Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Stage-Resolved Explainability Workload and Dependence-Aware Performance Assessment in $PM_{2.5}$ Forecasting: A Multi-Criteria Decision Framework

**Agastine A**

agastine@isbr.in

International School of Business and Research

**Badhri Narayanaa P**

International School of Business and Research

**Article**

**Keywords:** $PM_{2.5}$ forecasting, Dependence-aware validation, Explainable machine learning, Computational workload, Persistence benchmark, Pareto optimization

**Posted Date:** November 26th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11428615/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Operational environmental forecasting requires more than accurate predictions: model fitting, inference and post-hoc explanation create distinct computational demands, while serially correlated observations complicate performance comparisons. This study develops a stage-resolved explainability framework that combines forecast error, explanation workload, performance uncertainty and Pareto-based selection. An initial five-model evaluation used 420,768 hourly records from 12 Beijing monitoring stations. Random Forest achieved $R^2 = 0.9418$ and LightGBM $R^2 = 0.9397$; explanation-to-prediction time ratios for 50 observations were 130.9, 38.5 and 11.3 for Random Forest, LightGBM and XGBoost. An independent experiment fitted new Random Forest and LightGBM configurations and evaluated both against one-hour persistence on 16,558 matched test cases. Their root mean squared errors were 26.78 μg/m³, compared with 27.26 μg/m³ for persistence. Synchronized 24-hour moving-block resampling (1,000 replications) and paired heteroskedasticity- and autocorrelation-consistent tests did not demonstrate a statistically detectable difference in squared loss ($p > 0.05$). The experiments are interpreted separately because their model configurations and timing workloads differ. The proposed framework makes explanation frequency an explicit operational decision variable and prevents small forecast improvements from being evaluated independently of computational cost and temporal dependence.


# 1 Introduction

Hourly $PM_{2.5}$ forecasting supports short-horizon air-quality surveillance, exposure communication and environmental decisions. The task is statistically challenging because pollutant concentrations are persistent, meteorological drivers vary over time and observations from neighboring stations share atmospheric influences. A useful forecasting system must therefore be evaluated with chronological data partitioning and with uncertainty procedures that respect dependence. Reviews document the growth of machine-learning forecasting but also identify the continuing need for comparable validation and interpretable modelling (Joharestani et al. 2019; Zhou et al. 2024).

Predictive accuracy is only one component of operational performance. An environmental model may be retrained periodically, queried continuously and explained selectively during elevated-pollution episodes or routine audit. Training, direct inference and explanation consequently impose different computational workloads. Computational-efficiency research emphasizes the importance of distinguishing execution settings, implementation and measurement boundaries (Schwartz et al. 2020; Strubell et al. 2019; Patterson et al. 2022; Georgiou et al. 2022). These considerations become especially important when a small forecasting advantage is obtained at substantially greater explanation cost. Elapsed runtime measures computation under a defined environment; it is not itself an estimate of energy consumption or carbon emissions.

SHAP and permutation feature importance are widely used to examine environmental predictors and to communicate model behavior (Lundberg and Lee 2017; Barredo Arrieta et al. 2020; Houdou et al. 2024).

Applications to particulate matter illustrate the value of understanding persistence, co-pollutants and meteorological covariates (Birinci et al. 2026; Wang et al. 2026; Liu and Hang 2026). Two questions remain operationally important: how much additional computation is needed to explain model outputs, and whether differences in apparent forecasting skill survive an appropriate baseline and dependence-aware comparison. Agreement between attribution techniques offers a consistency check, not evidence of causal identification (Longo et al. 2024; Shao et al. 2026).

The Stage-Resolved Explainable AI (SRE-XAI) framework addresses these linked issues through explicit accounting of fitting, prediction and explanation stages, an explanation-frequency workload function and a finite-candidate Pareto rule. The empirical assessment consists of two clearly distinguished experiments. The first compares five architectures and reports explanation costs for three tree ensembles. The second provides an independent, timestamp-matched persistence comparison with moving-block uncertainty and paired loss tests for fitted Random Forest and LightGBM models. Keeping the experiments separate is necessary because an explanation ratio measured on one fitted implementation cannot be transferred to another. The combined evidence informs model-selection methodology without assuming universal ranking across cities, horizons or hardware.

The contribution is threefold. First, the evaluation boundary treats post-hoc explanation as a measurable computational stage rather than an incidental output of prediction. Second, a formal multi-objective selection rule accounts for predictive error and stage-specific workload without assigning unvalidated utility weights. Third, the empirical demonstration juxtaposes the original explainability benchmark with a separately reproducible, persistence-based and dependence-aware forecasting comparison. The contribution is an evaluation and decision protocol rather than a newly invented forecasting architecture.

Three research questions guide the study. RQ1: How do candidate forecasting architectures differ in out-of-time predictive skill, and how does their performance compare with zero-parameter persistence on matched observations? RQ2: How does post-hoc explanation time compare with direct-prediction time under the same defined workload? RQ3: How do temporal dependence, explanation frequency and competing performance–cost objectives affect defensible model selection?

## 2 Related work and research gap

# 2.1 Environmental machine learning and explanation

Environmental forecasting employs regression, bagged and boosted trees, neural models and hybrid systems. For structured pollutant and meteorological inputs, Random Forest, XGBoost and LightGBM capture nonlinearities and interactions (Breiman 2001; Chen and Guestrin 2016; Ke et al. 2017). Reported rankings are conditional on the target, horizon, predictor availability and validation strategy. At a one-hour horizon, persistence provides a demanding zero-parameter comparator: incremental learning-model value must be established on identical station–timestamp cases rather than inferred from high $R^2$ alone (Joharestani et al. 2019; Zhou et al. 2024).

Explanation methods answer a different question from prediction. SHAP provides additive attribution for a fitted output, whereas permutation feature importance measures the loss of predictive performance after perturbing a feature (Lundberg and Lee 2017; Barredo Arrieta et al. 2020). Air-quality applications emphasize meteorological and pollutant-driver interpretation (Houdou et al. 2024; Birinci et al. 2026; Wang et al. 2026). Their agreement may indicate a stable broad ranking, but correlated covariates, model dependence and the chosen explanation algorithm limit stronger interpretation. Computational demand must also be assessed explicitly rather than inferred from the appearance or complexity of attribution plots (Longo et al. 2024; Shao et al. 2026).

## 2.2 Computationally responsible machine learning

Computationally responsible machine learning broadens evaluation beyond forecast scores to algorithmic resources and deployment behavior (Schwartz et al. 2020; Lannelongue et al. 2021; Patterson et al. 2022). Execution time varies with implementation, workload size, threading and hardware; comparing fitting costs without specifying whether tuning is included can misstate total development effort (Georgiou et al. 2022; Verdecchia et al. 2023). For explainable systems, the additional explanation stage is particularly consequential because explanations may be generated for every prediction or only for selected decisions. Stage-resolved timing therefore complements, but cannot substitute for, direct electricity or emissions measurement.

Evaluation uncertainty also has a structural dimension. Environmental time series exhibit serial correlation and station-level common shocks, so independently resampling individual observations can understate uncertainty in performance differences. A paired resampling design should preserve aligned forecast outcomes; moving contiguous time blocks jointly across stations addresses part of this dependence structure. Such an assessment answers whether observed differences are distinguishable from temporal variation, while the computational audit answers whether a potential improvement warrants its execution burden. These dimensions should remain separate in a multi-criteria decision model.

## 2.3 Research gap and scientific contribution

The literature establishes three substantial but incompletely connected lines of work: accurate environmental prediction, explanation of model behavior and efficient machine learning. The specific gap is the absence, within an otherwise common forecasting workflow, of a transparent decision boundary that jointly considers a credible persistence comparator, dependence-aware error contrasts and the workload incurred when explanations are requested. An accuracy-only ranking overlooks explanation demand. A timing-only ranking overlooks forecast value. A marginal uncertainty interval cannot determine the uncertainty of a paired difference. The proposed framework combines these decisions while preserving their distinct estimands. Tables 1 and 2 position this contribution against representative research rather than claiming that related evaluation ideas have never appeared.

Recent environmental-statistics studies address leakage from clustered observations (Guignard et al. 2024), the influence of preprocessing on learning comparisons (Demir and Sahin 2024) and stochastic

variation in hydrological forecasting (Bogner and Padrón 2026). They support the need for careful experimental boundaries. The present application connects that concern with explainability-stage workload, using a formal Pareto criterion and a separately performed paired persistence analysis. Its generalizable contribution is the protocol; the numerical findings remain specific to the Beijing dataset and the recorded execution conditions.

