaluation provides a direct test of this failure mode, especially when the station identifier is removed and the target station is excluded from all parameter estimation.

Point forecasts alone are also insufficient for environmental decision-making. Probabilistic and diffusion-based air-quality models have highlighted the need to quantify forecast uncertainty (Chen et al., 2024; Murad et al., 2021). Conformal prediction offers distribution-free marginal coverage under exchangeability and can be applied to any point predictor (Angelopoulos and Bates, 2023). However, its nominal guarantees can weaken under covariate or distribution shift (Gibbs and Candès, 2021; Tibshirani et al., 2019). Measuring conformal coverage during spatial transfer can therefore serve not only as an uncertainty assessment but also as a diagnostic of whether the target station behaves like the source network.

Interpretability is equally important. SHAP values provide additive feature attributions for complex models (Lundberg and Lee, 2017), while feature-group ablation determines whether improved accuracy comes from pollutant history, meteorology, co-pollutants, or merely calendar effects. Together, these analyses help distinguish predictive association from mechanistic inference and clarify which sensors are most informative for monitored-network forecasting.

This study addresses these gaps using seven years of official hourly data from five REMMAQ stations. The objectives were to: (i) develop multi-horizon PM2.5 forecasts with strict chronological separation; (ii) select the model using validation data only and evaluate it on an untouched two-year test set; (iii) quantify uncertainty with split conformal prediction; (iv) assess strict LOSO transfer and spatial calibration failure; (v) explain predictions using SHAP and feature-group ablation; and (vi) test whether 7, 30, or 90 days of local observations can improve transferred forecasts through a simple, robust calibration. The study was designed to prioritize reproducibility and realistic generalization rather than maximize performance under a favorable split.

MATERIALS AND METHODS

Study area and monitoring network

Quito is located in the tropical Andes of Ecuador and extends longitudinally along an elevated,

topographically constrained urban corridor. Its atmospheric environment combines high solar irradiance, marked terrain effects, heterogeneous land use, and spatially varying traffic and combustion sources (Valencia et al., 2020). These conditions make it a relevant test bed for evaluating whether data-driven forecasting relationships transfer across monitoring locations.

Hourly records were obtained from REMMAQ, administered by the Secretaría de Ambiente of the Metropolitan District of Quito (Secretaría de Ambiente del Distrito Metropolitano de Quito, n.d.). Five stations were retained as the core panel because they provided the most continuous PM2.5 observations and complementary meteorological records during 2019–2025: Belisario, Carapungo, Centro, Cotocollao, and Tumbaco. Other stations were excluded from the main panel because entire years or key predictor groups were largely missing, which would have weakened chronological and spatial comparisons.

Data acquisition, harmonization, and quality control

The source archive included PM2.5, PM10, NO_2, O_3, CO, SO_2, air temperature, relative humidity, atmospheric pressure, precipitation, solar radiation, wind speed, and wind direction. Pollutant and meteorological files were stored separately and had different date conventions. Meteorological timestamps provided in Coordinated Universal Time were converted to America/Guayaquil local time (UTC−5) before hourly alignment. Station names and column labels were standardized, duplicated station-hour records were removed, and variables were screened for physically impossible values. No invalid PM2.5 values were removed during the final physical-range audit.

The resulting balanced station-time index contained 306,840 rows (61,368 hourly positions per station) from 1 January 2019 through 31 December 2025. Missing observations were retained as null values. The PM2.5 target was never imputed. Tree-based models used their native missing-value handling; Ridge regression used median imputation estimated from the training set. PM10 was excluded from the primary forecasting feature set because useful multiyear coverage was largely confined to Carapungo. Wind direction was converted to sine and cosine components; its absence at Centro was retained as missing rather than synthetically reconstructed.

Descriptive daily means were calculated only when at least 18 valid hourly PM2.5 observations were available, corresponding to 75% daily completeness. The percentage of valid days above 15 $\mu g\ m^{-3}$ was reported relative to the WHO 24-h guideline for health interpretation, not as a claim of regulatory non-compliance (World Health Organization, 2021).

Forecast targets and feature engineering

For each station and forecast origin t, the targets were PM2.5(t+h) for h = 1, 6, 12, and 24 h. The predictor set included the current PM2.5 measurement at t; PM2.5 lags of 1, 2, 3, 6, 12, 24, 48, and 72 h; rolling means over the previous 3, 6, 12, and 24 h; NO_2, O_3, CO, and SO_2; temperature, humidity, pressure, precipitation, solar radiation, wind speed, and wind-direction components; weekend status; and cyclic encodings of hour of day and day of year.

All rolling statistics ended at t−1, so they contained no target-period information. The current PM2.5 value was a legitimate real-time predictor available at the forecast origin. Cyclic variables were represented with sine and cosine transformations to preserve adjacency between 23:00 and 00:00 and between the end and beginning of the year. In pooled models, station identity was represented by binary indicators. Station identity was removed completely in the LOSO experiment.

Chronological experimental design

The temporal partition was fixed before model comparison: 2019–2022 for training, 2023 for model selection, early stopping, and conformal calibration, and 2024–2025 for final testing. Random splitting was not used. The two-year test period remained untouched during algorithm selection. This separation is important because random station-hour splits can place near-duplicate temporal states in both training and test sets and overstate prospective generalization performance (Figure 1).

Baselines and machine-learning models

Three forecasting baselines were implemented. Persistence predicted the future value using PM2.5(t). Seasonal persistence used the observation aligned with the same target hour on the preceding day. Station-month-hour climatology

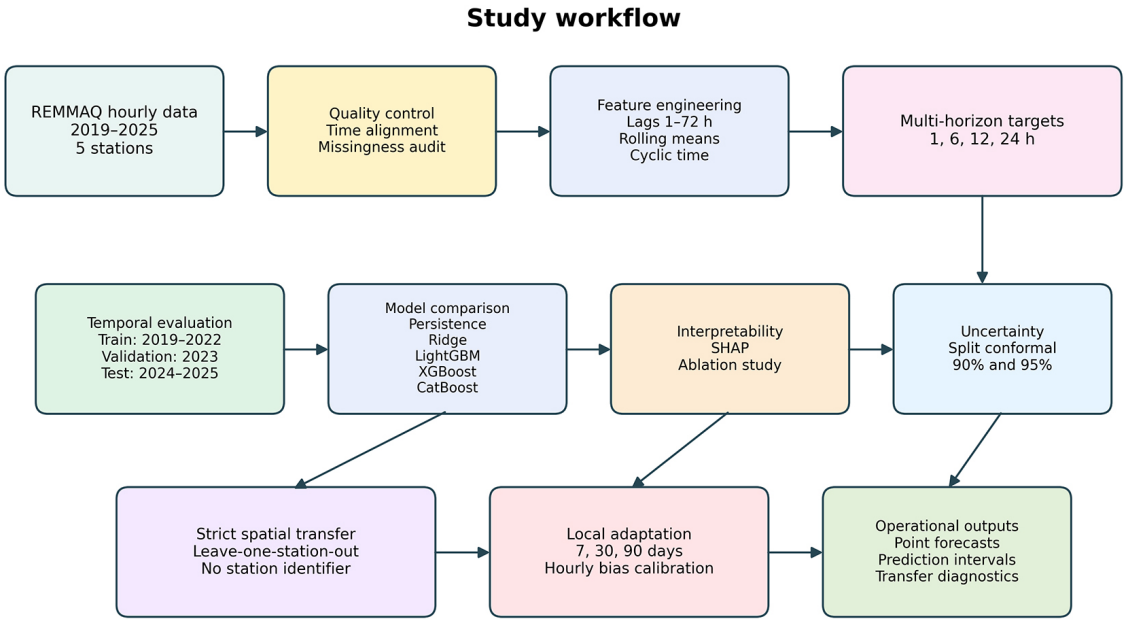

Figure 1. End-to-end study workflow. The design combines chronological testing, multi-horizon point forecasts, explainability, split-conformal uncertainty, strict leave-one-station-out transfer, local calibration, and decision-support diagnostics

was estimated exclusively from the training period and evaluated at the target timestamp. These baselines represent progressively stronger temporal references and prevent overstating the value of complex algorithms.

