

Article

Interpretable Machine Learning Models Using SHAP for Hourly and Daily-Maximum Carbon Monoxide Forecasting at Urban Air Quality IoT Monitoring Stations in Greece

Yiannis Kiouvrekis ^{1,2,*} , Christos Christakis ³ , Ioannis Tsilikas ⁴  and Theodor Panagiotakopoulos ⁵ 

¹ Mathematics, Computer Science and Artificial Intelligence Laboratory, University of Thessaly, 43100 Karditsa, Greece

² School of Business, University of Nicosia, 2417 Nicosia, Cyprus

³ Department of Environmental Sciences, University of Thessaly, 41500 Larissa, Greece; cchristakis@uth.gr

⁴ Department of Applied Physics and Mathematics, National Technical University of Athens, 15772 Athens, Greece; tsilikasgiannis@hotmail.com or itsilikas@uniwa.gr

⁵ Department of Management Science and Technology, University of Patras, 26504 Patras, Greece; tepanag@gmail.com or tpanagiotakop@upatras.gr

* Correspondence: kiouvrekis.y@uth.gr or yiannis.kiouvrekis@gmail.com

Abstract

Carbon monoxide (CO) remains an important urban air quality and traffic-exposure tracer despite rarely exceeding regulatory limits in modern European cities. This study presents an interpretable machine learning framework for short-term CO forecasting at two operational horizons—next hour and next-day maximum—applied identically to four automated monitoring stations spanning central-urban, urban, urban-background, and suburban typologies in the Greater Athens Area (2021–2024). Using only univariate CO history and calendar-derived features, four learners (support vector regression, random forest, gradient boosting, and a multilayer perceptron) were benchmarked against a naïve baseline under a chronological train/validation/test split. At the next-hour horizon, all learners outperformed naïve at every station, with gradient boosting achieving the best or joint-best skill (test $R^2 = 0.81$ – 0.90); Wilcoxon signed-rank tests confirmed that these small but consistent margins were statistically significant. The next-day-maximum task proved substantially harder ($R^2 = 0.44$ – 0.59), with the neural and random forest models overtaking gradient boosting and SVR losing competitiveness. A SHAP (Shapley Additive Explanations) analysis of the next-hour gradient-boosting model showed that the most recent hourly lag dominates the forecast, with an effect nearly an order of magnitude over any other predictor, with hour-of-day encoding and short-lag rolling statistics contributing secondary, sign-consistent effects—providing a transparent, mechanistic account of model behavior rather than a black-box skill score. Unlike ozone, a secondary pollutant whose predictability degrades toward the trafficked urban core, CO concentration and forecastability increase together at the traffic-dominated site, indicating that primary-pollutant forecasts are most reliable precisely where exposure is greatest. These findings support a horizon-specific, interpretable forecasting strategy for operational deployment on real-time monitoring networks.



Academic Editor: George Angelos Papadopoulos

Received: 6 July 2026

Revised: 27 July 2026

Accepted: 28 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: carbon monoxide; air quality forecasting; SHAP; explainable artificial intelligence; gradient boosting; urban air pollution; interpretable machine learning

1. Introduction

Carbon monoxide (CO) is a highly toxic gas produced as a product of the incomplete combustion of organic materials when combustion occurs under conditions of insufficient oxygen supply [1]. Such conditions prevail mainly in enclosed or inadequately ventilated spaces, where organic matter is not fully oxidized to carbon dioxide (CO₂), resulting in increased production of carbon monoxide. Significant sources of exposure include domestic combustion appliances such as stoves, fireplaces, braziers, grills, and burners, as well as industrial installations and internal combustion engines, particularly those of older-technology vehicles with elevated pollutant emissions [2]. Carbon monoxide is a colorless, odorless, tasteless, and non-irritating gas, characteristics that make its detection by humans practically impossible without the use of specialized detectors [3]. As a result, in an enclosed space its concentrations can rise significantly without being perceived by the people present, creating high-risk conditions for severe poisoning or even death.

The toxicity of carbon monoxide is due primarily to its extremely high affinity for hemoglobin, which is approximately 200–250 times greater than that of oxygen [4]. Upon binding, carboxyhemoglobin (COHb) is formed, which drastically reduces the delivery of oxygen to tissues and causes tissue hypoxia [5]. The clinical picture of acute poisoning depends on the concentration of the gas, the duration of exposure, and the particular characteristics of the exposed individual. The most common initial symptoms include headache, dizziness, weakness, nausea, vomiting, dyspnea, chest pain, and disturbances of mental status, such as confusion and impaired concentration. At higher concentrations or after prolonged exposure, seizures, loss of consciousness, coma, cardiac arrhythmias, and ultimately death due to severe hypoxia of vital organs may occur.

Exposure during sleep presents a particularly elevated risk, as individuals perceive neither the presence of the gas nor the early symptoms of poisoning. Likewise, the consumption of alcoholic beverages may mask the initial symptoms or reduce the ability to react in time, significantly increasing mortality. Furthermore, chronic exposure to low concentrations of carbon monoxide has been associated with cognitive disorders, memory impairment, concentration disturbances, depressive symptomatology, and other neuropsychiatric manifestations, which may persist even after removal from the source of exposure [6]. The prevention of carbon monoxide poisoning relies on the proper maintenance of combustion appliances, adequate ventilation of spaces, and above all the installation of certified carbon monoxide detectors, which can give early warning of dangerous gas concentrations and prevent potentially fatal incidents [7].

Contribution of This Study

This work's central contribution is not merely predictive accuracy but interpretable predictive accuracy: rather than treating the best-performing learner as a black box, we apply SHAP (Shapley Additive Explanations) [8] to decompose its next-hour forecasts into feature-level, additive contributions expressed in the physical units of the target (mg/m³). This yields two complementary layers of evidence. Global SHAP importance ranks predictors by their mean absolute contribution across the test set, quantitatively confirming that the model's skill is overwhelmingly attributable to the most recent hourly observation—consistent with, but far more precise than, the impurity-based importances typically reported in comparable studies. The SHAP beeswarm summary goes further, revealing the direction and shape of each feature's effect: high recent CO values push forecasts upward while low values pull them down, and the diurnal (hour-of-day) encoding contributes a smaller, sign-changing correction that traces the bimodal traffic cycle. Together, these analyses move the model from a validated but opaque predictor to a mechanistically interpretable one—showing that the learned model is, in effect, a nonlinear refinement

of naïve rather than an unexplained statistical artifact. This interpretability layer is what allows the study's second contribution, the traffic-core forecastability finding, to be read causally rather than merely correlatively: because SHAP identifies the same short-lag, traffic-timed features as dominant at the highest-concentration station, the coincidence of high CO and high predictability can be attributed to the strength and regularity of local combustion emissions rather than to an artifact of model selection. In this sense, SHAP is used here not as an add-on diagnostic but as the analytical bridge connecting model performance to the underlying physical and temporal structure of urban CO pollution.

2. Related Work

Short-term forecasting of carbon monoxide and related pollutants with data-driven models is an active field. A growing body of work applies recurrent and operator-based neural networks to CO prediction. For the industrial city of Sakarya in Turkey, daily CO was modeled with long short-term memory (LSTM) and bidirectional LSTM networks and benchmarked against artificial neural networks (ANNs), k -nearest neighbors, and support vector regression, using past CO and PM₁₀ values at lags of up to three days; the recurrent models were the most accurate ($R = 0.91$), ahead of the kernel-based SVR ($R = 0.88$) [9]. A seasonal gated recurrent unit (SGRU) model was developed for long-record CO data (2005–2021) from the Taiwan environmental authority, with Kalman filter preprocessing, and outperformed a range of statistical and recurrent baselines—including ARIMA, seasonal ARIMA, exponential smoothing, and seasonal LSTM—achieving the lowest error among the compared models [10]. Satellite observations have also been exploited: a Sentinel-5P CO product accessed through a cloud Earth-observation platform was forecast with an LSTM time-series model and validated against held-out observations [11]. Most recently, a complex neural-operator architecture (CoNOAir) was proposed for CO forecasting across India; it surpassed a Fourier neural operator for both hourly and 72 h horizons at national scale, captured extreme-pollution episodes, and exceeded $R^2 = 0.95$ for hourly predictions across the evaluated cities, with the authors framing the system as an early-warning tool [12]. These studies establish deep learning as a strong approach for CO, but they typically depend on large multivariate or remotely-sensed inputs and report skill without a naïve baseline.

A second strand benchmarks competing algorithms on a common dataset—the design we adopt. For Petaling Jaya in Malaysia, six regression families (linear, regression-tree, Gaussian process, tree ensembles, support vector, and ANN models) were compared for CO prediction from concentration and time inputs; a Matérn-5/2 Gaussian process regressor was the most reliable, attaining $R^2 \approx 0.97$ with low error across scenarios, although the authors flagged the single-area focus as a limitation on generalizability [13]. A multi-station Malaysian study predicted CO at 12 and 24 h horizons with decision trees, random forest, support vector machines, and an ANN multilayer perceptron, using meteorological and co-pollutant predictors (wind speed, humidity, O₃, NO_x, SO₂, NO₂); the ANN-MLP performed best, achieving an R^2 of up to 0.89, and the analysis was strengthened with an uncertainty assessment and a Taylor diagram [14]. The lag-scenario designs of the Sakarya and Petaling Jaya studies make this explicit, systematically varying the number of antecedent days supplied to the model [9,13].

3. Methods and Materials

3.1. Study Stations and Areas

Hourly ambient CO data were obtained from four automated stations of the national air pollution monitoring network in the Greater Athens Area, covering 1 January 2021 to 31 December 2024.

The stations span a gradient of local combustion intensity: a central-urban traffic site (PAT), an urban site (PIR)—located near the Piraeus port basin and therefore additionally exposed to shipping, freight-handling, and harbor-traffic emissions alongside ordinary road traffic—an urban-background residential site (SMY), and a suburban background site (GEO). Their characteristics and CO climatology are summarized in Table 1. Mean concentration decreases monotonically along the gradient—from 1.15 mg/m³ at the heavily trafficked PAT site to 0.44 mg/m³ at the suburban GEO site—directly reflecting proximity to traffic emissions (Section 4.1). No station recorded an exceedance of the 10 mg/m³ eight-hour limit over the four years.