Table 1
Research gap and contribution mapping

| Research stream | Typical evaluation focus | Unresolved issue | Contribution of this study |
|---|---|---|---|
| Environmental ML | Predictive accuracy and forecasting error | Operational explainability cost is rarely modelled | Chronological model comparison with stage-resolved accounting |
| Environmental XAI | Feature attribution and physical plausibility | Explanation generation is usually treated as cost-free | SHAP and PFI interpreted together with explanation workload |
| Green AI | Training, inference, energy, hardware, carbon | Explanation is rarely isolated as a separate stage | Prediction and explanation runtimes measured separately |
| Model selection | Single-metric ranking or accuracy-efficiency trade-off | Preference can change with explanation frequency | Operational cost scenarios and Pareto-based selection |

Table 2
Representative literature positioning and the stage-resolved research gap

| Study / stream | Application focus | Interpretability / efficiency focus | Stage-resolved XAI workload? | Role in present positioning |
|---|---|---|---|---|
| Joharestani et al. (2019) | $PM_{2.5}$ prediction | Comparative predictive modelling | Not a study focus | Establishes model-comparison context |
| Houdou et al. (2024) | Air-pollution forecasting review | Interpretable ML methods | Not a study focus | Shows growing environmental-XAI adoption |
| Birinci et al. (2026) | Multi-pollutant prediction in Istanbul | SHAP-based interpretation | Not reported as central criterion | Illustrates modern air-quality XAI use |
| Liu and Hang (2026) | UHI-$PM_{2.5}$ management | LightGBM-SHAP mechanisms | Not reported as central criterion | Illustrates policy-oriented interpretable ML |
| Wang et al. (2026) | $PM_{2.5}$ sources and meteorology | Explainable ML attribution | Not reported as central criterion | Illustrates environmental driver interpretation |
| Shao et al. (2026) | Air-pollution IML review | Interpretation scope and robustness | Review emphasis, not stage accounting | Motivates careful robustness language |
| Schwartz et al. (2020); Verdecchia et al. (2023) | Green / efficient AI | Computational efficiency | Explanation stage not isolated | Provides efficiency foundation |
| This study | One-hour-ahead $PM_{2.5}$ demonstration | SHAP + PFI + operational workload | Yes: fit, predict, explain separated | Connects XAI workload with multi-criteria model selection |

# 3 Stage-resolved explainability evaluation framework

# 3.1 Evaluation boundary

For an evaluated model m, the Stage-Resolved Explainable AI assessment associates six evidence dimensions with each prediction system: predictive quality, single-fit computation, direct inference workload, explanation workload, attribution consistency and evaluation robustness. The corresponding vector is defined in Eq. (1). Components are kept separate because a change in units or a policy-specific preference cannot legitimately be concealed inside an arbitrary composite score.

$$SRE(m) = \{Q\square\square\square d,\ Cfit,\ Cpred,\ Cxai,\ Econs,\ R\}$$

1

Here Qpred(m) denotes the predictive-quality metrics ($R^2$, RMSE and MAE); Cfit(m) the final-model fitting time; Cpred(m) direct-prediction workload; Cxai(m) post-hoc explanation workload; Econs(m) evidence of agreement between explanation procedures; and R(m) assessment robustness. MAPE is secondary because small observed concentrations magnify percentage errors. Econs is a consistency diagnostic, not a guarantee of attribution correctness; R includes the experiment-specific uncertainty and ablation procedures rather than a single numerical risk index.

# 3.2 Explanation overhead

For a fixed model and the same cases, define explanation overhead as the ratio of explanation-generation time to direct-prediction time (Eq. (2)). Both numerator and denominator must refer to the same workload and execution context. The ratio is dimensionless and describes relative computational burden; it is not an explanation-quality or energy metric.

$$O\square\square\square\,(m)\;=\;T\square\square\square\,(m)\;/\,T\square\square\square d\,(m)$$

2

An overhead of 100 indicates that generating explanations for a specified workload takes approximately one hundred times the direct-prediction runtime under the measured conditions. Comparisons across studies require matching explanation algorithms, samples, software and hardware. The original 50-case benchmark is retained as a controlled within-experiment comparison rather than extrapolated as a hardware-independent constant.

# 3.3 Operational explanation-frequency function

Explanation requests may be sparse or routine. For N predictions and an explanation fraction f between 0 and 1, (Eq. 3) separates fitting once, repeated inference and explanation on Nf cases. This first-order accounting formulation converts a measured workload into a transparent scenario calculation, not a benchmark of end-to-end deployment throughput.

$$C\square\square\square\square\square(m,\;N,\;f)\;=\;Cfit\,(m)\;+\;N\;\times\;\;tpred\,(m)\;+\;N\;\times\;\;f\;\times\;\;txai\,(m)$$

3

In (Eq. 3), Cfit(m) represents a single final model fit; tpred(m) and txai(m) are observed per-case averages calculated from the common 50-case benchmark. Forecast volume N and explanation frequency f are operational assumptions. The arithmetic projection assumes stable serial per-case cost and excludes hyperparameter tuning, caching, batch optimization and hardware scaling. These boundaries are essential when comparing frequency scenarios.

For production execution, the more general relationship Cdeploy(m,N,f,b,h) = Cfit(m) + Tpred(m,N,b,h) + Txai(m,Nf,b,h) recognizes batch size b and execution environment h. Its workload functions require

separate measurement under deployment conditions. No accelerator speedup, energy benefit or non-linear batch coefficient follows from the present observations.

## 3.4 Multi-criteria model-selection rule

Consider the finite set M of tree-based candidates for which explanation timing was measured. With larger values preferred on all coordinates, define g(m) = [$R^2$(m), −RMSE(m), −Cfit(m), −tpred(m), −txai(m)]. Model a dominates model b if $g_i(a) \geq g_i(b)$ for every i and at least one strict inequality holds. The Pareto set consists of candidates not dominated by another evaluated candidate. $R^2$ and RMSE describe complementary aspects of the same predictive task; the criterion set, sample and environment must therefore be reported explicitly. The decision rule is descriptive, makes no claim of statistical superiority and does not combine the independent-validation model estimates with original explanation timings.

## 3.5 Interpretation of uncertainty and operational risk

Three quantities must be distinguished. Forecast error measures predictive performance for observed concentrations; uncertainty in a model-level error statistic quantifies evaluation variability; and post-hoc explanation workload quantifies additional computation to interpret fitted outputs. Paired block-bootstrap intervals and autocorrelation-consistent tests pertain to differences in forecasting error, not to observation-level prediction intervals. Attribution agreement concerns ranking stability, not causal mechanisms. Keeping these dimensions separate protects the environmental interpretation of the multi-objective decision rule.

# 4 Materials and methods

## 4.1 Data source and forecasting problem

The empirical analysis used the Beijing Multi-Site Air Quality Dataset from the UCI Machine Learning Repository (Chen 2017). The dataset contains hourly observations from 12 monitoring stations between 1 March 2013 and 28 February 2017. Combining the station files produced 420,768 hourly records. The raw $PM_{2.5}$ field contained 412,029 valid observations; after the within-station one-step target shift, 412,017 records retained a valid next-hour target and were available for chronological partitioning.

The forecasting target was $PM_{2.5}$ at time t + 1 using information available at time t. Current-hour $PM_{2.5}$ was retained in the primary feature set because persistence is operationally relevant for one-hour-ahead forecasting. A separate ablation removed current-hour $PM_{2.5}$ to test whether predictive performance persisted through co-pollutants, meteorological conditions, temporal structure, and station identity.

Table 3
Data structure and predictor groups

| Component | Specification |
|---|---|
| Observations | 420,768 hourly records; 12 monitoring stations |
| Study period | March 2013 to February 2017 |
| Forecast target | $PM_{2.5}$ concentration at t + 1 |
| Pollutant predictors | $PM_{2.5}$, $PM_{10}$, $SO_2$, $NO_2$, CO, $O_3$ |
| Meteorological predictors | Temperature, pressure, dew point, rainfall, wind direction, wind speed |
| Temporal representation | Hour, month, day of week, sine and cosine cyclical terms |
| Spatial representation | Monitoring-station identifier |
| Final encoded feature space | 45 predictors |

# 4.2 Preprocessing and temporal leakage prevention

Station files were concatenated after constructing an hourly timestamp from year, month, day, and hour fields. Records were ordered by station and timestamp. The one-step target shift was applied within station so that the final observation from one station could not be linked to the first observation from another. Missing numerical predictors were imputed using medians estimated from the training period only. Wind direction and station identity were treated as categorical variables, producing 45 predictors after categorical encoding and cyclical time representation.

Hour, month, and day of week were represented using sine and cosine terms to preserve periodic proximity at temporal boundaries. The t + 1 target was excluded from every predictor available at t. High-pollution episodes were retained because they are part of the forecasting problem: $PM_{2.5}$ reached 999 μg/m³, with 95th and 99th percentiles of approximately 242 and 370 μg/m³. This decision allows RMSE and MAE to reflect performance under both routine and severe pollution conditions.