Ridge regression provided a regularized linear comparator. Three gradient-boosted tree algorithms were evaluated: LightGBM (Ke et al., 2017), XGBoost (Chen and Guestrin, 2016), and CatBoost (Prokhorenkova et al., 2018). Gradient boosting constructs an additive ensemble of weak learners that sequentially reduces residual loss. Ridge used $\alpha = 10$ after standardized median-imputed numerical predictors and one-hot station encoding. Tree models used a learning rate of 0.05, subsampling and feature-subsampling controls, and early stopping on 2023. CatBoost used depth 7, L2 regularization 3, random strength 0.5, a maximum of 650 iterations, and a 60-iteration early-stopping patience. The same model family was retained for all main analyses if it achieved the lowest validation RMSE.

CatBoost was selected for all four horizons using 2023 only. In the independent test period, the smallest post hoc RMSE was obtained by LightGBM at 1 and 6 h, XGBoost at 12 h, and Ridge at 24 h. These minima differed from CatBoost by only 0.003–0.013 μg m^{-3} and were not used to alter the pre-specified model choice. This preserves the independence of the final test evaluation.

Evaluation metrics and statistical comparison

Forecast accuracy was summarized by mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), symmetric mean absolute percentage error (sMAPE), and mean bias, defined as the mean of predicted minus observed concentrations. Elevated-concentration performance was evaluated for observations above 15 and 25 μg m^{-3}. The metrics were calculated using Equations 1–4.

$$MAE = (1/n)\,\Sigma_i\,|y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{[(1/n)\,\Sigma_i\,(y_i - \hat{y}_i)^2]} \quad (2)$$

$$R^2 = 1 - [\Sigma_i\,(y_i - \hat{y}_i)^2 / \Sigma_i\,(y_i - \bar{y})^2] \quad (3)$$

$$sMAPE = (100/n)\,\Sigma_i\,[2|y_i - \hat{y}_i| / (|y_i| + |\hat{y}_i|)] \quad (4)$$

Forecast losses were compared with a Diebold–Mariano-type out-of-sample predictive-ability test using squared-error loss differentials and a Newey–West-type autocorrelation correction with a lag related to the forecast horizon (Giacomini and White, 2006). Statistical significance was assessed at $\alpha = 0.05$, but practical effect size was interpreted alongside p-values because the large number of hourly observations can make negligible differences statistically significant.

Sampling uncertainty in the primary test RMSE was quantified with a synchronous moving-block bootstrap (Kreiss and Lahiri, 2012). Squared errors were aggregated by hourly timestamp across the five stations, 168-h circular blocks were resampled to preserve weekly serial dependence and cross-station co-movement, and 2,000 bootstrap replicates were generated with a fixed seed. Percentile 95% confidence intervals describe test-period sampling variability conditional on the fitted CatBoost model; they do not include hyperparameter-selection or model-training uncertainty.

Explainability and feature-group ablation

CatBoost predictions were interpreted with SHAP values (Lundberg and Lee, 2017). Mean absolute SHAP values were calculated for a reproducible random sample of 6000 test observations per horizon (random seed 2026). SHAP values were interpreted as contributions to model output, not causal effects.

Five cumulative feature configurations were compared using the same temporal split: (1) calendar variables only; (2) calendar plus PM2.5 history; (3) calendar, history, and meteorology; (4) the full model including gaseous pollutants; and (5) the full model without current PM2.5(t). The full model was compared with reduced configurations using the Diebold–Mariano test.

Split-conformal prediction intervals

Split conformal prediction was applied to absolute validation residuals (Angelopoulos and Bates, 2023). For nominal coverage $1-\alpha$, the finite-sample upper quantile q of $|y-\hat{y}|$ from 2023 was estimated, and the test interval was defined as $[\hat{y}-q, \hat{y}+q]$. Intervals were generated at 90% and 95% nominal coverage. Empirical coverage and mean width were reported.

Conformal prediction relies on exchangeability between calibration and test observations. Because station transfer can violate this condition, LOSO intervals were calibrated with 2023 residuals from the four source stations and evaluated at the held-out station. Departures from nominal coverage were interpreted as evidence of spatial distribution shift, consistent with conformal prediction under covariate shift (Gibbs and Candès, 2021; Tibshirani et al., 2019).

Strict spatial transfer

Generalization to an untrained station was evaluated with five LOSO folds. For each fold, 2019–2022 from four source stations were used for training, 2023 from those same stations for early stopping and interval calibration, and 2024–2025 from the held-out station for testing. The target station contributed no observations to model fitting or hyperparameter selection, and station identity was excluded.

The term zero-shot transfer refers to the absence of target-station labels during parameter estimation. The transferred predictor still used real-time and lagged measurements observed at the target station at the forecast origin. The experiment therefore evaluates parameter transfer across stations, not prediction at a completely sensor-free location. Spatial-transfer penalty was defined as the percentage RMSE change relative to the corresponding pooled CatBoost model evaluated at the same station.

Few-shot local calibration

To determine whether a small amount of local labelled data could correct spatial shift, the first 7, 30, or 90 days of 2024 at each held-out station were used to calibrate LOSO predictions. All strategies were evaluated on a common future period from 1 April 2024 through 31 December 2025, ensuring that longer calibration windows did not receive an easier test interval.

Three low-complexity corrections were considered: a global additive bias, a regularized linear slope-and-intercept correction, and a smoothed hourly bias. The hourly method estimated the mean residual for each target hour and shrank it toward the station-wide mean with a pseudocount of 12 observations. This approach was chosen over retraining a high-dimensional local model because preliminary few-day fits extrapolated unstable seasonal relationships. Adaptation was considered beneficial only when it improved a later period; no assumption was made that calibration must help every station.

Reproducibility

All data-processing, modelling, uncertainty, transfer, ablation, and adaptation steps were implemented in Python with fixed random seeds and pinned principal package versions.