The spatial configuration of the four stations is shown in Figure 1. All lie within the Athens basin and are separated by no more than 12 km, so that they share a common synoptic and boundary-layer regime; the gradient they span is therefore one of local emission intensity rather than of meteorological setting. It should be emphasized that the typology labels follow the functional classification of the monitoring network, which is based on the character and proximity of local emission sources rather than on radial distance from the city center: a site may be classified as urban background or suburban while remaining geographically central, provided it is not directly exposed to a major traffic axis.



Figure 1. Location and typology of the four CO monitoring stations of the national network in the Greater Athens Area. Symbol shape denotes site typology and symbol area is proportional to the mean CO concentration over 2021–2024. The PIR site lies adjacent to the Piraeus port basin.

Table 1. Characteristics of the four CO monitoring stations (2021–2024). Valid % is the percentage of valid hourly values; concentrations in mg/m³; ‘Max 8 h’ is the highest 8 h running mean; ‘Exc 8 h’ counts days exceeding the 10 mg/m³ eight-hour limit value.

Station	Typology	Period	Valid %	Mean	Median	Max	P95	Max 8 h	Exc 8 h
PAT	Central urban	2021–2024	97.3	1.15	1.0	11.4	2.8	6.38	0
PIR	Urban	2021–2024	93.1	0.87	0.8	5.2	1.8	3.84	0

Table 1. Cont.

Station	Typology	Period	Valid %	Mean	Median	Max	P95	Max 8 h	Exc 8 h
SMY	Urban background	2021–2024	96.9	0.62	0.5	6.1	1.4	3.51	0
GEO	Suburban *	2021–2024	75.8	0.44	0.3	5.3	1.4	3.51	0

* GEO has reduced completeness relative to the other three stations, concentrated in 2024, the year used for the test partition (Section 3.1).

3.2. Dataset and Variables

The dataset comprises four plain-text files, one per monitoring station (CO_PAT, CO_PIR, CO_SMY, CO_GEO), covering 1 January 2021 to 31 December 2024. Each file is stored in a wide daily format: one row per calendar day, beginning with the date followed by the 24-hourly mean CO concentrations of that day (Table 2). Carbon monoxide is measured by non-dispersive infrared (NDIR) absorption photometry, the EU reference method, and reported as hourly means in mg/m^3 . Missing observations—arising from instrument downtime, calibration, or quality-assurance rejection—are encoded by the sentinel value -9999 . Each station therefore contains $1461 \text{ days} \times 24 \text{ h} = 35,064$ potential hourly values.

Table 2. Variables of the raw daily station files (wide format; 25 fields per row).

Field	Variable	Type/Units	Description
1	date	string "DD-MM-YYYY"	Calendar date of the record
2–25	h00–h23	float, mg/m^3	The 24-hourly mean CO concentrations, from 00:00 to 23:00; -9999 denotes a missing hour

From these files, station-level descriptive variables were derived to characterize the CO climatology and data completeness (Table 3; these populate Table 1). All concentration variables are expressed in mg/m^3 .

Table 3. Station-level descriptive variables.

Variable	Units	Description
Station	—	Station code (PAT, PIR, SMY, GEO)
Typology	—	Site classification along the emission gradient (central urban/urban/urban background/suburban)
Period	—	Data span (2021–2024)
Valid	%	Percentage of valid (non-missing) hourly values out of 35,064
Mean	mg/m^3	Arithmetic mean of all valid hourly CO values
Median	mg/m^3	Median (50th percentile) of hourly CO
Max	mg/m^3	Maximum single hourly value observed
P95	mg/m^3	95th percentile of hourly CO (upper-tail/episodic level)
Max 8h	mg/m^3	Highest 8 h running mean over the period
Exc 8h	days	Number of days exceeding the $10 \text{ mg}/\text{m}^3$ maximum-daily 8 h-limit value

3.3. Engineered Predictor Variables

All model inputs are derived from the CO series and the calendar, computed strictly from past values within gap-free segments (Section 3.2). The next-hour task uses 17 predictors with the target being $y = \text{CO}(t)$ (Table 4); the next-day-maximum task uses 11 daily predictors with the target being the following day's maximum hourly CO (Table 5).

Table 4. Predictor variables for the next-hour task (target $y = \text{CO}(t)$; naïve baseline = lag1).

Variable	Units	Description
lag1–lag168	mg/m^3	CO observed 1, 2, 3, 6, 12, 24, 48, and 168 h before the target hour (short memory, daily, and weekly periodicity)
rmean3, rstd3	mg/m^3	Rolling mean and standard deviation over the previous 3 h (shifted 1 h)

Table 4. *Cont.*

Variable	Units	Description
rmean24, rstd24	mg/m ³	Rolling mean and standard deviation over the previous 24 h (shifted 1 h)
hour_sin, hour_cos	—	Cyclical encoding of hour of day, sin / cos($2\pi h/24$)
doy_sin, doy_cos	—	Cyclical encoding of day of year, sin / cos($2\pi d/365.25$)
weekend	—	Binary indicator (1 = Saturday/Sunday)

Table 5. Predictor variables for the next-day-maximum task (target = next day's maximum hourly CO, for days with ≥ 18 valid hours; naïve baseline = dmax_lag1).

Variable	Units	Description
dmax_lag1,2,3,7	mg/m ³	Daily maximum CO of 1, 2, 3, and 7 days before the target day
dmean_lag1,2,3	mg/m ³	Daily mean CO of 1, 2, and 3 days before the target day
dmax_roll7	mg/m ³	7-day rolling mean of the daily maximum (shifted 1 day)
doy_sin, doy_cos	—	Cyclical day-of-year encoding for the target day
weekend	—	Weekend indicator for the target day

Rationale for a Univariate Design

The predictor set defined above excludes meteorological, traffic, and co-pollutant covariates, all of which are known to modulate ambient CO. Three considerations motivated this choice. First, operational deployability. The pipeline is designed to run directly on the monitoring network's real-time data stream. A univariate model requires no external data feed, no data-sharing agreement with a meteorological provider, and has no additional mode of failure when an upstream source is delayed or unavailable, properties of practical consequence for a system intended to operate continuously within an environmental authority. Second, and more fundamentally, an operational forecast cannot use observed meteorology. At the day-ahead horizon, the meteorological state at the target time is not known and must itself be forecast, so a multivariate model deployed operationally would inherit the error of the numerical weather prediction fields on which it depends. Studies that quantify the benefit of meteorological covariates using observed rather than forecast fields therefore establish an optimistic upper bound on the attainable improvement rather than an operationally realizable one. Reporting the skill available from the pollutant's own history establishes the corresponding lower bound, against which any multivariate extension, and the cost of its external dependencies, can be assessed. Third, the design supports a controlled comparison. This study is a cross-pollutant counterpart to a companion analysis of ground-level ozone, employing an identical pipeline and feature set; introducing exogenous covariates would confound differences attributable to the target pollutant with differences attributable to the predictor set.

3.4. Autocorrelation and Partial Autocorrelation

To characterize the temporal memory of hourly CO and to motivate the lag structure of the engineered predictors, the series at each station was examined using the autocorrelation function (ACF) and the partial autocorrelation function (PACF) [15]. The ACF quantifies the overall linear association between a series and its own lagged values, without controlling for the effect of intermediate lags; for a series X_t , the ACF at lag k is defined as

$$\rho_k = \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t) \text{Var}(X_{t-k})}}, \quad (1)$$

where $\text{Cov}(\cdot, \cdot)$ and $\text{Var}(\cdot)$ denote the covariance and variance operators, respectively. Because ρ_k mixes the direct dependence of X_t on X_{t-k} with the indirect dependence transmitted through the intervening lags $X_{t-1}, \dots, X_{t-k+1}$, a slowly decaying or periodically

re-peeking ACF is characteristic of trend or seasonal structure rather than of any single lag in isolation.

The PACF isolates the direct association between X_t and X_{t-k} after removing the linear effect of the intermediate observations. It is defined as the partial correlation,

$$\phi_k = \frac{\text{Cov}(X_t, X_{t-k} \mid X_{t-1}, \dots, X_{t-k+1})}{\sqrt{\text{Var}(X_t \mid X_{t-1}, \dots, X_{t-k+1}) \text{Var}(X_{t-k} \mid X_{t-1}, \dots, X_{t-k+1})}}, \quad (2)$$

so that a significant PACF value at lag k indicates a contribution from X_{t-k} that is not already explained by the shorter lags. In this sense the ACF describes the aggregate correlation structure of the series, whereas the PACF is the conventional diagnostic for identifying candidate autoregressive orders. Together, the two functions provide complementary evidence for lag selection: the ACF reveals recurring periodicities, while the PACF indicates which of those lags carry genuinely independent information once shorter-range dependence has been accounted for. ACF and PACF were computed for each station's hourly CO series over lags of up to 72 h.