# 4.3 Chronological validation and predictive models

A chronological validation design was used instead of random partitioning. Observations before 1 January 2016 formed the training set (n = 292,150), observations from 1 January to 31 December 2016 formed the validation set (n = 103,151), and observations from 1 January to 28 February 2017 formed the unseen test set (n = 16,716). The final test period was not used to guide feature construction or model choice. This design preserves forecasting direction and reduces the risk that neighbouring observations from the same pollution episode occur on both sides of the test boundary. The one-hour persistence predictor is included as an input but a separately fitted persistence-only benchmark was not reported; this limits claims about incremental forecast skill over naïve persistence.

Five architectures were evaluated: Linear Regression, Random Forest, XGBoost, LightGBM, and a multilayer perceptron (MLP). Linear Regression provided a low-complexity baseline; Random Forest

represented bagged trees; XGBoost and LightGBM represented gradient-boosted tree designs; and the MLP provided a neural benchmark. All models used the same encoded predictors and identical unseen test observations. Predictive performance was summarized using $R^2$, RMSE, MAE, and MAPE. Because conventional MAPE can become unstable when observed $PM_{2.5}$ values are small, $R^2$, RMSE, and MAE are treated as the primary model-comparison metrics and MAPE is retained as a secondary descriptive metric rather than a decisive ranking criterion.

# 4.4 Explainability and runtime measurement

Tree SHAP was applied to Random Forest, XGBoost, and LightGBM. Global importance was summarized using mean absolute SHAP values. PFI supplied an independent model-agnostic ranking check by quantifying deterioration in prediction quality after feature permutation. Agreement was assessed through the ordering of dominant predictors rather than numerical equivalence between SHAP and PFI. Cross-method agreement is interpreted as explanation consistency and ranking robustness; it is not interpreted as causal truth or proof of attribution correctness.

Computational accounting distinguished a single final-model fit, direct prediction on 50 held-out cases, and SHAP generation for the same 50 cases. The timing comparisons concern these fixed workloads rather than total development time, which also includes model selection and tuning. Permutation feature importance (PFI) was measured independently on 200 cases, so its absolute runtime is not directly comparable with the 50-case SHAP workload. Explanation-frequency values between 0.01 and 1.00 were used only in arithmetic sensitivity calculations. Without independent execution-state replications, measured runtime differences are reported descriptively.

# 4.5 Predictive uncertainty and robustness tests

Aggregate test-metric uncertainty was characterized using 300 observation-level bootstrap resamples of the held-out predictions. This procedure does not preserve serial correlation or synchronized variation among monitoring stations. Accordingly, the reported 95% ranges are conditional on the observation-level resampling scheme and must not be interpreted as dependence-adjusted confidence intervals or predictive intervals for individual pollution events. Persistence-feature ablation re-estimated LightGBM and XGBoost after removing current-hour $PM_{2.5}$. Attribution consistency was evaluated using cross-model SHAP rankings and permutation-based error deterioration.

Observation-level bootstrap ranges describe the original five-model experiment only. Section 5.7 presents a distinct, independently executed paired validation using synchronized 24-hour moving-block resampling (1,000 replications) and HAC-based tests on the fitted Random Forest and LightGBM models. These results do not retroactively adjust the original experiment’s marginal intervals.

# 4.6 Reproducibility boundary

Comparability was maintained by holding the target definition, chronological partitions, encoded feature space, test period, and XAI workloads constant. SHAP timing used the same fixed 50-case workload for all three tree architectures, while PFI used a separate fixed 200-case workload and was interpreted only

within that procedure. Randomness affecting model estimation and bootstrap resampling was controlled through fixed seeds where supported. Runtime is reported only as measured computational burden; energy and $CO_2$-equivalent outcomes are excluded because direct power telemetry was not available. For full computational replication, the archived experiment record should report the exact processor, RAM, operating system, Python build, library versions, thread settings, and final model hyperparameters; these identifiers should be taken from the original run environment rather than reconstructed retrospectively.

# 5 Results

# 5.1 Predictive skill and bootstrap uncertainty

Reported $R^2$ values are dimensionless; RMSE and MAE are expressed in μg/m³. Runtime is expressed in seconds. Bootstrap bounds describe uncertainty in model-level scores on the test sample, not calibrated prediction intervals for individual observations.

Random Forest achieved the highest out-of-time predictive performance, with $R^2$ = 0.9418, RMSE = 27.09, and MAE = 12.66. LightGBM followed closely at $R^2$ = 0.9397, RMSE = 27.59, and MAE = 12.70. The MLP reached $R^2$ = 0.9400 but produced a higher MAE than LightGBM and required 6.55 s for the single final fit. XGBoost and Linear Regression recorded $R^2$ values of 0.9325 and 0.9302. MAPE values were comparatively high (41.08%-57.95%), which is consistent with the known sensitivity of percentage error when observed concentrations approach zero; accordingly, the substantive ranking is based primarily on $R^2$, RMSE, and MAE rather than MAPE alone.

Table 4
Out-of-time predictive performance and single-fit model-fitting runtime

| Model | $R^2$ | RMSE | MAE | MAPE (%) | Single-fit runtime (s) |
|---|---|---|---|---|---|
| Random Forest | 0.9418 | 27.09 | 12.66 | 41.08 | 12.64 |
| MLP | 0.9400 | 27.51 | 13.89 | 49.20 | 6.55 |
| LightGBM | 0.9397 | 27.59 | 12.70 | 41.47 | 2.54 |
| XGBoost | 0.9325 | 29.19 | 13.28 | 42.99 | 2.12 |
| Linear Regression | 0.9302 | 29.67 | 16.34 | 57.95 | 0.30 |

LightGBM attained 99.78% of the Random Forest $R^2$ and required approximately one-fifth of its single-fit runtime. The reported observation-level bootstrap 95% intervals for $R^2$ were 0.9353–0.9477 for Random Forest, 0.9336–0.9465 for LightGBM, 0.9328–0.9466 for the multilayer perceptron, and 0.9255–0.9381 for XGBoost. These marginal intervals overlap, although overlap alone cannot establish equivalence or the uncertainty of paired between-model differences.

# 5.2 Cross-method consistency of feature attribution

SHAP identified current-hour $PM_{2.5}$ as the dominant predictor for Random Forest, XGBoost, and LightGBM. The mean absolute SHAP value for current $PM_{2.5}$ was approximately 47.94 for Random Forest, 51.92 for XGBoost, and 49.94 for LightGBM. $PM_{10}$, dew point, wind speed, CO, temperature, $NO_2$, and cyclical temporal features formed a second tier of influence. The exact ordering changed across architectures, but the broad signal structure was stable: short-horizon persistence dominated, co-pollutants added substantial information, and meteorological and temporal variables contributed beyond pollutant history.

PFI supported the same broad ranking interpretation. Permuting current $PM_{2.5}$ produced the largest increase in error, with MAE deterioration of approximately 61.19 for Random Forest, 58.66 for XGBoost, and 61.08 for LightGBM. $PM_{10}$ was the strongest secondary variable for Random Forest, with an MAE deterioration of about 10.31. Convergence between SHAP and PFI reduces dependence on one explanation procedure and supports robustness of the dominant-feature ranking. It does not establish causal correctness of the explanations.

# 5.3 Stage-resolved explanation workload

The explanation workload produced a much sharper separation than predictive performance. For 50 standardized test cases, Random Forest required a median 0.01390 s for direct prediction and 1.8190 s for SHAP, corresponding to an explanation-overhead ratio of 130.9. LightGBM required 0.00103 s for prediction and 0.0396 s for SHAP, giving a ratio of 38.5. XGBoost required 0.00216 s for prediction and 0.0245 s for SHAP, giving the smallest proportional burden at 11.3. Similar predictive performance can therefore coexist with materially different explanation workloads.

Table 5
Stage-resolved computational cost of post-hoc explanation

| Model | Prediction time, 50 cases (s) | SHAP time, 50 cases (s) | O□□$_i$ | PFI time, 200 cases (s) |
|---|---|---|---|---|
| Random Forest | 0.01390 | 1.8190 | 130.9× | 0.627 |
| LightGBM | 0.00103 | 0.0396 | 38.5× | 0.079 |
| XGBoost | 0.00216 | 0.0245 | 11.3× | 0.143 |

PFI showed the same broad architecture-level efficiency pattern. LightGBM completed the 200-case permutation workload in 0.079 s, XGBoost in 0.143 s, and Random Forest in 0.627 s. The agreement across two explanation procedures indicates that the computational separation is not restricted to a single SHAP implementation.