The canonical processed dataset is Base_Datos_Unificada_REMMAQ_2019_2025.csv.gz (306,840 rows, 47 columns; SHA-256: d86e09c3e1ddde64c0203522b77142ffa00679f4167ca5f734736ebaea28babe). The synchronized clean and executed notebooks are REMMAQ_PM25_Reproduction_v1.0.ipynb and REMMAQ_PM25_Reproduction_v1.0_EXECUTED.ipynb; key numerical outputs, serialized models, predictions, and figure-generation artifacts are bundled as REMMAQ_PM25_Result_Artifacts_v1.0.zip. These materials are publicly archived in Zenodo as repository version 1.0.0 (https://doi.org/10.5281/zenodo.21958114). Raw observations remain attributable to the official REMMAQ repository (Secretaría de Ambiente del Distrito Metropolitano de Quito, n.d.).

RESULTS

Data coverage and exploratory patterns

The five-station panel comprised 306,840 station-hour rows. Overall PM2.5 availability was 96.41%, with station coverage ranging from 95.22% at Tumbaco to 97.82% at Cotocollao. Carapungo had the highest mean concentration (15.08 μg m^{-3}), whereas Tumbaco had the lowest (11.00 μg m^{-3}). The 95th percentile ranged from 24.70 μg m^{-3} at Tumbaco to 33.94 μg m^{-3} at Carapungo (Table 1).

The proportion of valid daily means above 15 μg m^{-3} was 45.20% at Carapungo, 42.93% at Belisario, 40.49% at Centro, 33.02% at Cotocollao, and 14.90% at Tumbaco. Diurnal profiles showed station-specific morning maxima between 07:00 and 09:00. The annual series also indicated substantial interannual variation, including a broad decrease in 2020 and station-dependent changes thereafter. These patterns support explicit temporal and spatial evaluation rather than a single random split.

PM2.5 showed the strongest global Spearman associations with NO_2 ($\rho = 0.39$) and SO_2 ($\rho = 0.38$). Its associations with solar radiation ($\rho = 0.21$), humidity ($\rho = -0.10$), and pressure ($\rho = -0.11$) were weaker. One-hour PM2.5 autocorrelation was approximately 0.60, and a daily signal reappeared at 24 h (approximately 0.38), justifying short lags, rolling histories, and daily cyclic predictors (Figure 2).

Chronological multi-horizon forecasting

CatBoost was selected at every horizon based solely on 2023 validation RMSE. On the 2024–2025 test period, CatBoost achieved RMSE values of 6.48, 7.78, 8.06, and 8.04 μg m^{-3} at 1, 6, 12, and 24 h, respectively (Table 2). The corresponding 95% moving-block-bootstrap intervals were [6.31, 6.66], [7.54, 8.01], [7.79, 8.32], and [7.76, 8.32] μg m^{-3} (Table 3). MAE values were 4.83, 5.89, 6.10, and 6.08 μg m^{-3}, R^2 values were 0.544, 0.340, 0.292, and 0.298, and sMAPE increased from 42.55% at 1 h to approximately 49–51% at longer horizons.

At 1 h, CatBoost reduced RMSE by 20.13% relative to persistence and 26.71% relative to climatology. At 6 and 12 h, persistence deteriorated sharply, and CatBoost reduced RMSE by 33.71% and 36.06%, respectively. At 24 h, persistence and seasonal persistence were identical by construction, and CatBoost reduced RMSE by 20.23%. Diebold–Mariano tests confirmed lower CatBoost squared-error loss than all three baselines at every horizon (all $p < 0.001$).

The three gradient-boosted tree models performed similarly, and Ridge was also competitive

Table 1. PM2.5 descriptive statistics and data completeness, 2019–2025

Station	Mean (μg m^{-3})	Median (μg m^{-3})	95th percentile (μg m^{-3})	Coverage (%)	Valid days >15 μg m^{-3} (%)
Belisario	14.61	12.78	33.78	96.20%	42.93%
Carapungo	15.08	13.32	33.94	95.81%	45.20%
Centro	14.32	12.76	31.33	96.98%	40.49%
Cotocollao	13.59	12.01	30.25	97.82%	33.02%
Tumbaco	11.00	9.70	24.70	95.22%	14.90%

Note: Concentrations are in μg m^{-3}. Daily means required at least 18 valid hourly measurements. The 15 μg m^{-3} threshold is the WHO 2021 24-h guideline and is used as a health reference, not a legal limit (World Health Organization, 2021).

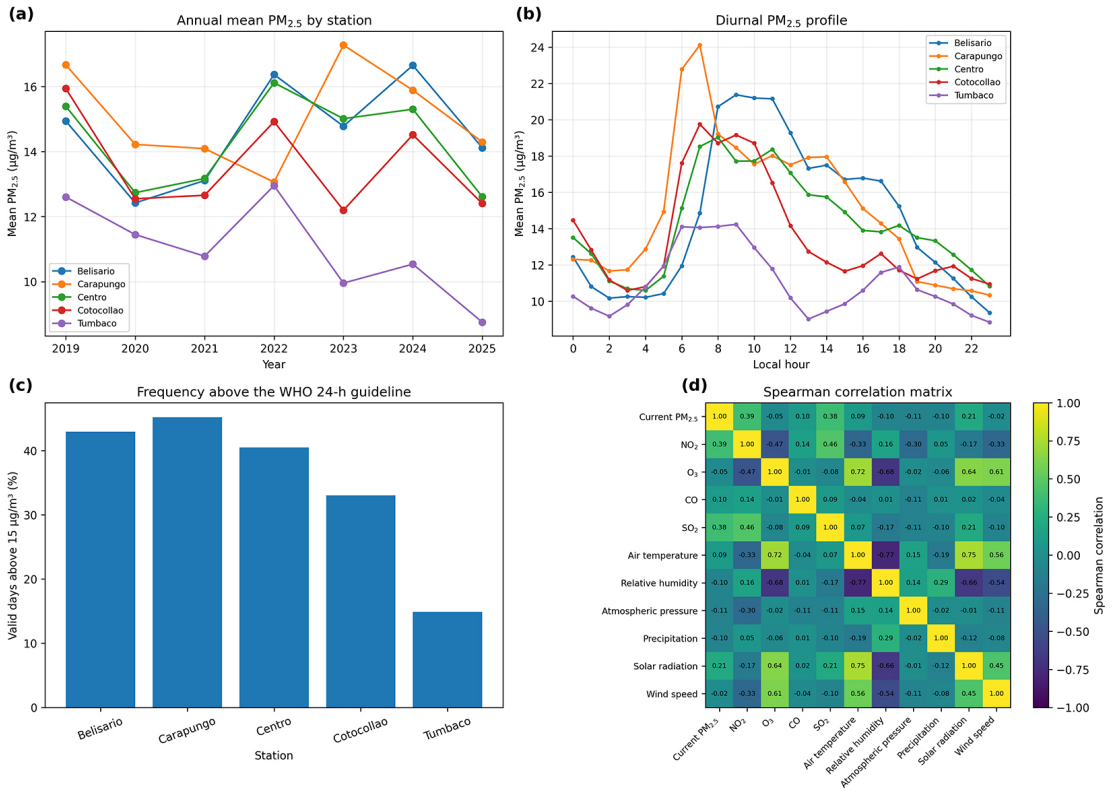

Figure 2. Exploratory PM2.5 patterns. (a) Annual mean by station; (b) diurnal profile; (c) frequency of valid daily means above the WHO 24-h guideline; and (d) Spearman correlation matrix. Correlations are descriptive and do not imply causality

at 24 h. Post hoc test-set minima were 6.470 µg m^{-3} for LightGBM at 1 h, 7.775 µg m^{-3} for LightGBM at 6 h, 8.056 µg m^{-3} for XGBoost at 12 h, and 8.034 µg m^{-3} for Ridge at 24 h. Because the model family was selected using validation data rather than test data, CatBoost remained the primary model. The very small 0.003–0.013 µg m^{-3} differences between CatBoost and the post hoc test minima indicate that the conclusions were not driven by a large performance gap among candidate implementations.