3.5. Machine Learning Models

Four supervised learning algorithms, spanning distinct modeling paradigms, were selected to complement the naïve baseline and to provide a representative cross-section of approaches commonly used in short-term air quality forecasting. Support vector regression (SVR) [16] with a radial basis function (RBF) kernel represents a kernel-based method that maps the input features into a higher-dimensional space to capture nonlinear relationships while controlling model complexity through a small number of hyperparameters (C, γ, ϵ). Random forest (RF) [17] is a bagging-based ensemble of decision trees, each trained on a bootstrap resample of the data with randomized feature subsampling at each split; predictions are obtained by averaging across trees, which reduces variance and provides a natural, impurity-based measure of feature importance. Gradient boosting (GBM) [18] is a boosting-based tree ensemble that builds trees sequentially, with each new tree fitted to the residual errors of the current ensemble; this typically yields higher predictive accuracy than bagging-based ensembles at the cost of greater sensitivity to hyperparameter choices such as learning rate and tree depth. Finally, a multilayer perceptron (MLP) [19] was included as a neural network reference, consisting of two fully connected hidden layers (64 and 32 units) with ReLU activations, trained with the Adam optimizer and early stopping to prevent overfitting; unlike the tree-based methods, the MLP requires standardized inputs and is capable of learning smooth, continuous nonlinear mappings between predictors and the target. Together, these four learners span linear-kernel, bagging, boosting, and neural network paradigms, allowing the comparison to probe whether any one modeling philosophy is consistently better suited to short-term CO forecasting, rather than restricting the analysis to a single algorithmic family. As a baseline against which the learned models were benchmarked, we adopted the naïve method, one of the simplest and most widely used forecasting approaches for level time series [20]. Under naïve forecasting, the one-step-ahead forecast is set equal to the most recently observed value, $\hat{y}_t = y_{t-1}$, which corresponds to a random-walk data-generating process, $y_t = y_{t-1} + \epsilon_t$. This choice of baseline is particularly appropriate for a series such as hourly CO, which—like temperature over very short intervals—exhibits strong short-term persistence: because the underlying process changes little from one hour to the next, the naïve forecast can already be highly accurate. Naïve carries no parameters to estimate, updates instantaneously to genuine shifts in level, but—by construction—has a memory of only one observation and therefore cannot filter out noise, propagating it directly into the forecast. For the next-day-maximum task, the corresponding baseline was the seasonal analogue of

this principle, using the previous day's maximum value as the forecast for the following day. Despite its simplicity, naïve (and its seasonal counterpart) is widely regarded as an indispensable forecasting benchmark: if a more complex model cannot outperform it, that model offers no practical forecasting value [20].

A further consideration concerns the availability and representativeness of meteorological inputs at the hourly resolution of this study. The analysis spans four years of hourly observations at each station, and station-located meteorological measurements are not available at matching resolution for the full record; gridded reanalysis or data from the nearest synoptic station could be substituted, but these are not collocated with the air quality monitors and would introduce a representativeness error at the street and neighborhood scale on which CO concentrations vary. In place of exogenous meteorological drivers, the four-year record provides repeated sampling of the diurnal and seasonal cycles through which meteorological variability principally acts, so that its systematic component is captured implicitly by the calendar encodings and by the rolling dispersion features. We regard the incorporation of collocated or bias-corrected meteorological covariates—together with the forecast meteorology that an operational day-ahead system would in any case require—as the clearest direction for future work, particularly for the next-day-maximum horizon, where dispersion plays the largest role.

3.6. Hyperparameter Grid Search

Each learner's hyperparameters were tuned by grid search over the space defined in Table 6, evaluating every combination and selecting the one that minimized the root-mean-square error on the chronological validation partition (the middle 15% of each series); the held-out test partition was never used for selection. The grid comprises 12 configurations for SVR, 8 for Random forest, 8 for GBM, and 4 for the MLP, i.e., 32 candidate configurations per station and horizon. Inputs were standardized using training-partition statistics for the kernel- and gradient-based learners (SVR, MLP), whereas the tree ensembles (RF, GBM) are scale-invariant; for tractability the SVR grid was fitted on a random training subsample. Settings outside the grid were held at fixed, conventional values—an RBF kernel for SVR, and the Adam optimizer with ReLU activations, a maximum of 400 epochs, and early stopping for the MLP. The search space in Table 6 was informed by a companion study of ground-level ozone that employs the same pipeline, sampling frequency, and feature construction: because the two pollutants share these characteristics, the region of hyperparameter space containing sensible configurations is common to both, and the ozone study was used to delimit the grid rather than to fix the configuration. This does not presuppose that the optimal configuration is identical for CO. Preliminary searches on shorter CO subsamples confirmed that the chosen domain was appropriate—the optimum lay within the grid rather than at its edge—and the final configuration for each station and horizon was then selected by evaluating every combination in Table 6 and retaining the one that minimized the root-mean-square error on the CO validation partition, with the held-out test partition never used for selection.

Table 6. Hyperparameter grid-search space for each model. Every listed combination was evaluated; the configuration adopted in this study (matching the companion ozone study and held fixed across stations) is shown in **bold**. “Combinations” is the number of grid points searched per station and horizon.

Model	Hyperparameter	Search Space (Adopted in Bold)	Combinations
Naïve	—	Parameter-free	1
SVR	kernel	RBF (fixed)	12
	C	{1, 10, 100}	
	γ	{Scale, 0.1}	
	ϵ	{0.1, 0.5}	

Table 6. Cont.

Model	Hyperparameter	Search Space (Adopted in Bold)	Combinations
Random forest	n_estimators	{150, 300}	8
	max_depth	{None, 20}	
	max_features	{1.0, sqrt}	
GBM (hist. gradient boosting)	max_iter	{300, 400}	8
	learning_rate	{0.05, 0.1}	
	max_leaf_nodes	{31, 63}	
MLP	hidden_layer_sizes	{{ (64, 32) , (128, 64)}	4
	alpha	{ 10^{-4} , 10^{-3} }	
Total candidate configurations per station and horizon			32

3.7. Machine Learning Framework

A single, identical pipeline (Figure 2) was applied to every station. Each wide daily file was reshaped to a continuous hourly series on a regular one-hour index and the sentinel converted to missing. Completeness ranged from 75.8% (GEO) to 97.3% (PAT). Short gaps of up to three consecutive hours were filled by time interpolation and flagged; longer gaps were left missing and used to split each series into contiguous segments, so no lagged feature or training sequence bridges a gap. Interpolated values were used only as predictors; no interpolated observation was used as a prediction target, and hours within long gaps were excluded from both training and evaluation through the segmentation described above.

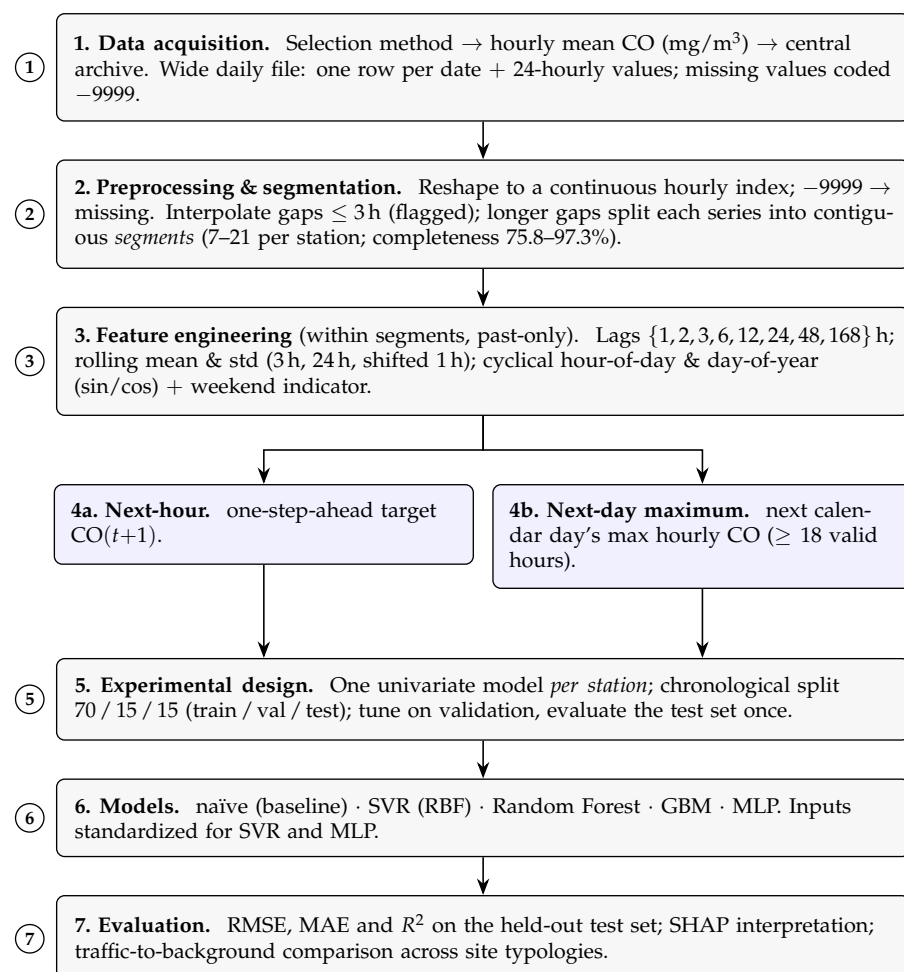


Figure 2. Overview of the univariate CO-forecasting pipeline, applied identically to every station: from data acquisition and gap-aware segmentation, through past-only feature engineering, to the two forecast horizons, per-station model training, and evaluation.

3.8. Interpretability (SHAP) Variables

The SHAP analysis introduces three quantities that are defined here. For a single prediction, the SHAP (Shapley) value ϕ_i of predictor i is its signed contribution to the model output, expressed in the units of the target (mg/m^3): it is the average marginal change in the predicted next-hour CO when feature i is added to every possible subset of the remaining features. A positive ϕ_i raises the forecast above the base value and a negative ϕ_i lowers it; the values satisfy local accuracy, so for each instance the base value (the mean model output over the background sample) plus the summed SHAP values equals the prediction. The *global importance* of a feature is the mean of its absolute SHAP values over the explained instances, $\frac{1}{N} \sum_n |\phi_i^{(n)}|$ (in mg/m^3). In the beeswarm summary, each point is one test instance; its horizontal position is the SHAP value; and its color encodes the *feature value* for that instance, min–max normalized between the 1st and 99th percentiles (blue = low, red = high); the vertical scatter within a row is density jitter only. The explanation is computed for the next-hour GBM model at the representative PAT station, and the twelve features with the largest mean $|\text{SHAP}|$ are displayed (Table 7).

Table 7. Variables appearing in the SHAP figures: the figure label, the corresponding engineered predictor, and its definition. Lag and rolling features are in mg/m^3 ; cyclical and weekend features are dimensionless.