# 5.4 Persistence-feature ablation

The ablation experiment tested whether the lower-cost boosting alternatives depended almost entirely on current-hour $PM_{2.5}$. Removing that feature reduced predictive performance but did not collapse the forecasting task. LightGBM retained $R^2 = 0.8947$, RMSE = 36.45, and MAE = 17.67, while XGBoost retained $R^2 = 0.8976$, RMSE = 35.94, and MAE = 17.71. Co-pollutants, meteorological conditions, station identity, and temporal structure therefore retained substantial predictive information after the strongest persistence feature was removed.

Table 6
Robustness evidence and stage-resolved model position

| Criterion | Random Forest | LightGBM | XGBoost |
|---|---|---|---|
| Primary test $R^2$ | 0.9418 | 0.9397 | 0.9325 |
| 95% bootstrap $R^2$ interval | 0.9353–0.9477 | 0.9336–0.9465 | 0.9255–0.9381 |
| Single-fit runtime (s) | 12.64 | 2.54 | 2.12 |
| SHAP overhead | 130.9× | 38.5× | 11.3× |
| $R^2$ without current $PM_{2.5}$ | Not part of ablation | 0.8947 | 0.8976 |

# 5.5 Explanation-frequency sensitivity

The explanation-frequency function was applied to a standardized workload of 1,000 predictions using the measured per-case prediction and SHAP times. Training cost is included once, and SHAP cost is included only for the proportion f of predictions requiring explanation. The scenarios are arithmetic projections from observed runtimes and provide an operational interpretation of the stage-resolved framework.

Table 7
Projected computation for 1,000 predictions at alternative explanation frequencies

| Model | f = 0.01 | f = 0.10 | f = 0.50 | f = 1.00 |
|---|---|---|---|---|
| Random Forest | 13.2818 s | 16.5560 s | 31.1080 s | 49.2980 s |
| LightGBM | 2.5685 s | 2.6398 s | 2.9566 s | 3.3526 s |
| XGBoost | 2.1681 s | 2.2122 s | 2.4082 s | 2.6532 s |

The sensitivity analysis makes the operational consequence of explanation frequency visible. Random Forest remains the accuracy leader, but its cumulative measured-computation requirement rises rapidly as explanation becomes more frequent. XGBoost gives the lowest projected cumulative runtime in all four scenarios, while LightGBM preserves a much smaller predictive gap to Random Forest. The ordering on runtime does not reverse across the tested values of f because Random Forest is slower both to fit and to explain; the substantive sensitivity lies in how much decision weight an application places on a small predictive gain relative to recurrent explanation workload.

# 5.6 Pareto model selection

Using predictive performance, single-fit model-fitting runtime, SHAP overhead, and PFI runtime as competing criteria, none of the three tree-based models is uniformly superior across all dimensions. Random Forest remains non-dominated because it provides the lowest predictive error and highest $R^2$, despite being the most expensive to fit and explain. XGBoost remains non-dominated because it minimizes proportional SHAP burden and single-fit runtime, despite a larger predictive error. LightGBM remains non-dominated because it combines near-leading accuracy with the lowest PFI runtime and substantially lower fitting and SHAP workloads than Random Forest. Figure 4 visualizes this trade-off directly: movement toward lower predictive error comes at sharply different explanation cost, and bubble size makes the fitting-runtime component visible. The Pareto set therefore preserves the substantive trade-offs rather than hiding them inside an arbitrary weighted score. This non-dominance finding is conditional on the chosen metrics and the tested set of candidate models; it does not establish global optimality.

# 5.7 Independent persistence benchmark and dependence-aware paired validation

An independent validation analysis used the 12 station-level air-quality CSV files to test whether the accuracy advantage of fitted learning architectures persists relative to a zero-parameter one-hour persistence forecast, $\hat{y}(t + 1) = y(t)$. This analysis uses the same chronological split boundaries and 45 encoded predictors as the original dataset specification, but fits independently configured Random Forest and LightGBM models. It is a separate experiment from the five-model comparison and explanation-runtime results in Sections 5.1–5.6.

The independent holdout contains 16,716 cases with valid next-hour outcomes. Persistence is defined only where current-hour $PM_{2.5}$ is also observed. All three models are therefore evaluated on 16,558 identical station–timestamp test cases. Each numerical comparison below uses that shared sample, avoiding changes of denominator across models.

Table 8
Matched-case one-hour $PM_{2.5}$ forecast accuracy; RMSE and MAE in μg/m³; 24-hour synchronized time-block bootstrap, 1,000 resamples The two fitted models yielded modestly lower matched-case RMSE point estimates than persistence. The 95% intervals describe variation of model errors under the specified synchronized resampling procedure. Overlap between marginal ranges is not a hypothesis test. Paired comparisons instead resample timestamp-aligned losses and maintain all stations within the same sampled time windows.

| Forecast method | Matched n | $R^2$ | RMSE | MAE | 95% block-bootstrap RMSE range |
|---|---|---|---|---|---|
| Persistence | 16,558 | 0.9414 | 27.26 | 12.58 | 21.00 to 32.93 |
| Random Forest | 16,558 | 0.9435 | 26.78 | 12.39 | 20.06 to 33.18 |
| LightGBM | 16,558 | 0.9434 | 26.78 | 12.44 | 19.93 to 33.39 |

Table 9
Dependence-aware paired comparisons on matched observations; RMSE intervals in μg/m³, squared-loss differences in (μg/m³)²

| Paired comparison | 95% RMSE difference interval | Mean squared-loss difference | HAC p value |
|---|---|---|---|
| Random Forest – persistence | -2.00 to 1.38 | -26.05 | 0.595 |
| LightGBM – persistence | -2.19 to 1.66 | -25.91 | 0.643 |
| LightGBM – Random Forest | -0.42 to 0.44 | 0.14 | 0.990 |

The synchronized 24-hour moving-block bootstrap uses 1,000 replications and retains cross-station alignment within each sampled time interval. Paired differences in squared loss are also tested with heteroskedasticity- and autocorrelation-consistent standard errors using 24 hourly lags. Every paired RMSE interval includes zero and each HAC p value exceeds 0.05. The comparisons do not demonstrate a statistically established reduction in error under the evaluated resampling and test settings. They do not establish equality either. The results are specific to the selected test period, block length, model configurations and HAC specification.

The independently fitted LightGBM and Random Forest models used explicitly defined estimators and fixed random seeds. The recorded single-fit times were approximately 2.23 s and 7.05 s, respectively, under the execution environment used for this run. Those measurements are neither matched SHAP timings nor hardware-normalized measures of energy. In particular, the original manuscript's explanation overhead ratios must not be applied to the new configurations.

Table 10
Measured runtimes for the independently fitted models; training and prediction workloads differ and are not directly comparable to the original 50-case SHAP benchmark

| Fresh model | Single final-model fit (s) | Prediction, full holdout (s) |
|---|---|---|
| LightGBM | 2.234 | 0.032 |
| Random Forest | 7.050 | 0.025 |

A strong persistence benchmark reflects the temporal structure of one-hour pollution. The principal implication is that interpretation workload should be justified by demonstrated forecast benefit and a stated operational need, particularly when the competing forecasts are statistically difficult to separate. The preceding tests strengthen the forecast-comparison component of the SRE-XAI evaluation boundary but do not establish cross-city transferability, measured energy savings or newly benchmarked explanation performance.

# 6 Discussion

# 6.1 Explanation as an operational forecasting stage

The new paired persistence comparison in Section 5.7 adds an important qualification to operational model selection: short-horizon pollutant persistence can achieve accuracy close to that of fitted architectures, and dependence-aware inference may not establish their incremental advantage. Separately, the original explanation benchmark shows that interpretable reporting has measurable computational cost. A model should therefore earn its operating complexity through supported predictive gains and a clear explanation requirement. The original and independent experiments have different model configurations and must not be combined into a single empirical timing frontier.

# 6.2 Why model preferences are conditional

Predictive fit, explanation runtime and implementation-dependent throughput describe competing dimensions rather than a single notion of optimality. Inferences about a preferred model must be conditional on explicit operational priorities, a fixed workload and a defined candidate set. A change in batch size, hardware, explanation method or tolerance for forecasting error can alter that decision boundary. Formal Pareto efficiency does not by itself establish statistical superiority or transferable deployment performance.

The implication is that explainable-model selection should be conditional on use. A monitoring system that explains only exceptional events can tolerate a different architecture than a governance system that produces an explanation for every prediction. The SRE-XAI formulation makes that distinction explicit through f, the explanation frequency. This operational variable is absent from accuracy-only model comparison even though it can materially change cumulative workload.