Performance during elevated concentrations

Approximately 35.2–35.4% of test observations exceeded 15 µg m^{-3} and 11.2–11.3% exceeded 25 µg m^{-3}, depending on the horizon-specific complete-case set. Error increased substantially in these strata (Table 4). For observations above 15 µg m^{-3}, MAE rose from 5.78 µg m^{-3} at 1 h to 7.82–7.83 µg m^{-3} at 12–24 h. Mean bias was negative at every horizon, ranging from −3.84 to

Table 2. Test RMSE (µg m−3) for all forecasting models

Model	1 h RMSE (µg m^{-3})	6 h RMSE (µg m^{-3})	12 h RMSE (µg m^{-3})	24 h RMSE (µg m^{-3})
Persistence	8.117	11.734	12.607	10.079
Seasonal persistence	10.054	10.029	10.029	10.079
Climatology	8.846	8.821	8.830	8.852
Ridge	6.887	8.239	8.191	8.034
LightGBM	6.470	7.775	8.058	8.048
XGBoost	6.498	7.805	8.056	8.051
CatBoost†	6.483	7.779	8.061	8.040

Note: † CatBoost was selected using 2023 validation data before evaluating 2024–2025. Test-set minima were not used to alter the model choice. Seasonal persistence equals persistence at 24 h.

Table 3. CatBoost point-forecast performance with block-bootstrap uncertainty

Horizon (h)	N	MAE (µg m−3)	RMSE (µg m^{-3}) [95% CI]	R^2	sMAPE (%)	Bias (µg m^{-3})
1	82,695	4.827	6.483 [6.309, 6.657]	0.544	42.55%	0.828
6	81,329	5.887	7.779 [7.543, 8.010]	0.340	49.28%	0.485
12	80,996	6.103	8.061 [7.794, 8.322]	0.292	50.82%	0.370
24	81,164	6.078	8.040 [7.764, 8.323]	0.298	50.60%	0.356

Note: MAE, RMSE, confidence limits, and bias are in µg m−3. Confidence intervals were obtained from 2.000 synchronous 168-h moving-block bootstrap replicates and are conditional on the fitted model.

−7.01 µg m^{-3}, indicating systematic underprediction when concentrations were elevated.

The upper tail was more challenging. For observations above 25 µg m^{-3}, RMSE was 11.64 µg m^{-3} at 1 h and increased to 15.08–16.18 µg m^{-3} at 6–24 h. The model underpredicted 88.24% of these cases at 1 h and 96.07–98.29% at the longer horizons; mean errors ranged from −8.38 to −13.88 µg m^{-3}. Thus, overall RMSE concealed a pronounced loss of sensitivity to high-pollution episodes, which limits the use of these forecasts as stand-alone warning signals.

Explainability and pooled uncertainty

SHAP patterns changed systematically with horizon. At 1 h, current PM2.5 was the dominant feature, followed by the 24-h rolling mean, NO_2, SO_2, and O_3. At 6 h, the 24-h rolling mean and cyclic hour were dominant, followed by current PM2.5, NO_2, and weekend status. At 12 h, cyclic hour, the 12- and 24-h rolling means, and the 12-h lag were most prominent. At 24 h, cyclic hour, current PM2.5, the 3-h rolling mean, and weekend status were among the leading predictors. These results indicate a transition from immediate state dependence at short lead times toward shared daily structure at longer horizons.

Pooled conformal coverage was close to nominal. At the 90% level, empirical coverage ranged from 89.11% at 1 h to 90.43% at 24 h; mean interval width increased from 19.62 to 24.94 µg m^{-3}. At the 95% level, coverage ranged from 95.15% to 95.51%, with mean widths from 25.69 to 32.07 µg m^{-3}. Thus, uncertainty widened with horizon, while calibration remained stable when source and target stations were represented in model development (Figure 3, Table 5).

Strict leave-one-station-out transfer

Under strict LOSO evaluation, aggregated CatBoost RMSE was 6.86, 8.05, 8.28, and 8.15 µg m^{-3} at 1, 6, 12, and 24 h (Table 6). Relative to persistence, these values corresponded to RMSE reductions of 15.48%, 31.36%, 34.36%, and 19.13%. CatBoost outperformed persistence in 19 of 20 station-horizon combinations.

The exception was Tumbaco at 1 h, where CatBoost RMSE was 6.79 µg m^{-3} and persistence RMSE was 5.51 µg m^{-3}. Tumbaco also showed negative R^2 under zero-shot transfer despite comparatively low absolute errors, indicating that the model did not reproduce its lower and less variable local distribution. Carapungo and Belisario had the largest absolute errors, reaching approximately 9.3 µg m^{-3} at intermediate horizons.

The mean RMSE penalty relative to pooled CatBoost decreased from 9.00% at 1 h to 2.22% at 24 h. This average concealed marked

Table 4. CatBoost performance during elevated hourly PM2.5 concentrations

Horizon (h)	N >15	MAE >15 (µg m^{-3})	Bias >15 (µg m^{-3})	N >25	RMSE >25 (µg m^{-3})	Bias >25 (µg m^{-3})	Underprediction >25 (%)
1	29.276	5.78	-3.84	9.365	11.64	-8.38	88.2%
6	28.719	7.39	-6.30	9.131	15.08	-12.63	96.1%
12	28.497	7.82	-7.01	9.073	16.16	-13.88	98.2%
24	28.617	7.83	-7.01	9.141	16.18	-13.88	98.3%

Note: Thresholds refer to observed hourly PM2.5 and are operational diagnostic strata, not hourly compliance limits. Errors and bias are in µg m−3; negative bias denotes underprediction.