Label in Figure	Variable	Definition
CO lag 1 h	lag1	CO 1 h before the target hour
CO lag 2 h	lag2	CO 2 h before
CO lag 3 h	lag3	CO 3 h before
CO lag 6 h	lag6	CO 6 h before
CO lag 12 h	lag12	CO 12 h before
CO lag 24 h	lag24	CO 24 h before (same hour, previous day)
CO lag 48 h	lag48	CO 48 h before
CO lag 168 h	lag168	CO 168 h before (same hour, previous week)
roll mean 3 h	rmean3	Mean CO over the previous 3 h (shifted 1 h)
roll std 3 h	rstd3	Std. dev. of CO over the previous 3 h (shifted 1 h)
roll mean 24 h	rmean24	Mean CO over the previous 24 h (shifted 1 h)
roll std 24 h	rstd24	Std. dev. of CO over the previous 24 h (shifted 1 h)
hour (sin)	hour_sin	$\sin(2\pi h/24)$, h = hour-of-day
hour (cos)	hour_cos	$\cos(2\pi h/24)$
day-of-yr (sin)	doy_sin	$\sin(2\pi d/365.25)$, d = day-of-year
day-of-yr (cos)	doy_cos	$\cos(2\pi d/365.25)$
weekend	weekend	Binary indicator, 1 = Saturday/Sunday

4. Results

The adopted model configuration, identified by validation-based tuning in the companion study and held fixed here, is summarized in Table 8; all selected values were stable across stations, with only the GBM iteration count differing by horizon in the companion study.

Table 8. Principal model settings. SVR was fitted on a random training subsample for tractability.

Model	Configuration
Naïve	Last value (next hour); previous-day maximum (next day)
SVR	RBF kernel, $C = 10$, γ = scale, standardized inputs
Random forest	150 trees, unbounded depth
GBM	Histogram gradient boosting, 400 iterations, $\text{lr} = 0.05$
MLP (NN ref.)	Two hidden layers (64, 32), standardized inputs, early stopping

4.1. Spatial and Typological CO Patterns

Mean CO decreased systematically from the traffic core to the suburban background (Figure 3, Table 1): PAT (1.15 mg/m³) > PIR (0.87) > SMY (0.62) > GEO (0.44). This ordering is the mirror image of the urban–rural ozone gradient and identifies the trafficked center, rather than the periphery, as the locus of CO exposure. The diurnal cycle (Figure 4) is bimodal at every site, with a morning and an evening maximum that coincide with the two traffic peaks; the amplitude is largest at PAT and smallest at GEO. The seasonal cycle (Figure 5) peaks in winter and reaches its minimum in summer at every station—the opposite phase to ozone—consistent with heating emissions, a shallow winter boundary layer, and a weaker oxidative sink.

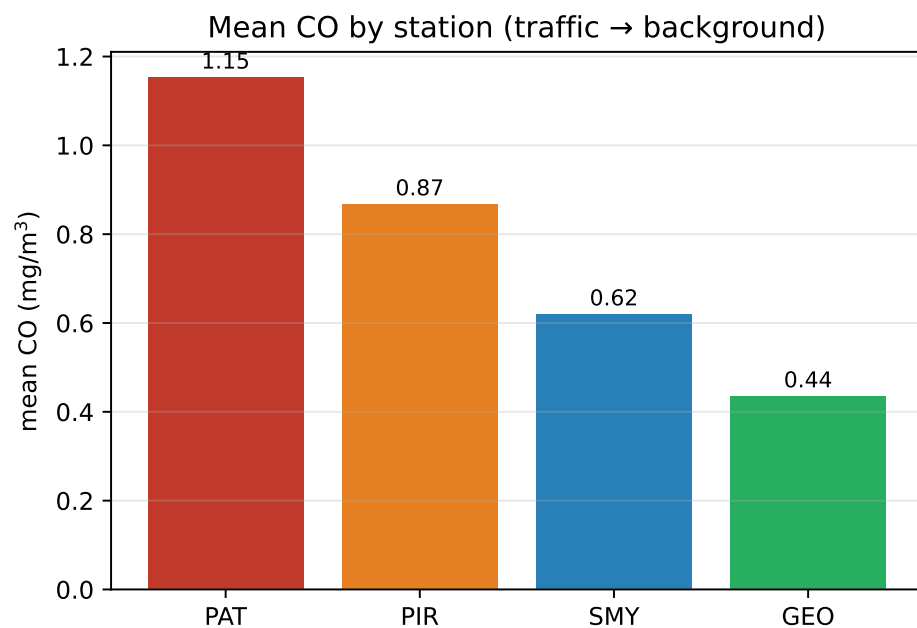


Figure 3. Mean CO concentration by station, ordered from the traffic core (PAT) to the suburban background (GEO).

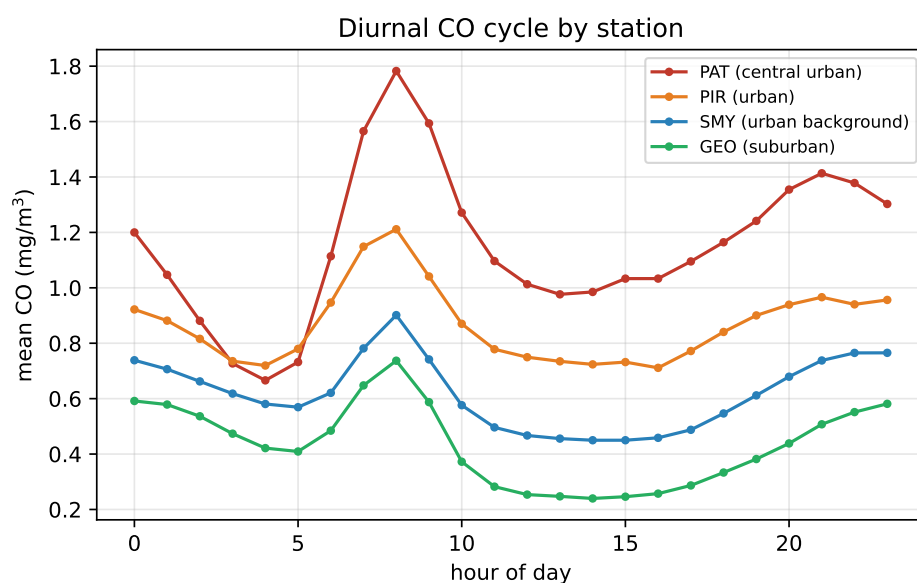


Figure 4. Mean diurnal CO cycle by station. The bimodal morning/evening structure tracks the two traffic rush hours; amplitude is largest at the central-traffic PAT site.

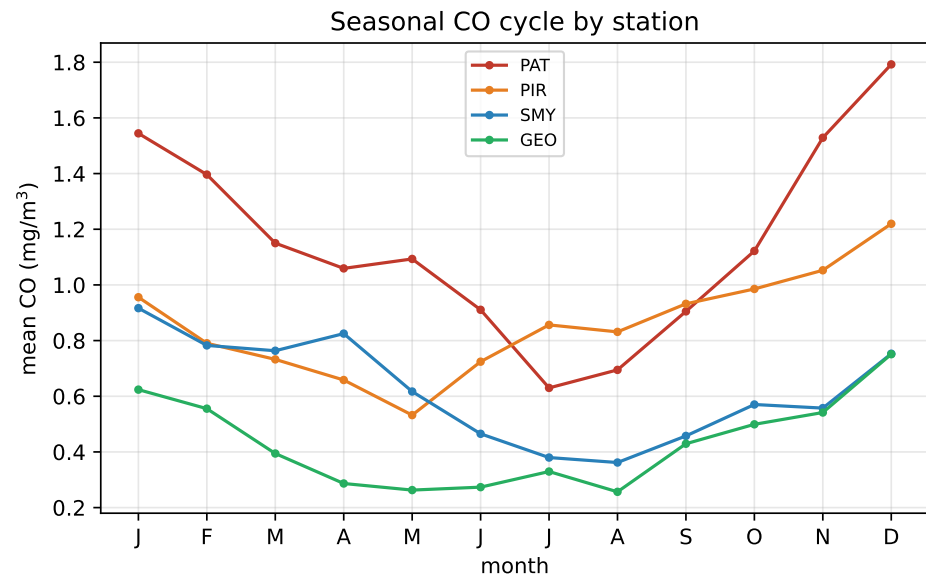


Figure 5. Mean monthly CO by station, showing the winter maximum and summer minimum—the opposite seasonal phase to ground-level ozone.

4.2. Temporal Structure

The autocorrelation and partial-autocorrelation functions (Figure 6, shown for the representative PAT site) reveal strong periodicity at 24 and 48 h and a partial autocorrelation dominated by the most recent hour with a secondary daily echo. The lag-1 autocorrelation is high at every station ($\rho_1 = 0.88\text{--}0.91$, mean: 0.90). This structure motivates the chosen lag set and explains the strength of the naïve baseline at the next-hour horizon.

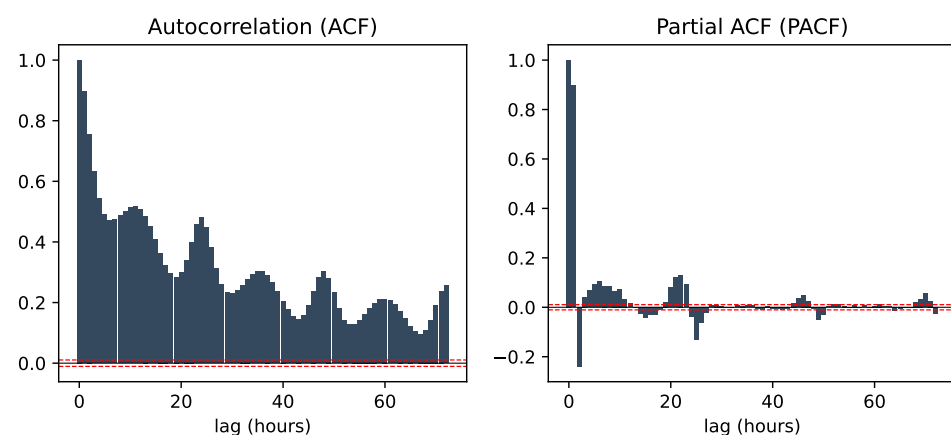


Figure 6. Autocorrelation (ACF) and partial autocorrelation (PACF) of hourly CO for a representative station, showing the dominant 24 h periodicity and lag-1 dominance.

