eng

MDPI

Article

Reliability-Oriented Predictive Emission Monitoring of Gas Turbines Under Multiyear Distribution Shift and High-Emission Tail Risk

Omar Rodríguez-Abreo [1,2,*], Brenda S. Dublan-Barragán [3], José Marcos Zea Pérez [3], Carlos Alberto González-Gutiérrez [4,*], Luis Angel Iturralde Carrera [2] and Juvenal Rodríguez-Reséndiz [2]

[1] Programa de Ingeniería en Datos e Inteligencia Artificial, Universidad Politécnica de la Zona Metropolitana de Guadalajara, Tlajomulco de Zúñiga 45670, Mexico
[2] Facultad de Ingeniería, Universidad Autónoma de Queretaro, Santiago de Queretaro 76010, Mexico; liturralde28@alumnos.uaq.mx (L.A.I.C.); juvenal@uaq.edu.mx (J.R.-R.)
[3] Sistemas Inteligentes, Ingeniería Mecatrónica, Universidad Politécnica de Querétaro, Santiago de Querétaro 76240, Mexico; brenda.dublan@upq.mx (B.S.D.-B.); marcos.zea@upq.mx (J.M.Z.P.)
[4] Engineering Department, Universidad del Valle de Mexico, Santiago de Querétaro 76230, Mexico
* Correspondence: omar.rodriguez@upzmg.edu.mx (O.R.-A.); calberto.gonzalez@uvmnet.edu (C.A.G.-G.)

Abstract

Reliable deployment of predictive emission monitoring systems (PEMSs) requires evaluation under future operating conditions and explicit attention to high-emission events. This study presents a pollutant-specific, risk-constrained framework for predicting Gas Turbine CO and NO_x emissions under multiyear distribution shift. A dataset containing 36,733 observations was divided chronologically: 2011–2013 were used for model development, 2014 for configuration selection and calibration, and 2015 as a locked final test. Matched random within-prefix and forward temporal validation were compared, and Extra Trees was selected for both pollutants. Conventional, tail-weighted, covariate-shift-weighted, and combined configurations were evaluated under global-RMSE budgets of 2%, 5%, and 10%. Random within-prefix validation produced RMSE estimates 25.0% lower for CO and 34.0% lower for NO_x than matched temporal validation. In 2015, CO tail weighting produced modest changes, reducing global RMSE by 0.9% and Q_{99} mean underprediction loss by 2.0%, without statistically resolved Q_{99} improvement. For NO_x, tail weighting reduced Q_{99} RMSE by up to 5.5% while increasing global RMSE by 5.8%. Linear calibration reduced NO_x global RMSE from 13.246 to 6.333 mg/m^3 but worsened Q_{99} RMSE. Overall, temporal generalization, global calibration, and high-emission performance were distinct criteria, and no adaptation strategy was uniformly preferable.

Keywords: predictive emission monitoring systems; gas turbines; carbon monoxide; nitrogen oxides; temporal distribution shift; tail-risk prediction; covariate-shift adaptation; importance weighting; underprediction; machine learning

Check for updates

Academic Editor: Antonio Gil Bravo

Received: 4 September 2026
Revised: 28 September 2026
Accepted: 2 October 2026
Published: 9 October 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.

1. Introduction

Gas turbines play a central role in combined-cycle power generation because of their high power density, operational flexibility, and capacity to respond rapidly to variations in electricity demand. Their combustion process produces carbon monoxide (CO) and nitrogen oxides (NO_x), whose concentrations depend on ambient conditions, fuel–air interactions, combustion temperature, turbine load, and the dynamic state of the equipment. CO is closely associated with incomplete combustion, whereas NO_x formation is strongly

influenced by flame temperature and oxygen availability. Accurate and continuous estimation of these pollutants is therefore essential for environmental management, combustion supervision, operational optimization, and regulatory compliance [1–3].

Predictive emission monitoring systems (PEMSs) provide a software-based approach for estimating pollutant concentrations from routinely measured operational and environmental variables. These systems learn the relationship between turbine conditions and reference emission measurements, enabling continuous prediction from variables already available in the plant control system. Neural networks, gradient boosting methods, and modern tabular machine learning approaches have demonstrated the feasibility of accurate data-driven emission estimation in industrial environments [4–6]. Consequently, the central challenge is no longer only predictive capacity under historical data, but also maintaining reliable performance when operating conditions evolve over time.

The Gas Turbine CO and NO_x Emission Data Set provides a valuable benchmark for studying this problem under multiyear operating conditions [7]. It contains 36,733 observations collected between 2011 and 2015 from a gas turbine, with nine ambient and process variables describing atmospheric and operating conditions. Kaya et al. introduced this dataset and established a benchmark PEMS formulation for CO and NO_x prediction [1]. Subsequent studies have applied symbolic regression, gradient boosting, deep forests, stacked ensembles, and Transformer-based models, obtaining increasingly accurate predictions [2,3,6,8,9]. However, these advances do not by themselves establish how the models behave when evaluated prospectively across different years.

Multiyear deployment makes the temporal design of model evaluation a central consideration. Long-term analyses of combined-cycle gas-turbine emissions have identified systematic trends and year-dependent offsets associated with evolving operating conditions [10]. A random partition distributes observations from different periods across the training and evaluation sets, thereby measuring interpolation among temporally mixed operating conditions. A chronological partition instead measures prospective generalization from historical observations to a later operational period. Structured cross-validation therefore provides a more appropriate framework when the intended use involves future operating conditions [11].