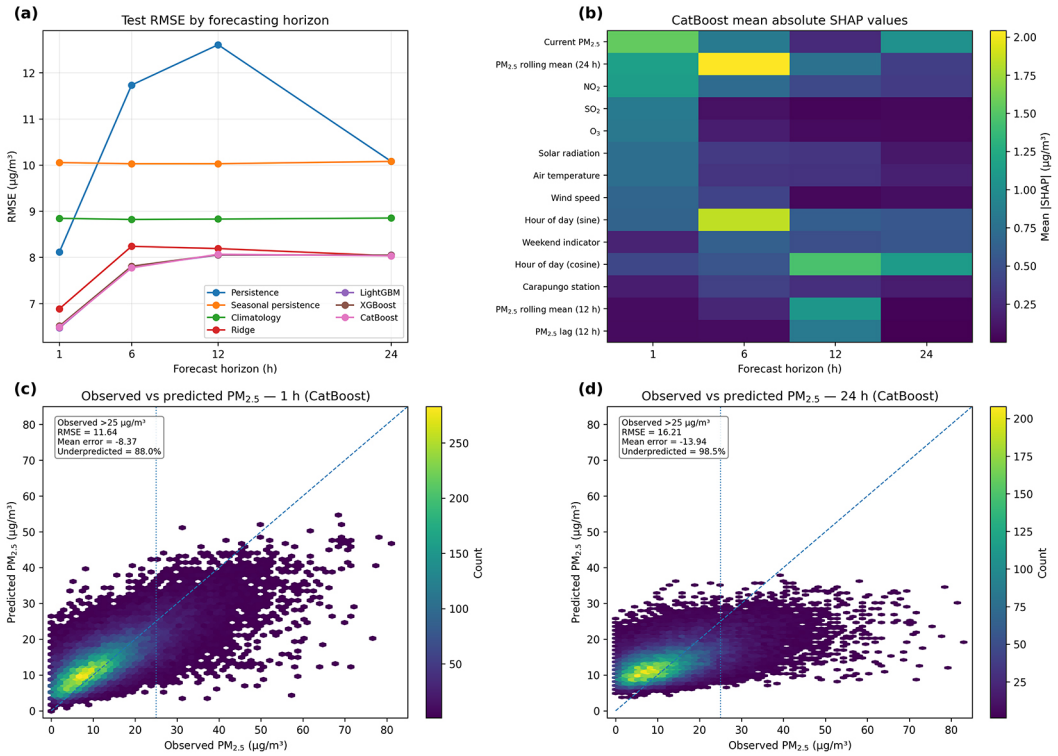

Figure 3. Pooled temporal forecasting results. (a) Test RMSE for baselines and machine-learning models; (b) CatBoost mean absolute SHAP values; (c) observed versus predicted PM2.5 at 1 h; and (d) observed versus predicted PM2.5 at 24 h. The dotted vertical line marks 25 µg m^{-3} observed PM2.5, and the dashed diagonal denotes perfect agreement

Table 5. Pooled split-conformal interval performance

Horizon (h)	90% coverage (%)	90% width (µg m^{-3})	95% coverage (%)	95% width (µg m^{-3})
1	89.11%	19.62	95.15%	25.69
6	89.81%	24.01	95.34%	31.00
12	90.15%	24.78	95.33%	31.80
24	90.43%	24.94	95.51%	32.07

Note: Interval widths are in µg m^{-3}. Residuals were calibrated on 2023 and evaluated on the independent 2024–2025 test period.

heterogeneity: Tumbaco penalties were 38.62%, 24.15%, 16.95%, and 10.09%, whereas Carapungo LOSO RMSE was slightly lower than the pooled station-specific RMSE at all horizons. Consequently, transfer difficulty was station-dependent rather than a uniform cost of withholding one site.

Uncertainty under spatial distribution shift

Conformal intervals calibrated on source-station residuals were less reliable after transfer. At 90% nominal coverage, aggregated LOSO coverage was 87.12%, 88.71%, 89.16%, and 90.03% for 1, 6, 12, and 24 h, respectively. At 95% nominal coverage, it ranged from 94.13% to 95.17%. Short-horizon undercoverage indicates that calibration residuals from source stations understated errors at some held-out stations.

The coverage deficit was most pronounced at locations with larger transfer penalties. This finding is consistent with the fact that ordinary split conformal prediction assumes calibration and test exchangeability (Angelopoulos and Bates, 2023; Tibshirani et al., 2019). In this application, empirical coverage served as a direct diagnostic of local mismatch: near-nominal performance was recovered at longer horizons, where shared cyclic structure contributed more and station-specific immediate dynamics contributed less (Figure 4).

Table 6. Aggregate strict leave-one-station-out performance

Horizon (h)	MAE LOSO (μg m^{-3})	RMSE LOSO (μg m^{-3})	R^2 LOSO	Persistence RMSE (μg m^{-3})	RMSE reduction (%)
1	5.146	6.860	0.489	8.117	15.48%
6	6.090	8.054	0.292	11.734	31.36%
12	6.275	8.275	0.254	12.607	34.36%
24	6.164	8.150	0.279	10.079	19.13%

Note: Each fold trained on four stations and tested on the fifth. The held-out station identifier and labels were not used for model fitting.

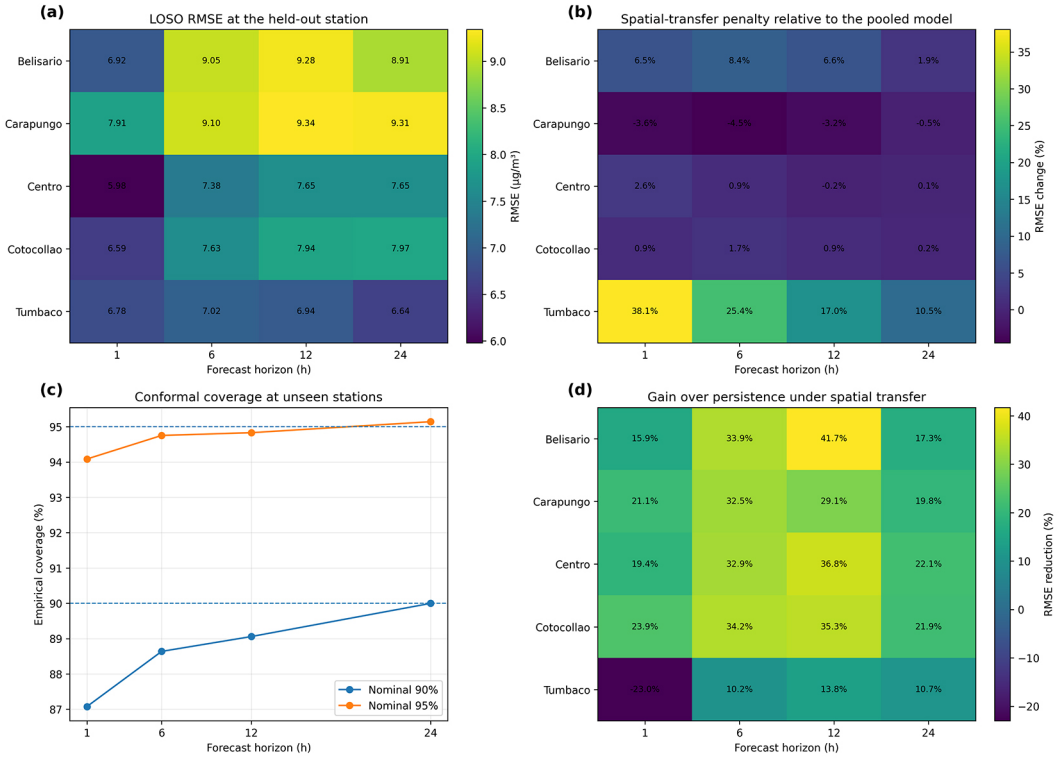

Figure 4. Strict spatial-transfer evaluation. (a) CatBoost LOSO RMSE at each held-out station; (b) RMSE change relative to the corresponding pooled station model; (c) empirical conformal coverage at unseen stations; and (d) RMSE reduction relative to persistence. Negative values in panel (b) indicate that the LOSO model had lower RMSE than the pooled station-specific model

Feature-group ablation

The calendar-only model produced RMSE values between 8.71 and 8.77 μg m^{-3} across horizons. Adding PM2.5 history reduced RMSE by 23.23% at 1 h, 9.42% at 6 h, 7.51% at 12 h, and 7.96% at 24 h. This was the largest incremental improvement in every case (Table 7).