4.3. Next-Hour Forecasting

At the next-hour horizon, all ML models outperformed naïve at every station (Table 9, Figure 7). Gradient boosting (GBM) achieved the best or joint-best R^2 at most stations, with R^2 ranging from 0.815 at the urban PIR site to 0.899 at the urban-background SMY site; random forest and the MLP were consistently close behind. The margin over naïve, however, was modest (for example, R^2 rising from 0.832 to 0.888 at PAT; GBM naïve skill scores of 0.16–0.27 across sites), reflecting the very high one-step autocorrelation of CO. The example time series (Figure 8) and the predicted-versus-observed scatter (Figure 9) show that the bimodal daily structure and the sharp morning/evening peaks are well reproduced.

Table 9. Next-hour (one-step-ahead) forecasting performance on the held-out test partition. RMSE in mg/m^3 . The best R^2 per station is in bold.

Station	Naïve		SVR		RF		GBM		MLP	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
<i>Central urban</i>										
PAT	0.362	0.832	0.336	0.855	0.305	0.881	0.295	0.888	0.301	0.884
<i>Urban</i>										
PIR	0.208	0.737	0.222	0.699	0.180	0.803	0.175	0.814	0.175	0.815
<i>Urban background</i>										
SMY	0.186	0.809	0.169	0.841	0.143	0.887	0.135	0.899	0.144	0.886
<i>Suburban</i>										
GEO	0.240	0.772	0.238	0.775	0.195	0.849	0.184	0.866	0.201	0.839

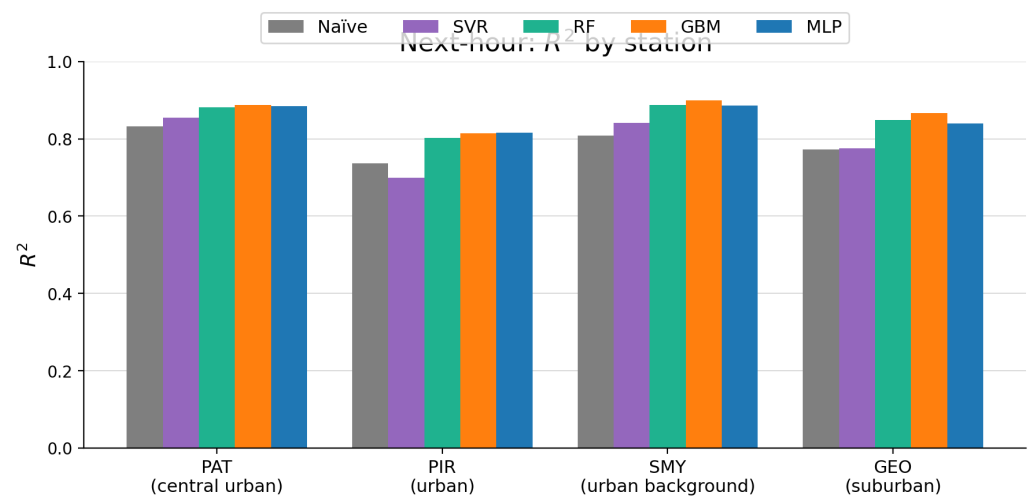


Figure 7. Test-set R^2 by station and model for the next-hour task. Stations are ordered from the traffic core (left) to the suburban background (right).

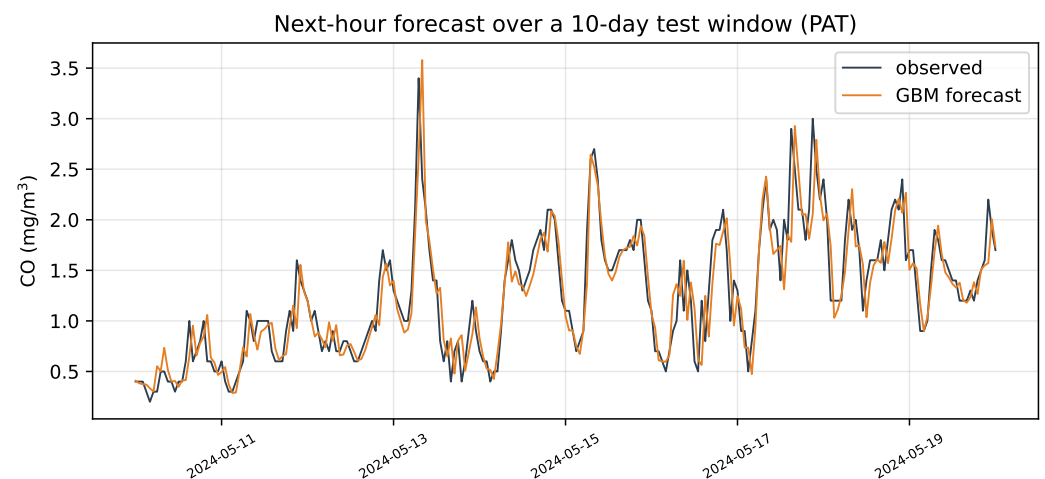


Figure 8. Observed and next-hour-forecast CO over a representative ten-day window of the test period (PAT). The window is chosen for legibility; aggregate skill is given in Table 9.

4.4. Next-Day-Maximum Forecasting

The next-day-maximum task was substantially harder, with test R^2 between 0.44 and 0.59 for the best model at each station (Table 10, Figure 10). Here, the relative advantage of the ML models over naïve was much larger than at the next-hour horizon, and the ranking among learners changed: the neural reference (MLP) was best at three of the four stations and random forest was best at the fourth (PIR), whereas support vector regression

underperformed, falling below naïve at PIR. This horizon-dependent change in the leading algorithm was consistent across the network and is a central finding of the study.

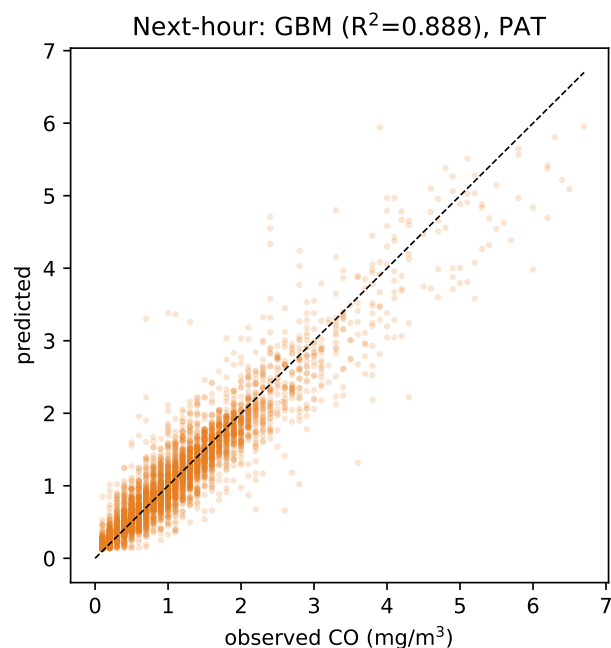


Figure 9. Predicted versus observed next-hour CO on the test set for the best model (GBM) at a representative station; the dashed line is the 1:1 reference.

Table 10. Next-day-maximum forecasting performance on the held-out test partition. RMSE in mg/m^3 . The best R^2 per station is in bold.

Station	Naïve		SVR		RF		GBM		MLP	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
<i>Central urban</i>										
PAT	1.112	0.291	0.944	0.490	0.894	0.541	0.994	0.434	0.842	0.594
<i>Urban</i>										
PIR	0.581	0.203	0.664	−0.040	0.489	0.437	0.543	0.304	0.508	0.393
<i>Urban background</i>										
SMY	0.659	0.360	0.676	0.327	0.589	0.489	0.612	0.448	0.551	0.553
<i>Suburban</i>										
GEO	0.750	0.384	0.776	0.340	0.759	0.368	0.712	0.444	0.696	0.468

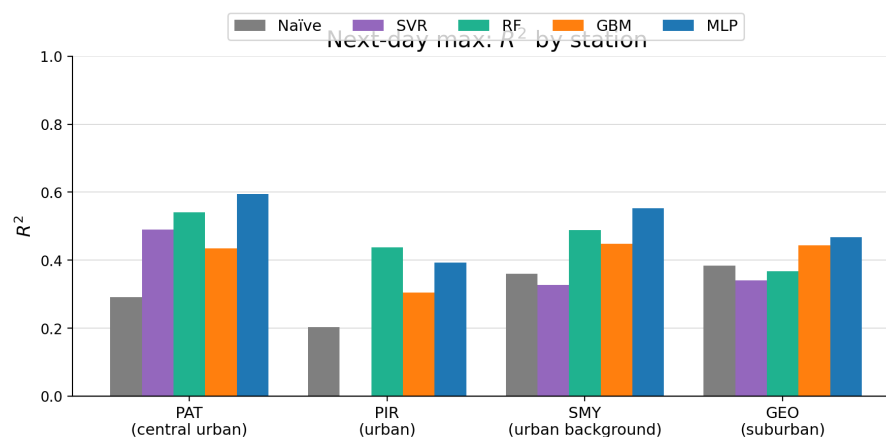


Figure 10. Test-set R^2 by station and model for the next-day-maximum task.