# 6.3 Environmental interpretation of feature importance

Across the three tree architectures, current-hour $PM_{2.5}$ remained the dominant predictor, consistent with strong short-horizon persistence. $PM_{10}$, dew point, wind speed, CO, temperature, $NO_2$, and cyclical temporal features formed a second tier. The ablation experiment is important because it tests whether the strong results are merely a persistence artifact: removing current $PM_{2.5}$ reduced accuracy, yet both boosting models retained $R^2$ near 0.90. This supports the predictive relevance of co-pollutants, meteorological conditions, station effects, and temporal structure. The SHAP-PFI agreement strengthens confidence in the stability of the broad ranking, while remaining an associational explanation rather than causal identification.

# 6.4 Methodological contribution

The main contribution is a reusable evaluation boundary rather than a new forecasting algorithm. SRE-XAI can be transferred to other environmental tasks whenever the predictive pipeline includes post-hoc explanation. The minimum design requires a fixed predictive workload, fixed explanation workload, common evaluation data, stage-specific timing, and explicit rules for how predictive quality and explanation burden enter model selection. The framework is modular: tuning cost, direct energy

measurement, memory use, additional explanation-robustness diagnostics, or alternative XAI procedures can be incorporated when those measurements are available. Transferability of the framework should not be confused with numerical generalizability of the Beijing runtime results.

## 6.5 External validity and transportability

The empirical evidence comes from one metropolitan dataset, one one-hour forecast horizon, and one historical period. The study therefore supports a framework-level claim and a task-specific empirical demonstration, not universal architecture ranking. A strong external-validation extension should repeat the same timing protocol, feature-availability rule, explanation workload, and Pareto criteria on additional cities or environmental forecasting tasks. Replication is especially important because model complexity, pollutant dynamics, feature distributions, software implementations, and hardware can alter both predictive and explanation runtimes. This bounded claim is intentional: the contribution is the stage-resolved decision protocol, while cross-city invariance remains an empirical question.

## 6.6 Implications for environmental decision support

Within the tested Beijing workload, Random Forest is defensible for accuracy-dominant applications in which explanations are infrequent and marginal predictive gain is valued highly. XGBoost is attractive for explanation-intensive applications in which proportional SHAP burden and cumulative runtime are primary constraints. LightGBM provides the strongest accuracy-computation compromise in this experiment because it approaches Random Forest accuracy while sharply reducing fitting and explanation workload. These are conditional decision profiles, not universal model recommendations; they should be recalibrated whenever the target, horizon, city, hardware, model implementation, XAI method, or explanation frequency changes.

## 6.7 Limitations and reproducibility

The empirical benchmark represents one metropolitan region, one historical period, and one forecast horizon. This is the principal external-validity limitation. The framework is transferable, but the numerical runtimes, Pareto positions, and architecture ordering should not be generalized beyond the tested workload without replication. Runtime depends on processor architecture, software implementation, thread scheduling, memory behaviour, final hyperparameters, and sample size. SHAP and PFI were evaluated, while other explanation methods can scale differently. Hyperparameter-search cost was not included; reported Cfit values are single-fit final-model runtimes rather than total model-development cost.

The study establishes computational burden, not environmental footprint. Direct energy assessment would require matched power measurement for fitting, prediction, and explanation on controlled hardware. Carbon assessment would require documented electricity intensity in addition to measured energy. These quantities are intentionally excluded from the present empirical claims so that runtime is not misrepresented as electricity consumption or $CO_2$-equivalent emissions.

The current analysis uses one chronological train-validation-test design. Internal validity is strengthened by within-station target shifting, training-period-only imputation, a common encoded feature space, an identical unseen test period, bootstrap uncertainty, persistence-feature ablation, cross-method ranking checks, and explanation-frequency sensitivity analysis. External validation should repeat the same SRE-XAI protocol across additional cities, rolling-origin windows, forecast horizons, and hardware environments. A future multi-environment study should also archive full hardware/software identifiers and include tuning cost, memory demand, and direct energy measurement so that computational and environmental claims can be evaluated separately.

# 7 Conclusions

This study examined whether the model with the highest one-hour-ahead $PM_{2.5}$ forecasting accuracy is also the operationally preferable explainable model once post-hoc explanation is treated as an explicit computational stage. Five predictive architectures were evaluated on 420,768 hourly observations from 12 Beijing monitoring stations using chronological out-of-time validation. Random Forest achieved the highest test $R^2$ (0.9418), while LightGBM reached $R^2 = 0.9397$ with a substantially lower single-fit runtime. The separation widened after explanation was introduced: SHAP-to-prediction overhead ratios were 130.9 for Random Forest, 38.5 for LightGBM, and 11.3 for XGBoost, and PFI showed the same broad architecture-level efficiency pattern.

The central contribution is a stage-resolved decision framework distinguishing model fitting, prediction and explanation workloads. The independent validation provides a zero-parameter persistence comparison on common test cases and uses synchronized 24-hour block resampling and paired HAC tests. Although the fitted Random Forest and LightGBM models had slightly lower matched-case point-estimate forecast errors than persistence, the statistical comparisons did not establish robust superiority. Original explanation-timing ratios are retained as a distinct experiment, not transferred to the new model configurations. These findings support conditional multi-criteria selection rather than universal model rankings.

# Declarations


### Funding

The authors declare that no funds, grants, or other financial support were received for the conduct of this research or the preparation of the manuscript.


### Competing Interests

The authors declare that they have no relevant financial or non-financial interests to disclose.

### Author Contributions

**Agastine A:** Conceptualization, methodology, data acquisition, data curation, formal analysis, investigation, computational implementation, validation, visualization, writing of the original draft, and manuscript preparation.

**Badri Narayanaa P:** Research supervision, methodological guidance, critical review, and editing of the manuscript.

Both authors reviewed and approved the final manuscript and agree to be accountable for the integrity and accuracy of the work.

### Data Availability

The study uses the publicly available Beijing Multi-Site Air Quality Dataset from the UCI Machine Learning Repository. The dataset contains hourly air-quality and meteorological observations collected from 12 monitoring stations in Beijing, China, between March 2013 and February 2017. The source dataset is publicly accessible at:

https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data

Dataset DOI: https://doi.org/10.24432/C5RK5G

### Ethics Approval

Ethics approval was not required because the study involved secondary analysis of publicly available environmental monitoring data and did not involve human participants or animals.

### Consent to Participate

Not applicable.

### Consent for Publication

Not applicable.

## References


1. Ahmad, M., Rappenglück, B., Osibanjo, O. O. & Retama, A. A machine learning approach to investigate the build-up of surface ozone in Mexico City. *J. Clean. Prod.* **379**, 134638. https://doi.org/10.1016/j.jclepro.2022.134638 (2022).
2. Alvarez, F. & Smith, M. Using Shapley values to assess the impact of temporary traffic restrictions on $NO_2$ levels in Madrid urban area. *Int. J. Environ. Sci. Technol.* **18**, 3343–3356. https://doi.org/10.1007/s13762-020-03091-3 (2021).
3. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion*. **58**, 82–115.

https://doi.org/10.1016/j.inffus.2019.12.012 (2020).

4. Birinci, E., Ekmekcioğlu, Ö., Ozdemir, H. & Deniz, A. Interpretable machine learning framework for air quality prediction in Istanbul using Shapley additive explanations (SHAP). *Stoch. Env. Res. Risk Assess.* **40**, 37. https://doi.org/10.1007/s00477-026-03168-4 (2026).
5. Bogner, K. & Padrón, R. S. Improving sub-seasonal hydrological forecasts utilizing the randomness in Deep Learning models. *Stoch. Environ. Res. Risk Assess.* **40**, 12. https://doi.org/10.1007/s00477-025-03138-2 (2026).
6. Bolón-Canedo, V., Morán-Fernández, L., Cancela, B. & Alonso-Betanzos, A. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing* **599**, 128096. https://doi.org/10.1016/j.neucom.2024.128096 (2024).
7. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
8. Chen, S. Beijing Multi-Site Air Quality [Dataset]. UCI Machine Learning Repository. (2017). https://doi.org/10.24432/C5RK5G
9. Chen, T., Guestrin, C. & Mining XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data 785–794. (2016). https://doi.org/10.1145/2939672.2939785
10. de Vries, A. The growing energy footprint of artificial intelligence. *Joule* **7**, 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004 (2023).
11. Demir, S. & Sahin, E. K. The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: a study on undrained shear strength prediction. *Stoch. Environ. Res. Risk Assess.* **38**, 3273–3290. https://doi.org/10.1007/s00477-024-02745-9 (2024).
12. Georgiou, S., Kechagia, M., Sharma, T., Sarro, F. & Zou, Y. Green AI: Do deep learning frameworks have different costs? Proceedings of the 44th International Conference on Software Engineering, 1082–1094. (2022). https://doi.org/10.1145/3510003.3510221
13. Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surveys*. **51**, 93. https://doi.org/10.1145/3236009 (2018).
14. Guignard, F., Ginsbourger, D., Levy Häner, L. & Herrera, J. M. Some combinatorics of data leakage induced by clusters. *Stoch. Environ. Res. Risk Assess.* **38**, 2815–2828. https://doi.org/10.1007/s00477-024-02715-1 (2024).
15. Houdou, A. et al. Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol Air Qual. Res.* **24**, 230151. https://doi.org/10.4209/aaqr.230151 (2024).
16. Joharestani, M. Z., Cao, C., Ni, X., Bashir, B. & Talebiesfandarani, S. $PM_{2.5}$ prediction based on Random Forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* **10**, 373. https://doi.org/10.3390/atmos10070373 (2019).