Meteorology reduced RMSE by a further 3.20% at 1 h, with smaller incremental gains of 1.19%, 0.42%, and 0.50% at longer horizons. Adding gaseous pollutants changed test RMSE by only 0.06%, 0.31%, 0.08%, and −0.11% at 1, 6, 12, and 24 h, respectively; the Diebold–Mariano test indicated a significant benefit only at 6 h ($p < 0.001$), with no significant difference at 1, 12, or 24h. Removing current PM2.5 increased RMSE by 5.60% at 1 h and 1.22% at 24 h, whereas the differences at 6 and 12 h were below 0.4%, showing that lagged and rolling measurements retained much of the relevant history beyond the immediate horizon.

Few-shot local calibration

The smoothed hourly-bias correction was the most consistent low-complexity adaptation

Table 7. Feature-group ablation: test RMSE (μg m^{-3})

Feature configuration	1 h RMSE (μg m^{-3})	6 h RMSE (μg m^{-3})	12 h RMSE (μg m^{-3})	24 h RMSE (μg m^{-3})
Calendar only	8.735	8.713	8.765	8.772
+ PM2.5 history	6.705	7.892	8.107	8.073
+ meteorology	6.491	7.798	8.073	8.033
Full model	6.487	7.774	8.067	8.042
Full without current PM2.5	6.850	7.804	8.065	8.139

Note: Feature groups are cumulative except for the final row, which removes current PM2.5 from the full model.

strategy. On the common future evaluation period, 7 local days reduced aggregate zero-shot RMSE by 4.10%, 2.49%, 1.92%, and 0.91% at 1, 6, 12, and 24 h. Improvements increased with 30 days and reached 7.65%, 6.32%, 5.05%, and 2.94% with 90 days (Table 8). Most of the aggregate gain occurred within the first 30 days.

Tumbaco benefited most. With 90 days, its RMSE decreased by 29.99%, 20.35%, 16.75%, and 12.31% across the four horizons. Belisario also improved consistently. Adaptation was not universally beneficial: Centro deteriorated at 6, 12, and 24 h after 90 days, with statistically significant increases in squared-error loss, whereas its 1-h change was small and not significant. These results support validation-gated calibration rather than automatic application of every local correction (Figure 5).

DISCUSSION

Principal findings

This study combined several evaluation layers that are often assessed separately in air-quality forecasting. A single model family, selected without looking at the final test set, provided competitive multi-horizon forecasts, near-nominal pooled uncertainty, meaningful feature attributions, and substantial gains over persistence under strict spatial transfer. At the same time, the framework exposed weaknesses that a conventional random split would conceal: Tumbaco failed at the 1-h LOSO horizon, short-horizon conformal intervals undercovered at unseen stations, high concentrations were systematically underpredicted, and local adaptation occasionally deteriorated performance.

The chronological CatBoost RMSE of 6.48 μg m^{-3} at 1 h increased to approximately 8 μg m^{-3} at longer horizons, while R^2 declined from 0.544 to about 0.30. This decline is expected as forecast information decays. The small difference between 12- and 24-h error likely reflects the strong daily cycle and the fact that each horizon contains a different set of complete target observations. It should not be interpreted as a universal advantage of 24-h forecasting.

Published PM2.5 studies frequently report higher R^2 or lower RMSE, but direct ranking is inappropriate because concentration ranges, cities, instruments, horizons, splits, and inclusion criteria differ (Chadalavada et al., 2025; Zhou et al., 2024). The present estimates were obtained from a two-year future test and a separate unseen-station experiment. These stricter conditions trade optimistic headline accuracy for a more credible assessment of prospective generalization risk.

Table 8. Aggregate RMSE after smoothed hourly local calibration

Horizon (h)	Zero-shot RMSE (μg m^{-3})	7-day RMSE (μg m^{-3})	30-day RMSE (μg m^{-3})	90-day RMSE (μg m^{-3})	90-day reduction (%)
1	6.862	6.580	6.419	6.337	7.65%
6	8.037	7.837	7.615	7.529	6.32%
12	8.262	8.103	7.937	7.845	5.05%
24	8.151	8.077	7.983	7.912	2.94%

Note: All strategies were evaluated on the common period 1 April 2024–31 December 2025.

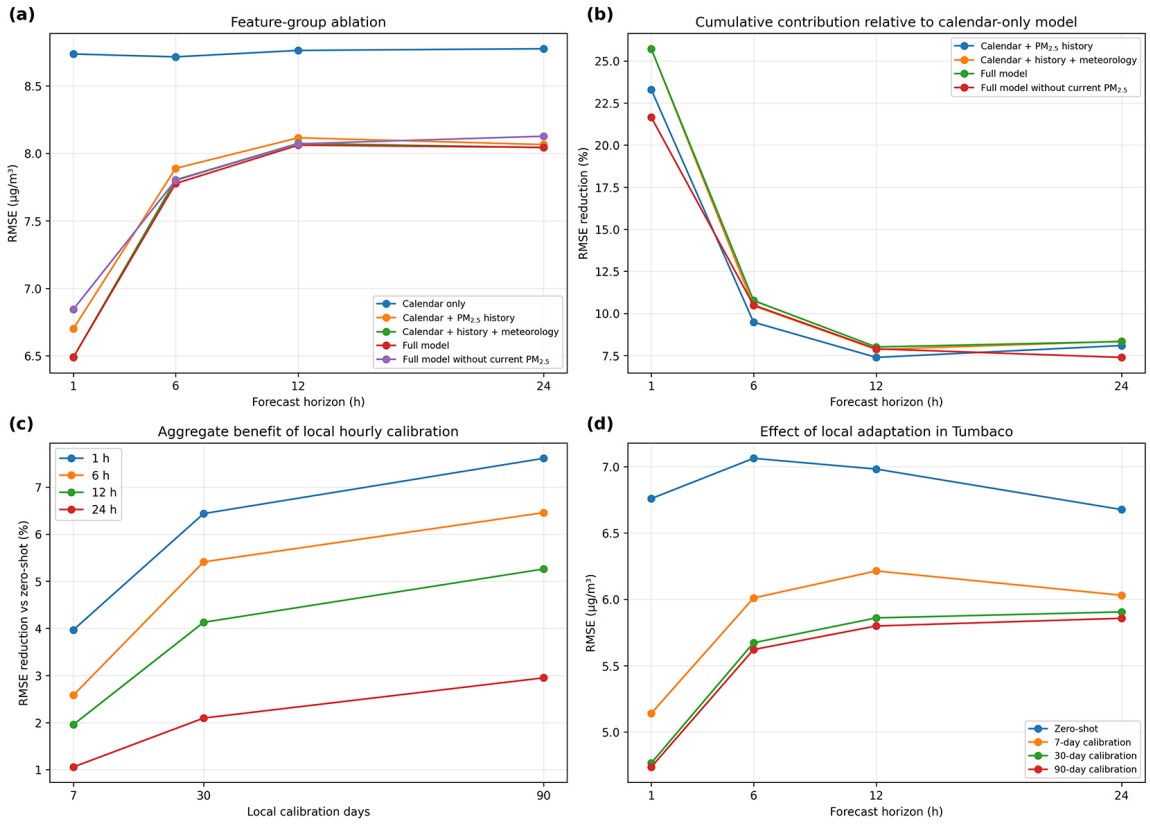

Figure 5. Feature contribution and local adaptation. (a) RMSE for cumulative feature groups; (b) cumulative improvement over the calendar-only model; (c) aggregate benefit of 7, 30, and 90 local calibration days; and (d) Tumbaco RMSE before and after local calibration.