4.5. Statistical Significance of Close R^2 Differences

Because several models reached very similar R^2 at the next-hour horizon (Table 9)—in particular GBM, RF and the MLP differ by $\Delta R^2 \leq 0.03$ at every station—we tested whether these small differences are statistically significant rather than due to chance. For each station we applied the Wilcoxon signed-rank test, a non-parametric paired test that makes no normality assumption, to the per-sample absolute errors of the two models on the same held-out test set (Table 11). GBM significantly outperformed random forest at all four stations ($p < 0.001$), confirming that its advantage, though numerically small, is systematic rather than incidental. Against the MLP the outcome is mixed: GBM was significantly more accurate at the central-urban, urban-background, and suburban sites, but significantly less accurate at the urban PIR site, where the MLP is the preferred next-hour model—consistent with the per-station best models in Table 9. We note that, as the hourly errors are themselves temporally autocorrelated, the reported p -values are best read as indicative; the consistency of the GBM-versus-RF result across all stations nonetheless supports GBM as the preferred general-purpose next-hour model.

Table 11. Statistical significance of the close next-hour R^2 differences, assessed with the Wilcoxon signed-rank test on paired per-sample absolute errors (held-out test set). ΔR^2 is the GBM advantage over the competitor; absolute R^2 values are in Table 9. Significance: *** $p < 0.001$.

Station	R^2_{GBM}	GBM vs. RF		GBM vs. MLP	
		ΔR^2	p -Value	ΔR^2	p -Value
Central urban					
PAT	0.888	+0.008	<0.001 ***	+0.005	<0.001 ***
Urban					
PIR	0.814	+0.011	<0.001 ***	−0.001	<0.001 ***
Urban background					
SMY	0.899	+0.012	<0.001 ***	+0.014	<0.001 ***
Suburban					
GEO	0.866	+0.017	<0.001 ***	+0.027	<0.001 ***

4.6. Error Analysis and Feature Importance

The residual distributions (Figure 11) are centered near zero for every model; naïve carries the heaviest tails—consistent with its higher RMSE—while the ML models are comparable across the interquartile range, so their advantage lies mainly in suppressing the large errors of the sharp morning and evening peaks. A feature-importance analysis of the random forest (Figure 12) shows that the one-hour lag dominates the next-hour forecast, with the hour-of-day encoding and the 24 h rolling statistics as secondary contributors—consistent with the autocorrelation structure and with the strength of the naïve baseline.

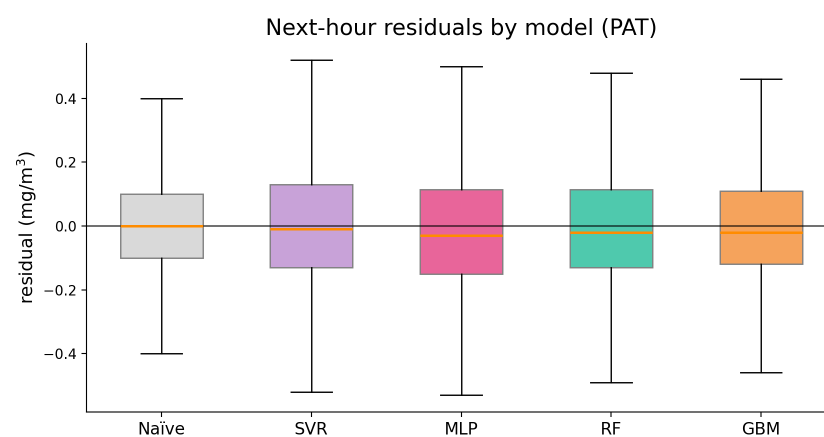


Figure 11. Distribution of next-hour prediction residuals by model at a representative station (PAT).

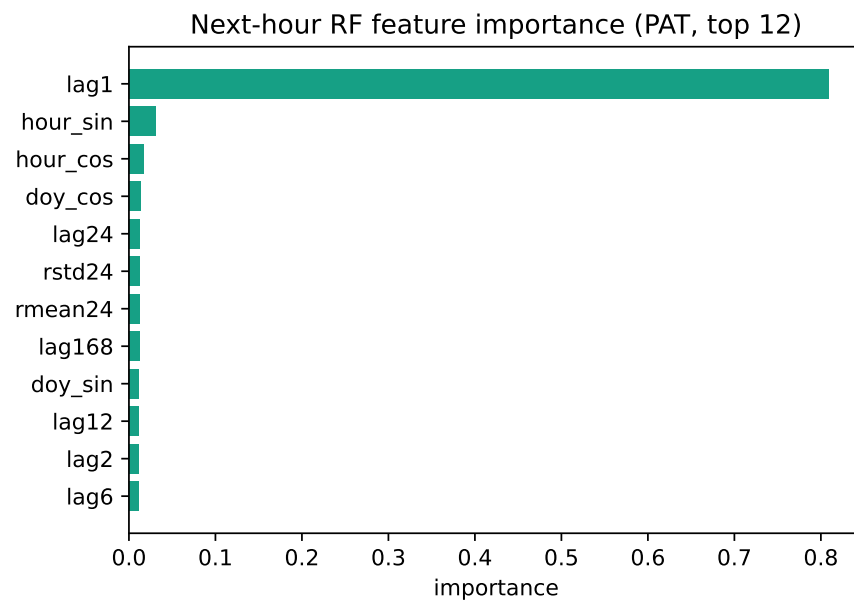


Figure 12. Random forest feature importance for the next-hour task (PAT), dominated by the most recent (1 h) lag.

4.7. SHAP-Based Interpretation

The Shapley analysis of the next-hour GBM model (Figures 13 and 14) refines the impurity-based picture. The one-hour lag is by far the dominant driver, with a mean absolute SHAP value roughly seven times that of any other feature, followed by the hour-of-day encoding and then a group of shorter lags and the 24 h rolling statistics. The beeswarm plot (Figure 14) shows the expected monotonic mechanism: high recent-CO values (red) push the forecast strongly upward while low values (blue) pull it down, with the widest spread of any feature. The hour-of-day terms contribute a smaller, sign-changing effect that encodes the bimodal traffic cycle. This confirms that the next-hour model operates essentially as a learned, nonlinear refinement of naïve—explaining both its high skill and the limited margin over the naïve baseline.

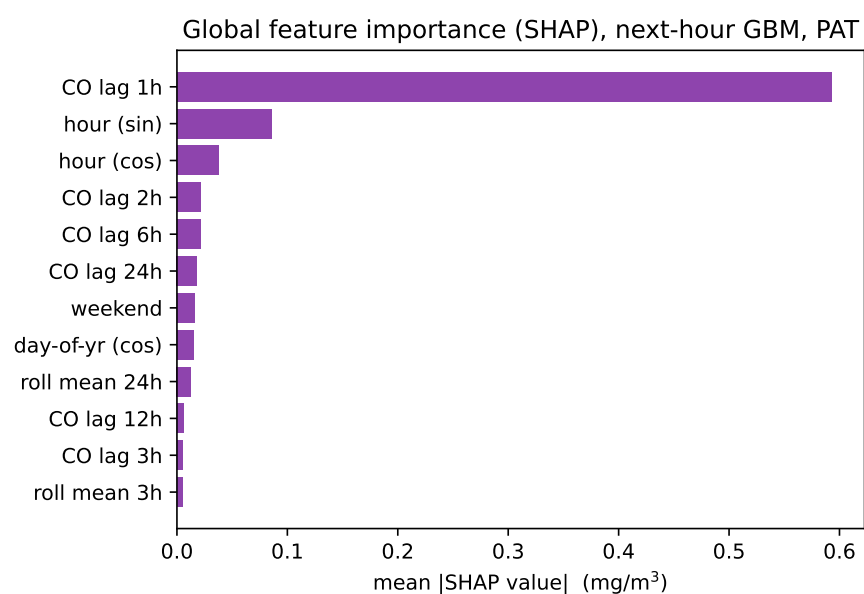


Figure 13. Global feature importance from SHAP (mean absolute Shapley value) for the next-hour GBM model at PAT. The one-hour lag dominates.

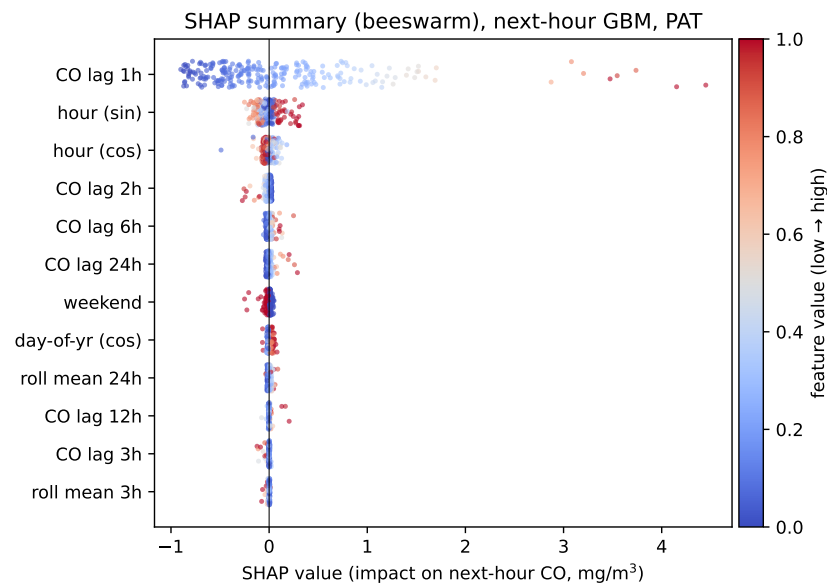


Figure 14. SHAP summary (beeswarm) for the next-hour GBM model at PAT. Each point is one test instance; color encodes the feature value (blue low, red high) and the horizontal position its signed impact on the forecast.

The SHAP analysis was carried out for the next-hour GBM model at all four stations. The global feature-importance ranking was found to be identical across the network: at every site the one-hour lag is the dominant predictor, with its effect roughly an order of magnitude greater than any other feature; it is followed by the hour-of-day encoding, the shorter lags, and the 24 h rolling statistics, in that order and with the same signs. The absolute magnitudes differ between sites, scaling with the local concentration level, but the structure of the explanation does not. Because the four decompositions are effectively indistinguishable in their ordering and interpretation, we present the analysis for the representative central-urban PAT station only (Figures 13 and 14) and omit the remaining three for concision; their inclusion would add near-identical figures without further interpretive content. The consistency of the decomposition across the traffic-to-background gradient indicates that the mechanism identified at PAT—a learned, nonlinear refinement of persistence in which recent CO dominates and the diurnal encoding contributes a smaller, sign-changing correction—is a property of the CO signal rather than a feature of any single station.