17. Ke, G. et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 3146–3154 (2017).
18. Lannelongue, L., Grealey, J. & Inouye, M. Green Algorithms: Quantifying the carbon footprint of computation. *Adv. Sci.* **8**, 2100707. https://doi.org/10.1002/advs.202100707 (2021).
19. Lannelongue, L. & Inouye, M. Carbon footprint estimation for computational research. *Nat. Reviews Methods Primers*. **3**, 9. https://doi.org/10.1038/s43586-023-00202-5 (2023).
20. Liu, Q. & Hang, T. Seasonal synergistic management of urban heat island effect and $PM_{2.5}$ pollution: Insights from interpretable LightGBM-SHAP machine learning model. *Environ. Impact Assess. Rev.* **116**, 108129. https://doi.org/10.1016/j.eiar.2025.108129 (2026).
21. Longo, L. et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inform. Fusion*. **106**, 102301. https://doi.org/10.1016/j.inffus.2024.102301 (2024).
22. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30**, 4765–4774 (2017).
23. Patterson, D. et al. The carbon footprint of machine learning training will plateau, then shrink. *Computer* **55**, 18–28. https://doi.org/10.1109/MC.2022.3148714 (2022).
24. Peykani, P., Emrouznejad, A., Ghanidel, S., Javadi-Sisi, I. & Mirjalili, S. Green Artificial Intelligence: A comprehensive review of metrics, tools, challenges, trends, and future prospects. *Arch. Comput. Methods Eng.* https://doi.org/10.1007/s11831-026-10546-2 (2026).
25. Santos, S., Ottoni, A. L. C., Borgo, R., Ferreira, D. & Nepomuceno, E. A systematic review of Green Machine Learning: Practices and challenges for sustainability. *Artif. Intell. Rev.* **59**, 132. https://doi.org/10.1007/s10462-026-11515-8 (2026).
26. Schwartz, R., Dodge, J., Smith, N. A. & Etzioni, O. Green AI. *Commun. ACM*. **63**, 54–63. https://doi.org/10.1145/3381831 (2020).
27. Shao, Y. et al. Interpretable machine learning for ambient air pollution modeling: A systemic review. *Atmos. Res.* **336**, 108869. https://doi.org/10.1016/j.atmosres.2026.108869 (2026).
28. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. (2019). https://doi.org/10.18653/v1/P19-1355
29. Verdecchia, R., Sallou, J. & Cruz, L. A systematic review of Green AI. *WIREs Data Min. Knowl. Discov.* **13**, e1507. https://doi.org/10.1002/widm.2022.1507 (2023).
30. Wang, X. et al. Explainable machine learning reveals the effect of sources and meteorology on $PM_{2.5}$ pollution in a typical city of Southern Sichuan, China. *Atmos. Environ.* **372**, 121881. https://doi.org/10.1016/j.atmosenv.2026.121881 (2026).
31. Zhou, S., Wang, W., Zhu, L., Qiao, Q. & Kang, Y. Deep-learning architecture for $PM_{2.5}$ concentration prediction: A review. *Environ. Sci. Ecotechnology*. **21**, 100400. https://doi.org/10.1016/j.ese.2024.100400 (2024).

# Figures

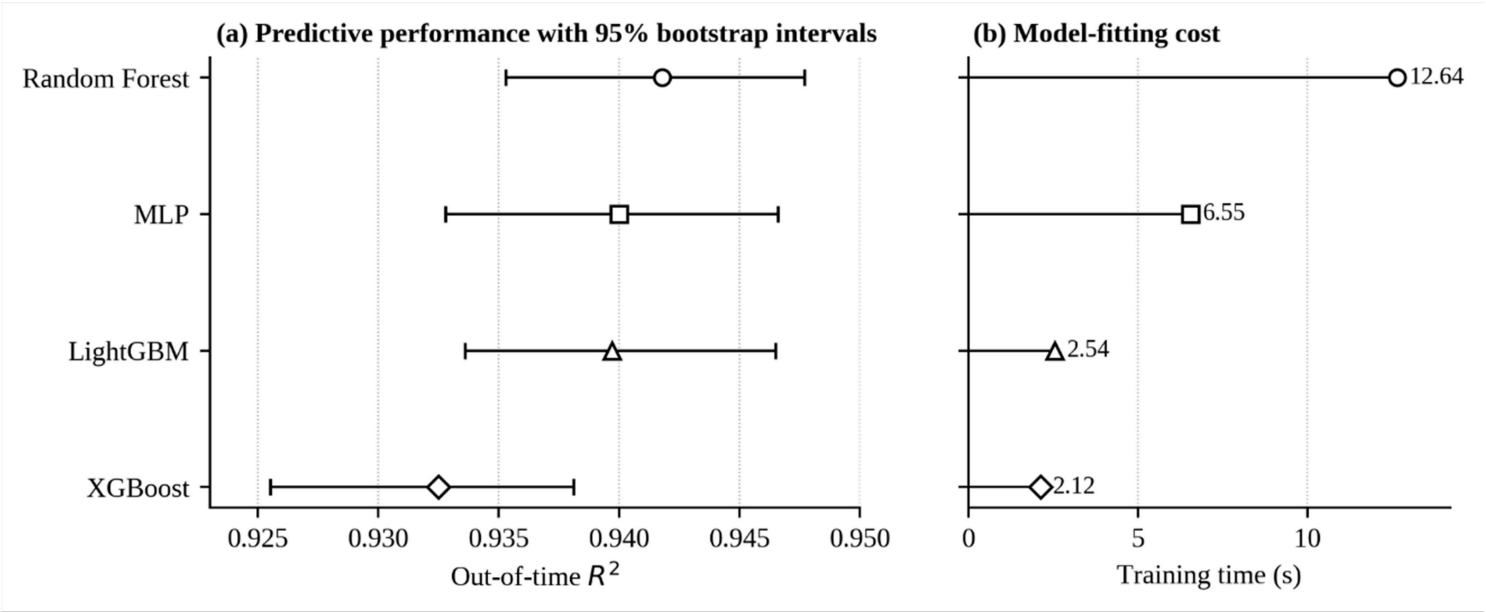


**Figure 1**

Predictive fit and fitting time across candidate models. Left: out-of-time test R² with reported 95% observation-level bootstrap intervals (not adjusted for temporal or cross-station dependence). Right: elapsed time in seconds for fitting a single final model; higher R² and lower runtime are preferred