Elevated concentrations expose systematic underprediction

Performance during high concentrations was materially worse than the aggregate scores suggest. At observed PM2.5 above 25 µg m^{-3}, RMSE increased to 11.64–16.18 µg m^{-3}, mean bias was strongly negative, and the percentage of underpredictions increased from 88.2% at 1 h to approximately 98.3% at 24 h. The flattening visible in Figure 3c–d is therefore not merely graphical scatter; it represents systematic regression toward the central concentration range.

This upper-tail limitation is critical when forecasts are considered for alert-related decision support because missing a high episode can be more consequential than making a similar-sized error under ordinary conditions. Squared-error optimization and the dominance of moderate observations can favor central accuracy at the expense of rare extremes. Future methodological development should evaluate regime-weighted loss functions, quantile or expectile objectives, extreme-event classification coupled with regression, asymmetric conformal intervals, and event-based recall. In its present form, the model should be interpreted as a forecasting and diagnostic framework rather than a stand-alone warning system, and tail-specific metrics should accompany aggregate RMSE in any decision-support use.

Temporal memory dominates short-term forecasting

The ablation and SHAP results agreed that recent PM2.5 history was the main information source. At 1 h, adding history reduced calendar-only RMSE by 23.23%, and current PM2.5 was the largest SHAP contributor. This is consistent with the high one-hour autocorrelation and with the persistence of emission and dispersion conditions over short periods. The importance of rolling histories and cyclic hour at 6–24 h indicates that the model combines recent state with recurrent daily structure.

Meteorology added its largest incremental benefit at 1 h (3.20%). Current wind,

temperature, humidity, and radiation can influence dilution, atmospheric stability, photochemistry, and boundary-layer development; however, the global monotonic correlations were modest. The improvement obtained by nonlinear models despite weak pairwise correlations illustrates why correlation screening alone is insufficient. The diminishing meteorological gain at longer horizons may also reflect the use of observed-at-origin rather than future meteorological forecasts. Incorporating lead-time-aligned numerical weather predictions would be a relevant extension for prospective forecasting applications.

Gaseous pollutants provided only small additional test improvements. Their statistically significant benefit at 6 h suggests complementary source or chemical information, whereas differences at 1, 12, and 24 h were not significant in the synchronized analysis. A reduced sensor configuration may therefore remain viable where gaseous analyzers are unavailable or costly. Nevertheless, feature attributions should not be interpreted as causal source apportionment; SHAP describes the trained predictor, not atmospheric mechanisms (Lundberg and Lee, 2017).

Spatial transfer reveals domain shift

The LOSO results demonstrate that temporal generalization and spatial generalization are distinct. CatBoost reduced persistence RMSE in 19 of 20 station-horizon cases, showing that transferable relationships existed across the network. Yet Tumbaco at 1 h was better served by persistence, and its transfer penalty reached 38.62%. Tumbaco also had the lowest mean and 95th-percentile concentrations, which indicates a distribution different from the higher-concentration source stations. A pooled model can partially represent such differences through station indicators; a zero-shot model cannot.

By contrast, Carapungo sometimes achieved lower LOSO RMSE than the pooled station-specific model. This does not imply that withholding data is inherently advantageous; rather, it suggests that the pooled model may have learned station-specific noise or that the source network provided regularization. The finding reinforces the need to report station-level outcomes instead of relying only on an aggregate score.

Transfer-learning research has proposed complex decomposition and neural strategies for data-limited stations (Jairi et al., 2024; Yao et al., 2024). The present work shows that a strong tree ensemble with strict experimental controls can provide a competitive and interpretable baseline. More sophisticated graph, domain-adaptation, or representation-learning methods should be compared against this LOSO benchmark rather than against random splits.

Uncertainty is sensitive to station shift

Pooled split-conformal intervals were close to nominal, but 90% LOSO coverage fell to 87.12% at 1 h. This is not a failure of the conformal algorithm under its assumptions; it reflects the violation of exchangeability between source calibration residuals and target-station errors (Angelopoulos and Bates, 2023; Tibshirani et al., 2019). For decision-support use, this means that a nominal interval can be overconfident precisely where the point model transfers poorly.

The recovery of near-nominal coverage at 24 h may be related to the increased role of daily cyclic structure shared across stations. However, wider intervals at longer horizons reduce precision. Future work should investigate weighted conformal methods, station-similarity weighting, adaptive conformal inference, or calibration with a small target-station sample (Gibbs and Candès, 2021; Tibshirani et al., 2019). Reporting both coverage and width is essential because coverage can be achieved trivially with excessively broad intervals.

Local calibration under limited target-station data

The few-shot experiment provides a practical bridge between zero-shot transfer and fully local retraining. A smoothed hourly correction was deliberately simple: it required few parameters, captured local diurnal bias, and avoided extrapolating high-dimensional seasonal relations from only a few days. Ninety days reduced aggregate 1-h RMSE by 7.65% and Tumbaco RMSE by approximately 30%, with much of the benefit achieved in the first month.

The deterioration observed at Centro is equally informative. Local corrections estimated during one season can become stale or reverse during later conditions. Consequently, local calibration should be regarded as conditional rather than automatically beneficial. In

a monitored-network decision-support setting, candidate corrections should be evaluated on a protected later period and retained only when they improve both aggregate and elevated-concentration diagnostics.

Implications for monitored-network decision support

The results support a tiered decision-support interpretation for urban air-quality networks. A pooled CatBoost model can provide short-term point forecasts with uncertainty intervals at established stations, and the same parameterization can be evaluated at an excluded station when real-time local PM2.5 and predictor histories are available. Local hourly calibration can then be considered after labelled observations accumulate, while empirical interval coverage should be monitored by station to detect distribution shift. These steps do not establish stand-alone alert reliability because high-concentration episodes remained systematically underestimated.

The feature analyses also inform sensor-prioritization questions. Recent PM2.5 history was most important for immediate forecasts, while meteorological measurements provided the clearest short-horizon incremental gain. Gaseous co-pollutants added smaller and horizon-specific improvements. These findings describe predictive information content within the present network and should not be interpreted as a prescription for regulatory monitoring design.

Strengths and limitations

Major strengths include the official multi-year dataset, explicit timestamp harmonization, high PM2.5 completeness, untouched two-year future testing, validation-only model selection, multiple baselines, strict LOSO transfer, uncertainty calibration, statistical comparisons, dependent-data bootstrap intervals, explicit elevated-concentration diagnostics, SHAP, ablation, and a common evaluation period for few-shot adaptation. The correction of model selection to rely exclusively on 2023 is particularly important: it prevents the final test from influencing the reported primary algorithm.