SHAP explains the model in terms of the features it was given: for each prediction it distributes the output across the input variables, showing how much each one contributes and in which direction. It does not point to emission sources directly. The bimodal traffic cycle, in particular, is not carried by any single feature and cannot be read from one SHAP value alone; it emerges from several features acting together: the hour-of-day encoding, which contributes positively during the morning and evening peaks and negatively between them; together with the short lags and the daily and weekly lags that carry the recurring emission schedule. This is precisely why the feature set combines lagged, rolling, and calendar terms rather than relying on a single predictor: the diurnal structure is represented jointly across them. The beeswarm summary shows the expected pattern, high recent-CO values push the forecast up and low values pull it down, while the hour-of-day terms show the mixed, sign-changing effect characteristic of a time-of-day contribution.

4.8. Accuracy Across Site Typologies: Cross-Matrix

To compare areas and accuracy directly, Table 12 and Figure 15 cross site typology with model. Two patterns stand out. At the next-hour horizon, every typology is forecast well

(R^2 between 0.81 and 0.90) and gradient boosting is the best or joint-best column throughout. At the next-day horizon, the picture changes sharply: accuracy falls to $R^2 \approx 0.4$ –0.6, the neural reference and random forest lead, and SVR is no longer competitive. Notably, the trafficked PAT site is not the least forecastable; at the next-day horizon it is in fact the most forecastable ($R^2 = 0.594$), so that high concentration and high predictability coincide at the source.

Table 12. Cross-matrix of forecasting accuracy (test R^2) by station/typology and model, for both horizons. The single station per fine typology class is a limitation (see text).

Station	Next Hour					Next-Day Max				
	Naïve	SVR	RF	GBM	MLP	Naïve	SVR	RF	GBM	MLP
PAT (central urban)	0.832	0.855	0.881	0.888	0.884	0.291	0.490	0.541	0.434	0.594
PIR (urban)	0.737	0.699	0.803	0.814	0.815	0.203	−0.040	0.437	0.304	0.393
SMY (urban background)	0.809	0.841	0.887	0.899	0.886	0.360	0.327	0.489	0.448	0.553
GEO (suburban)	0.772	0.775	0.849	0.866	0.839	0.384	0.340	0.368	0.444	0.468

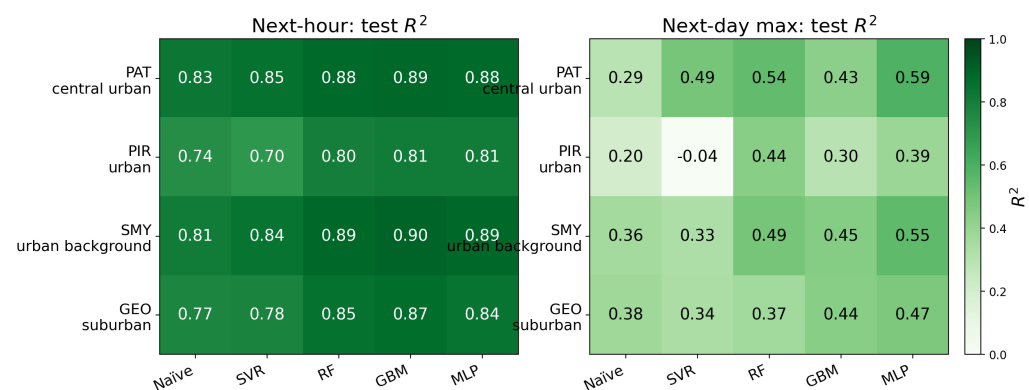


Figure 15. Cross-matrix heatmap of test R^2 by station/typology (rows) and model (columns) for the next-hour (left) and next-day-maximum (right) tasks. Darker is more accurate.

Figure 16 presents the same accuracies geographically. Two features are apparent. First, next-hour skill is high and spatially uniform: the light bars are of comparable height at all four sites, so that the strong short-range autocorrelation of CO supports reliable nowcasting irrespective of typology or position within the basin. Second, next-day skill is both lower and more spatially variable, and—contrary to what would be expected for a secondary pollutant—it is not lowest at the trafficked core. The tallest dark bar is at PAT, the site with the highest mean concentration, while the shortest is at the port-adjacent PIR site. The figure therefore renders the central result of Section 4.9 in its spatial form: at the day-ahead horizon, the location of greatest exposure was also the location of greatest predictability, an ordering consistent with the strength and regularity of the local emission signal, though with a single station per typology class this ordering is a site-level observation rather than an established property of the classes.

4.9. Effect of Site Typology on Predictability

Across both horizons, predictability did not decline toward the trafficked center—the opposite of the ozone pattern. At the next-hour horizon, all four typologies were forecast comparably well (best-model R^2 between 0.81 and 0.90), with the urban site PIR the least predictable, plausibly reflecting its proximity to port and shipping activity, emission sources that are more episodic and less tightly coupled to the regular commuter-traffic cycle than at the other three stations. At the next-day horizon, the central-urban PAT site was in fact the *most* forecastable (best-model $R^2 = 0.594$), ahead of the urban-background and suburban

sites, while PIR again trailed. Thus, for a primary pollutant, the strong, regular, high-amplitude emission signal at the traffic core makes the location of greatest exposure also the most forecastable. This both inverts the ozone result—where the heavily titrated central-urban site was the least predictable—and is reinforced by the variance-normalization property of R^2 , since the high-variance traffic core offers a more explainable signal than the low-amplitude background sites.

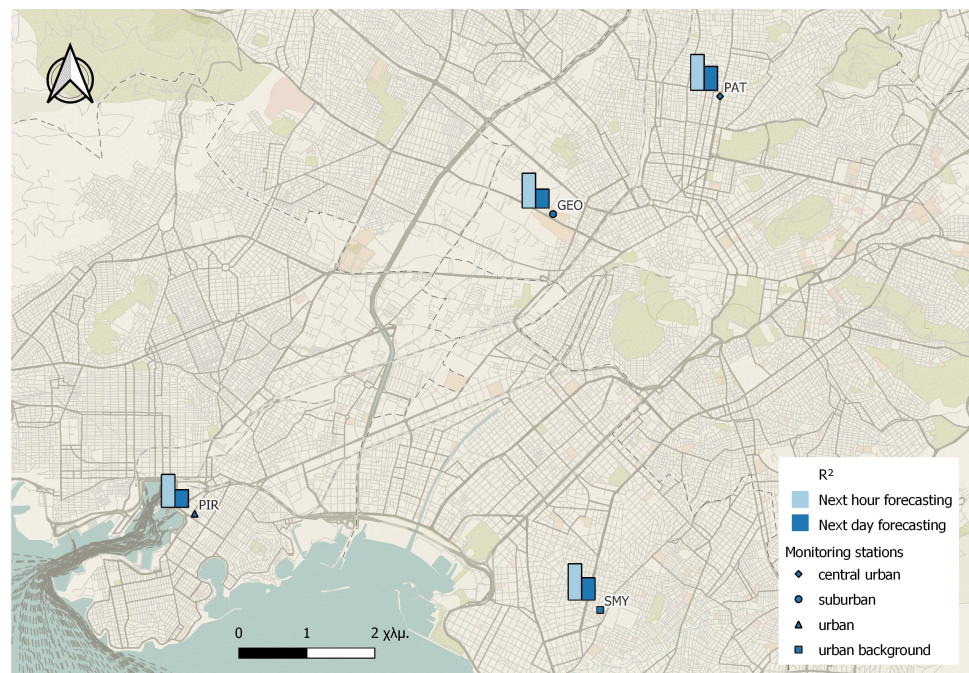


Figure 16. Spatial distribution of forecasting skill. Paired bars give the test-set R^2 of the best-performing model at each station for the next-hour (light) and next-day-maximum (dark) horizons. Bar heights are drawn on a common scale from 0 to 1. Symbol shape denotes site typology as in Figure 1.

5. Discussion

5.1. Principal Findings

Three results emerge consistently. First, at the next-hour horizon every ML model outperformed naïve at every site, with gradient boosting the best or joint-best model and test R^2 between 0.81 and 0.90; the margin over naïve was nonetheless small. Second, the next-day-maximum task was markedly harder (R^2 between 0.44 and 0.59), the relative gain from ML was larger, and the leading algorithm changed, with the neural reference and random forest leading, and SVR underperforming. Third, mean CO and its forecastability did not decline toward the city center; the trafficked PAT site was the most concentrated and, at the next-day horizon, the most forecastable. That these patterns hold across four stations using a single univariate pipeline suggests they are properties of the CO signal rather than artifacts of any one site.

5.2. Why the Best Model Depends on the Horizon

Hourly CO is strongly autocorrelated (Figure 6), so the next-hour target is dominated by the most recent observation; naïve is already a strong baseline and the lag-1 feature dominates the learned models (Figure 12). In this regime the ML learners win by modeling modest nonlinear corrections around a near-persistent signal, but the margin is small (naïve skill scores of 0.16–0.27). That this margin is nonetheless real and not noise is confirmed by the Wilcoxon test (Section 4.5): GBM significantly beats random forest at every station. The

next-day maximum, by contrast, is a smoother, lower-signal-to-noise target that requires genuine extrapolation rather than extension of the latest value; here, the flexible neural reference and the tree ensembles generalized better, while the fixed-hyperparameter SVR struggled and at one site fell below naïve. The practical implication is that an operational system should not commit to a single algorithm: the appropriate model is horizon-specific.

5.3. Linking the Emission Regime to Forecastability

The most informative aspect of these results is the contrast with ozone. For a secondary pollutant, titration by fresh nitric oxide suppresses ozone in the urban core and also degrades its short-term predictability, so the trafficked center is both the lowest and the least forecastable. For CO—a primary pollutant emitted at the same locations—the relationship inverts: the trafficked core carries the strongest, most regular, emission-driven signal, so it is both the most concentrated and (at the day-ahead horizon) the most forecastable. Concentration level and forecastability are therefore coupled in opposite directions for the two pollutants. This matters operationally because it means CO forecasting is most reliable precisely where exposure is greatest.