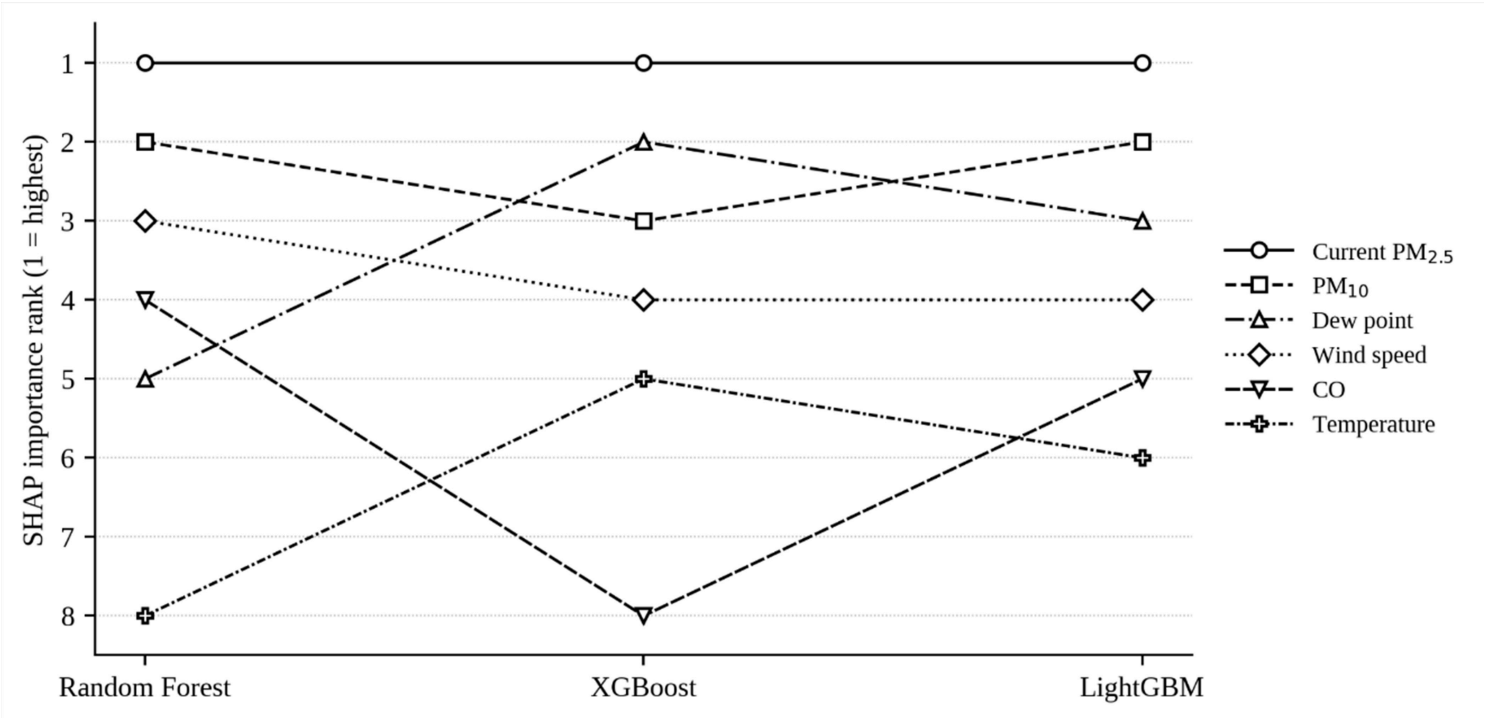


**Figure 2**

SHAP importance-rank comparison across Random Forest, XGBoost and LightGBM. The ordinate gives the within-model rank for six influential predictors, with rank 1 indicating the largest mean absolute attribution. Connecting lines show changes in ordering, not causal effects

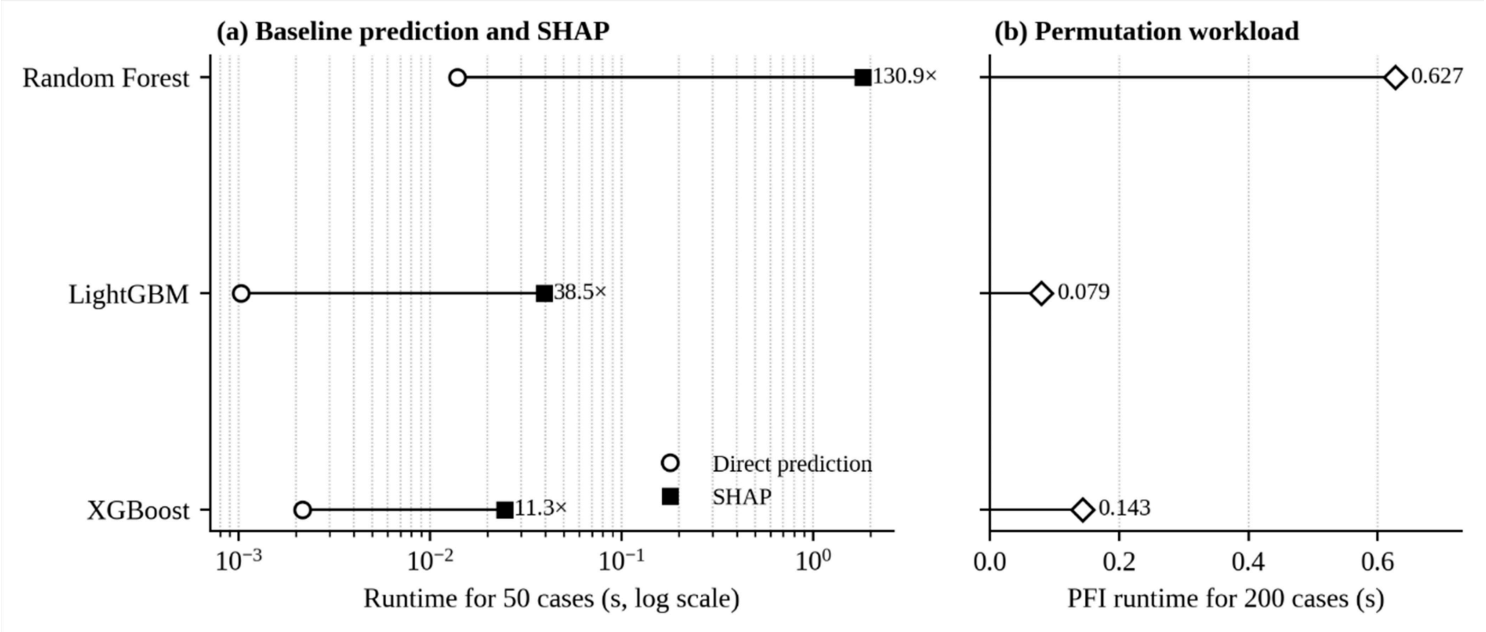


**Figure 3**

Prediction and post-hoc explanation workloads for three tree-based models. Left: elapsed prediction and SHAP times for the same 50-case workload, displayed on a logarithmic scale. Right: permutation feature importance (PFI) time in seconds for 200 cases; these workload sizes must not be directly equated

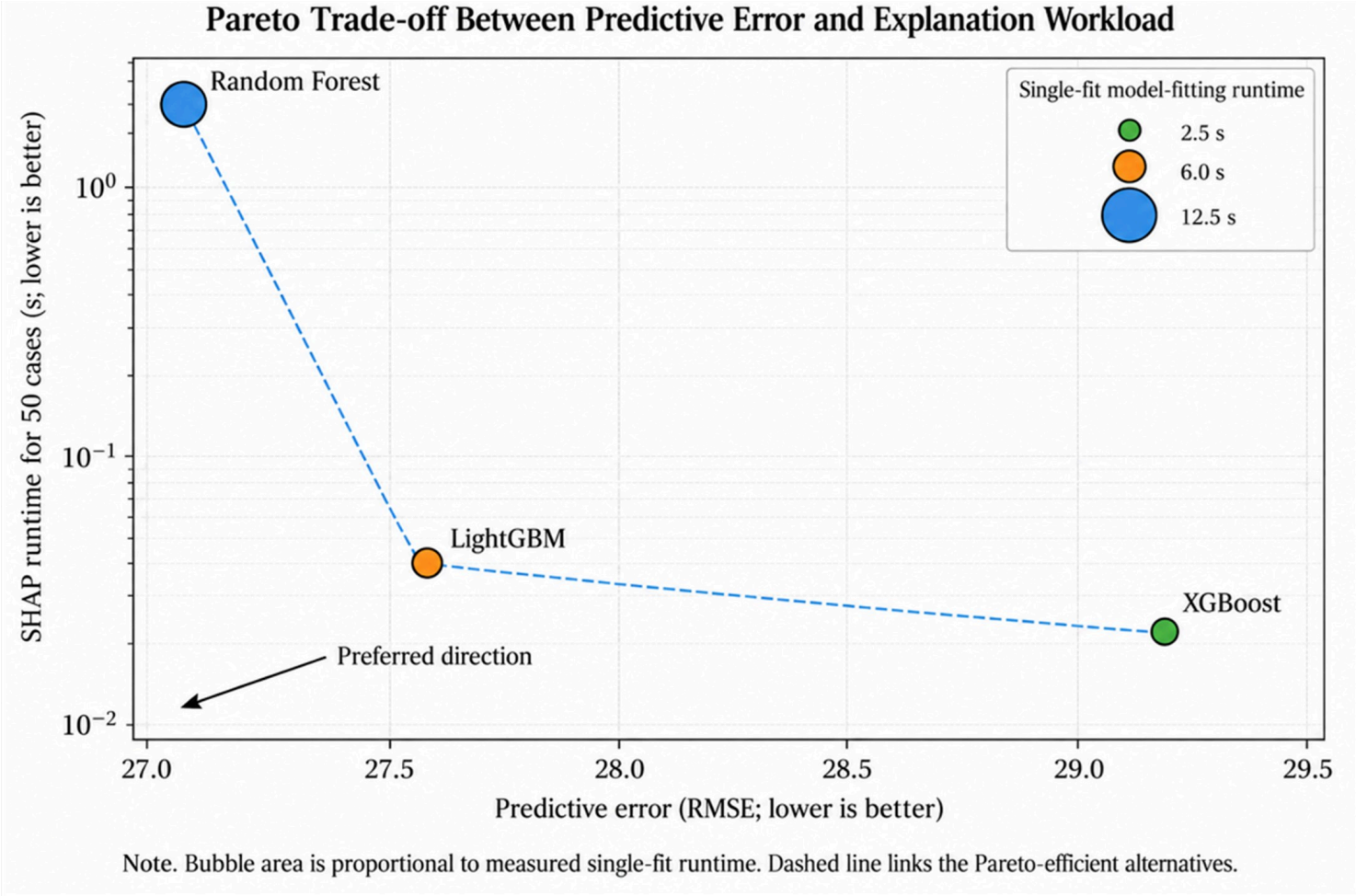


**Figure 4**

Pareto view of the forecasting-error and explanation-workload trade-off for three tree-based models. Horizontal axis: test RMSE in µg/m³ (lower is better). Vertical axis: SHAP runtime in seconds for 50 cases (logarithmic scale; lower is better). Bubble area represents single-fit model-fitting runtime; the displayed candidate set and criteria define the comparison