Several limitations should be considered. First, the main panel included five stations because other sites had substantial missing years. The spatial sample is therefore informative for Quito but insufficient to establish general transfer performance across Latin American cities. Second, the model did not include traffic counts, emission inventories, wildfire indicators, satellite aerosol optical depth, mixing-layer height, or future numerical-weather forecasts. Third, the zero-shot experiment still required current and lagged target-station measurements; it was not a virtual-monitoring experiment without a PM2.5 sensor. Fourth, symmetric conformal intervals cannot represent asymmetric upper-tail risk and were not conditioned on pollution regime. Fifth, early-stopping parameters were tuned on one validation year, and no nested temporal cross-validation was conducted because preserving a long independent test period was prioritized. Sixth, the block-bootstrap intervals quantify sampling variability of the fixed fitted model over dependent test periods but do not incorporate training-seed, hyperparameter, or model-selection uncertainty. Finally, the high-concentration strata contained fewer observations and were defined by fixed diagnostic thresholds rather than a formal extreme-value model. Most importantly for practical interpretation, the strong negative bias and 88.2–98.3% underprediction rate above 25 μg m^{-3} mean that the present framework should not be treated as a stand-alone high-pollution warning system.

Finally, the analysis is predictive. Associations of NO_2, SO_2, meteorology, and time variables with SHAP values cannot identify causal emission sources or intervention effects. Causal or chemical-transport analyses would require additional design and data.

CONCLUSIONS

An explainable and uncertainty-aware CatBoost framework produced reproducible hourly fine particulate matter forecasts across five stations in Quito. Under strict chronological separation, test RMSE ranged from 6.48 to 8.06 μg m^{-3} for 1–24 h horizons, with stable moving-block-bootstrap intervals and near-nominal pooled conformal coverage. The upper concentration tail remained the principal weakness: when observed concentrations exceeded 25 μg m^{-3}, RMSE rose to 11.64–16.18 μg m^{-3} and 88.2–98.3% of cases were underpredicted. This limitation prevents the present model from being interpreted as an independently reliable alerting system.

Recent particulate-matter history was the dominant source of forecast skill, while meteorology contributed most at the 1-h horizon. Strict leave-one-station-out transfer reduced RMSE relative to persistence in 19 of 20 station-horizon combinations, but Tumbaco exhibited marked spatial shift, including 1-h underperformance and conformal undercoverage. Smoothed hourly calibration with 90 local days reduced aggregate zero-shot RMSE by 2.94–7.65% and reduced Tumbaco RMSE by 12.31–29.99%, whereas deterioration at Centro demonstrated that local adaptation was station-dependent. Taken together, the results support the framework as a transparent forecasting, uncertainty-quantification, and transfer-diagnostic tool, with explicit caution required for high-concentration events.

Acknowledgments

The author acknowledges the Secretaría de Ambiente del Distrito Metropolitano de Quito for maintaining public access to the historical REMMAQ air-quality and meteorological observations used in this study.

REFERENCES

1. Ali, N., Kausar, A., Vambol, S., Kozub, P., Kozub, S., Hrynzovskyi, A. (2024). Environmental retrieval during COVID-19 lockdown in the megacity Karachi, Pakistan. *Ecological Questions, 35*(3), 1–16. https://doi.org/10.12775/EQ.2024.032
2. Alvarez-Mendoza, C. I., Teodoro, A. C., Torres, N., Vivanco, V. (2019). Assessment of remote sensing data to model PM10 estimation in cities with a low number of air quality stations: A case study in Quito, Ecuador. *Environments, 6*(7), 85. https://doi.org/10.3390/environments6070085
3. Angelopoulos, A. N., Bates, S. (2023). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101
4. Chadalavada, S., Faust, O., Salvi, M., Seoni, S., Raj, N., Raghavendra, U., Gudigar, A., Barua, P. D., Molinari, F., Acharya, U. R. (2025). Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review. *Environmental Modelling & Software, 185*, 106312. https://doi.org/10.1016/j.envsoft.2024.106312
5. Chen, K., Li, G., Li, H., Wang, Y., Wang, W., Liu, Q., Wang, H. (2024). Quantifying uncertainty: Air quality forecasting based on dynamic spatial-temporal denoising diffusion probabilistic model. *Environmental Research, 249*, 118438. https://doi.org/10.1016/j.envres.2024.118438
6. Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
7. Feng, X., Zhang, X., Henne, S., Zhao, Y. B., Liu, J., Chen, T. L., Wang, J. (2024). A hybrid model for enhanced forecasting of PM2.5 spatiotemporal concentrations with high resolution and accuracy. *Environmental Pollution, 355*, 124263. https://doi.org/10.1016/j.envpol.2024.124263
8. Gibbs, I., Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems, 34*, 1660–1672.
9. Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6), 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x
10. Gutiérrez-Avila, I., Arfer, K. B., Carrión, D., Rush, J., Kloog, I., Naeger, A. R., Grutter, M., Páramo-Figueroa, V. H., Riojas-Rodríguez, H., Just, A. C. (2022). Prediction of daily mean and one-hour maximum PM2.5 concentrations and applications in Central Mexico using satellite-based machine-learning models. *Journal of Exposure Science & Environmental Epidemiology, 32*, 917–925. https://doi.org/10.1038/s41370-022-00471-4
11. Jairi, I., Ben-Othman, S., Canivet, L., Zgaya-Biau, H. (2024). Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants. *Environmental Technology & Innovation, 36*, 103793. https://doi.org/10.1016/j.eti.2024.103793
12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.
13. Kreiss, J.-P., Lahiri, S. N. (2012). Bootstrap methods for time series. In *Handbook of Statistics* (Vol. 30, pp. 3–26). Elsevier. https://doi.org/10.1016/B978-0-444-53858-1.00001-6
14. Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
15. Murad, A., Kraemer, F. A., Bach, K., Taylor, G. (2021). Probabilistic deep learning to quantify uncertainty in air quality forecasting. *Sensors, 21*(23), 8009. https://doi.org/10.3390/s21238009
16. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: Unbiased

boosting with categorical features. *Advances in Neural Information Processing Systems, 31*, 6638–6648.

17. Secretaría de Ambiente del Distrito Metropolitano de Quito. (n.d.). *Red Metropolitana de Monitoreo Atmosférico de Quito: Historical air-quality and meteorological data*. Retrieved August 1, 2026, from https://datosambiente.quito.gob.ec/
18. Tibshirani, R. J., Barber, R. F., Candès, E. J., Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems, 32*.
19. Valencia, V. H., Hertel, O., Ketzel, M., Levin, G. (2020). Modeling urban background air pollution in Quito, Ecuador. *Atmospheric Pollution Research, 11*(4), 646–666. https://doi.org/10.1016/j.apr.2019.12.014
20. World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.
21. Yao, B., Ling, G., Liu, F., Ge, M. F. (2024). Multi-source variational mode transfer learning for enhanced PM2.5 concentration forecasting at data-limited monitoring stations. *Expert Systems with Applications, 238*, 121714. https://doi.org/10.1016/j.eswa.2023.121714
22. Zhou, S., Wang, W., Zhu, L., Qiao, Q., Kang, Y. (2024). Deep-learning architecture for PM2.5 concentration prediction: A review. *Environmental Science and Ecotechnology, 21*, 100400. https://doi.org/10.1016/j.ese.2024.100400