5.4. Comparison with Previous Work

The next-hour skill obtained here ($R^2 \geq 0.81$ at every site, univariately) is consistent with the broader comparative literature in which lag features dominate and gradient-boosting and tree-ensemble methods are a strong baseline. The more modest next-day skill is also consistent with reports that longer-horizon forecasting benefits substantially from meteorological drivers, which we deliberately excluded to keep the pipeline deployable from a single analyzer. The finding that naïve is a strong next-hour baseline for CO—harder to beat than for ozone—adds a pollutant-specific nuance to that literature.

5.5. Implications for Early-Warning and Operational Use

Because ambient CO now rarely approaches the regulatory limit, the operational value of these models lies less in threshold warning than in real-time nowcasting, exposure mapping, and the use of CO as a combustion tracer. Next-hour nowcasting is reliable across all site types and is well suited to real-time public information and traffic-related exposure estimation. Day-ahead prediction of the daily maximum is feasible but should use the neural reference or random forest rather than SVR, and should attach greater uncertainty to its estimates given the lower skill at this horizon. The univariate design makes both directly transferable to the network's real-time data stream.

5.6. Limitations

Several limitations frame these conclusions. The study is univariate; adding meteorological variables (temperature, wind speed, boundary-layer height) and traffic or co-pollutant indicators would most plausibly improve the harder next-day task and is the clearest next step. The GBM results used a histogram-based implementation as a stand-in for XGBoost, and the deep recurrent (LSTM) model anticipated for the next-hour task was not run; both should be added to complete the comparison. SVR was evaluated with fixed hyperparameters, mirroring a companion ozone study; tuning C and γ per station might recover some of its next-day performance. The analysis uses one pollutant from four stations in a single conurbation, with a single station per fine typology class, so the typology comparison cannot be averaged within class; the GEO site has reduced completeness concentrated in the test year. Finally, the models were trained per station, so spatial transferability between sites was not assessed.

6. Conclusions

This study developed a single, reproducible univariate pipeline and applied it to compare machine learning models for short-term carbon monoxide forecasting at four stations spanning a traffic-to-background gradient in the Greater Athens Area, across two operational horizons. At the next-hour horizon, every learner outperformed the naïve baseline at every station, with gradient boosting achieving the best or joint-best skill using only CO history (test R^2 of 0.81–0.90); this advantage, though numerically modest given the strong autocorrelation of hourly CO, was shown by the Wilcoxon signed-rank test to be systematic rather than incidental. The SHAP analysis of the gradient-boosting model clarified why this margin over naïve is small yet real: the model operates essentially as a learned, nonlinear correction around the one-hour lag, which dominates the forecast, with an effect roughly an order of magnitude greater than any other feature, with the diurnal cycle and short rolling statistics contributing secondary, physically interpretable adjustments. Forecasting the next day's maximum concentration proved considerably more difficult, with test R^2 between 0.44 and 0.59, and the ranking of models shifted accordingly—the multilayer perceptron and random forest took the lead while support vector regression lost competitiveness, underscoring that no single algorithm is optimal across both horizons. A further, and arguably more consequential, finding is that CO concentration and forecastability move together rather than in opposition: the trafficked central-urban station carried both the highest mean concentration and, at the day-ahead horizon, the highest predictive skill, in direct contrast to the behavior typically reported for secondary pollutants such as ozone, whose predictability degrades where titration by fresh emissions is strongest. Because this pattern held consistently across the four stations, analyzed with an identical univariate pipeline, it appears to reflect a genuine property of the CO emission and dispersion regime rather than an artifact specific to any one site. Taken together, these results support an operational forecasting strategy that is both horizon-specific and interpretable—gradient boosting for next-hour nowcasting and a neural or random forest model for day-ahead maxima—deployable directly on a real-time monitoring network. Future work should extend the feature set with meteorological and traffic covariates, complete the model comparison with XGBoost and a recurrent (LSTM) architecture, explore multi-step and probabilistic forecast formulations, and broaden the station and pollutant coverage to test whether the coupling between concentration and forecastability observed here for CO generalizes to other primary pollutants.

Author Contributions: Conceptualization, Y.K. and T.P.; methodology, Y.K.; software, C.C.; validation, T.P.; formal analysis, Y.K. and I.T.; investigation, Y.K. and T.P.; data curation, Y.K. and T.P.; writing—original draft preparation, Y.K.; writing—review and editing, Y.K. and T.P.; visualization, I.T.; supervision, Y.K. and T.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw hourly carbon monoxide measurements analyzed in this study are publicly available from the Greek Ministry of Environment and Energy through the National Air Pollution Monitoring Network at <https://ypen.gov.gr/perivallon/poiotita-tis-atmosfairas/dedomena-metriseon-atmosfairikis-rypansis/> (accessed on 5 July 2026). The processed datasets and the analysis code generated during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CO	Carbon monoxide
ML	Machine learning
MLP	Multilayer perceptron
SVR	Support vector regression
RF	Random forest
GBM	Gradient boosting machine
RMSE	Root-mean-square error
R^2	Coefficient of determination
SHAP	Shapley Additive Explanations
ACF	Autocorrelation function
PACF	Partial autocorrelation function
doy	Day of year
WHO	World Health Organization

References

1. World Health Organization. *WHO Guidelines for Indoor Air Quality: Selected Pollutants*; World Health Organization Regional Office for Europe: Copenhagen, Denmark, 2010.
2. Agency for Toxic Substances and Disease Registry (ATSDR). *Toxicological Profile for Carbon Monoxide*; NCBI Bookshelf, Bookshelf ID: NBK153693; ATSDR Toxicological Profiles; Agency for Toxic Substances and Disease Registry (US): Atlanta, GA, USA, 2012.
3. Das, S. Chapter 20—Toxic gases. In *Toxicology Cases for the Clinical and Forensic Laboratory*; Ketha, H., Garg, U., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 387–396. [\[CrossRef\]](#)
4. McMahon, K.; Launico, M.V. *Carbon Monoxide Toxicity*; StatPearls [Internet]; NCBI Bookshelf ID: NBK430740; StatPearls Publishing: Treasure Island, FL, USA, 2025.
5. Jasani, S. Chapter 147—Smoke Inhalation. In *Small Animal Critical Care Medicine*, 2nd ed.; Silverstein, D.C., Hopper, K., Eds.; W.B. Saunders: St. Louis, MO, USA, 2015; pp. 785–788. [\[CrossRef\]](#)
6. Amitai, Y.; Zlotogorski, Z.; Golan-Katzav, V.; Wexler, A.; Gross, D. Neuropsychological Impairment From Acute Low-Level Exposure to Carbon Monoxide. *Arch. Neurol.* **1998**, *55*, 845–848. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Moberg, M.E.; Hamilton, E.B.; Zeng, S.M.; Bryazka, D.; Zhao, J.T.; Feldman, R.; Abate, Y.H.; Abbasi-Kangevari, M.; Abdurehman, A.M.; Abedi, A.; et al. Global, regional, and national mortality due to unintentional carbon monoxide poisoning, 2000–2021: Results from the Global Burden of Disease Study 2021. *Lancet Public Health* **2023**, *8*, e839–e849. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Lundberg, S.M.; Lee, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874. [\[CrossRef\]](#)
9. Feizi, H.; Sattari, M.T.; Prasad, R.; Apaydin, H. Comparative analysis of deep and machine learning approaches for daily carbon monoxide pollutant concentration estimation. *Int. J. Environ. Sci. Technol.* **2023**, *20*, 1753–1768. [\[CrossRef\]](#)
10. Yang, C.H.; Chen, P.H.; Wu, C.H.; Yang, C.S.; Chuang, L.Y. Deep learning-based air pollution analysis on carbon monoxide in Taiwan. *Ecol. Inform.* **2024**, *80*, 102477. [\[CrossRef\]](#)
11. Verma, A.; Ranga, V.; Vishwakarma, D.K. Forecasting of Satellite Based Carbon-Monoxide Time-Series Data Using a Deep Learning Approach. In Proceedings of the 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 11–12 February 2023; pp. 1–4. [\[CrossRef\]](#)
12. Bedi, S.; Tiwari, K.; Prathosh, A.P.; Kota, S.H.; Krishnan, N.M.A. A neural operator for forecasting carbon monoxide evolution in cities. *npj Clean Air* **2025**, *1*, 2. [\[CrossRef\]](#)
13. Almubaidin, M.A.; binti Ismail, N.S.; Latif, S.D.; Ahmed, A.N.; Dullah, H.; El-Shafie, A.; Sonne, C. Machine learning predictions for carbon monoxide levels in urban environments. *Results Eng.* **2024**, *22*, 102114. [\[CrossRef\]](#)
14. Latif, S.D.; Almalayih, M.; Yafouz, A.; Ahmed, A.N.; Zaini, N.; Irwan, D.; AlDahoul, N.; Sherif, M.; El-Shafie, A. Prediction of atmospheric carbon monoxide concentration utilizing different machine learning algorithms: A case study in Kuala Lumpur, Malaysia. *Environ. Technol. Innov.* **2023**, *32*, 103387. [\[CrossRef\]](#)
15. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021.
16. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
17. Ho, T.K. Random Decision Forests. In Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95), Montreal, QC, Canada, 14–16 August 1995; pp. 278–282. [\[CrossRef\]](#)
18. Friedman, J.H. *Stochastic Gradient Boosting*; Technical report; Stanford University: Stanford, CA, USA, 1999.

19. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2006. [[CrossRef](#)]
20. Svetunkov, I. *Forecasting and Analytics with the Augmented Dynamic Adaptive Model (ADAM)*; OpenForecast; Chapman and Hall/CRC: Boca Raton, FL, USA, 2023; Chapter 3.3.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.