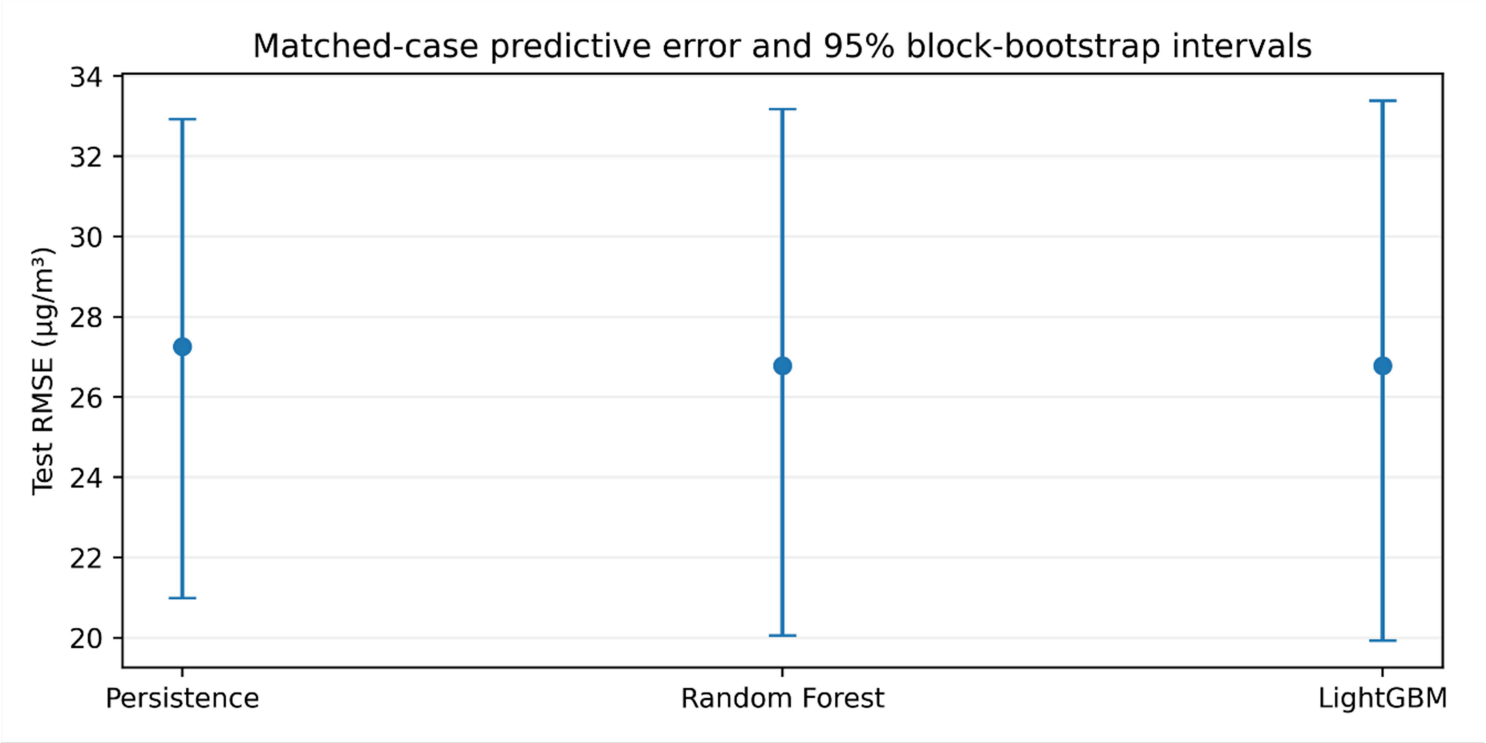


**Figure 5**

Matched-case RMSE for persistence and fitted learning models, with 95% synchronized moving-block bootstrap intervals (24-hour blocks; 1,000 replications). Lower RMSE indicates smaller forecast error

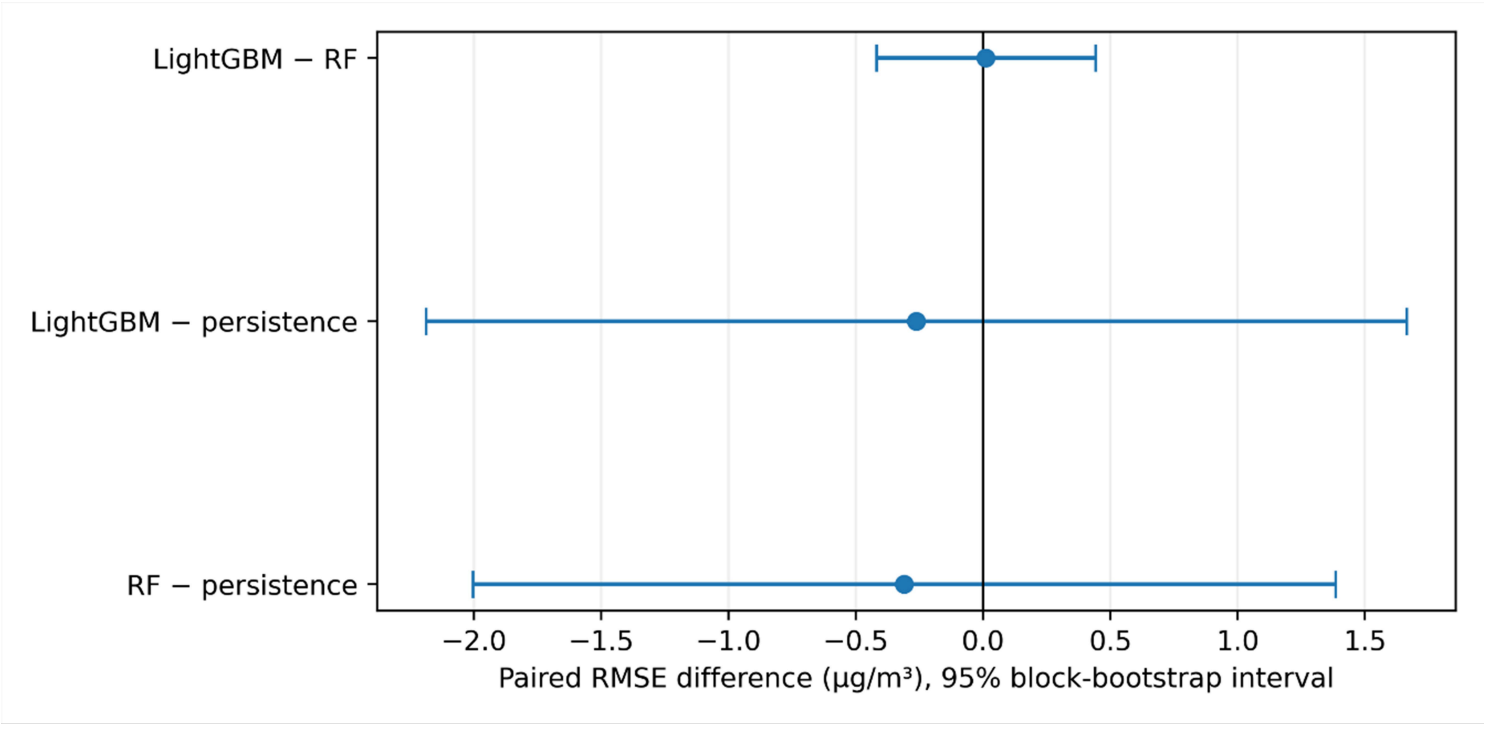


**Figure 6**

Paired RMSE-difference intervals from synchronized 24-hour moving-block resampling. Intervals crossing zero do not demonstrate a directional performance difference at the corresponding interval level