Temporal evolution can also produce dataset shift, broadly defined as a change in the data-generating distribution between model development and deployment [12]. Covariate shift is one possible case, in which the predictor distribution changes while the conditional response relationship remains stable. Importance weighting and discriminative domain learning provide established mechanisms for adaptation under this assumption [13–16]. Their practical usefulness, however, depends on adequate source–target overlap: severe domain separation can produce unstable density-ratio estimates and very small effective sample sizes. Moreover, year-dependent prediction bias may indicate that covariate shift alone is insufficient to describe the prospective prediction problem. Simpler alternatives such as recent-history training or labeled recalibration should therefore be compared under the same chronological framework rather than assuming that importance weighting is necessarily preferable.

Operational reliability also depends on predictive behavior in the upper part of the emission distribution. High-emission observations are relatively infrequent, yet they may carry greater environmental and operational relevance than observations near the center of the response distribution. Imbalanced-regression and extreme-event studies therefore motivate explicit evaluation of rare response regions [17,18]. For emission monitoring, global accuracy alone may conceal substantial error at source-defined Q_{90}, Q_{95}, or Q_{99} levels. In addition, reducing underprediction can increase overprediction elsewhere, so both symmetric tail error and asymmetric prediction costs should be examined.

Recent reliability-oriented PEMS research has incorporated adaptive learning and conformal prediction to generate uncertainty-aware estimates and prioritize model recalibration [19]. A remaining gap is the prospective comparison of temporal validation, domain adaptation, tail-oriented weighting, recent-history training, and recalibration within a single leakage-controlled protocol that also accounts for sparse-tail selection uncertainty.

Accordingly, this study develops a reliability-oriented predictive emission monitoring framework for CO and NO_x under multiyear distribution shift. Models are developed using observations from 2011 to 2013, while 2014 is used for configuration selection and calibration and 2015 constitutes a locked final test period. Matched-size random within-prefix and forward temporal validation are compared to quantify the optimism associated with temporally mixed evaluation. Distribution shift is assessed using univariate and multivariate diagnostics, common-support and importance-weight analyses, and residual diagnostics.

Predictive performance is evaluated globally and within source-defined Q_{90}, Q_{95}, and Q_{99} regions. Four learning configurations are examined: conventional learning (BASE), tail-weighted learning (TAIL), covariate-shift weighting (SHIFT), and their combination (SHIFT–TAIL). Configuration selection is performed on 2014 under prespecified global-RMSE allowances of 2%, 5%, and 10%, and the selected configurations are then frozen for the 2015 evaluation. The analysis is complemented by recent-history and quantile-regression baselines, 2014-based calibration, sensitivity to alternative tail objectives, moving-block selection-stability analysis, and paired moving-block inference. The contribution of the study is therefore an integrated prospective protocol for comparing adaptation strategies and their global–tail trade-offs rather than the proposal of a universally superior weighting method.

2. Materials and Methods

2.1. Data Source and Chronological Experimental Design

The experiments used the Gas Turbine CO and NO_x Emission Data Set from the UCI Machine Learning Repository [7]. The measurements were collected from a gas turbine located in northwestern Turkey between 1 January 2011 and 31 December 2015. The database contains 36,733 observations distributed as follows: 7411 observations from 2011, 7628 from 2012, 7152 from 2013, 7158 from 2014, and 7384 from 2015.

Nine operational variables were used as predictors: ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), air-filter differential pressure (AFDP), gas-turbine exhaust pressure (GTEP), turbine inlet temperature (TIT), turbine after temperature (TAT), turbine energy yield (TEY), and compressor discharge pressure (CDP). Carbon monoxide (CO) and nitrogen oxides (NO_x) were modeled as independent regression targets. CO was excluded from the NO_x predictor matrix, and NO_x was excluded from the CO predictor matrix. Calendar year and the within-file row index were used exclusively to preserve the supplied ordering and were not included as predictors.

The five annual CSV files were concatenated according to year and their original within-file row order. Column names were standardized, and the nine predictors and two targets were converted to numeric values. The data audit included annual sample counts, missing and nonfinite values, and exact duplicates across the measurement columns. No missing or nonfinite values were detected. Nine rows were flagged as exact duplicates across the measurement columns and were retained because identical aggregated sensor records may correspond to valid repeated operating conditions. Lagged and rolling predictors were not constructed because the distributed data do not include individual timestamps. Consequently, adjacency refers to the supplied row order and does not imply a verified constant time interval between consecutive records.

Annual predictor distributions were summarized using the arithmetic mean and sample standard deviation within each year (Table 1). These descriptive quantities were used to characterize the dataset and were not used for model selection or preprocessing.

Table 1. Annual descriptive statistics of the operational predictors. Values are reported as mean ± sample standard deviation.

Predictor	Unit	2011 ($n = 7411$)	2012 ($n = 7628$)	2013 ($n = 7152$)	2014 ($n = 7158$)	2015 ($n = 7384$)
AT	°C	17.11 ± 7.43	18.40 ± 7.66	17.60 ± 6.86	18.22 ± 6.99	17.23 ± 8.10
AP	mbar	1014.17 ± 6.29	1012.26 ± 6.38	1012.00 ± 6.29	1012.38 ± 5.97	1014.51 ± 6.90
AH	%	79.17 ± 13.47	79.07 ± 14.11	80.46 ± 14.13	82.14 ± 13.00	68.65 ± 13.54
AFDP	mbar	4.09 ± 0.66	4.31 ± 0.83	3.70 ± 0.81	3.91 ± 0.71	3.60 ± 0.61
GTEP	mbar	25.66 ± 4.33	25.18 ± 4.01	25.11 ± 4.35	25.74 ± 3.68	26.13 ± 4.47
TIT	°C	1084.73 ± 16.13	1082.89 ± 16.85	1081.57 ± 17.39	1078.84 ± 16.57	1078.97 ± 19.76
TAT	°C	544.50 ± 8.29	546.26 ± 7.33	545.78 ± 7.36	547.64 ± 4.61	546.64 ± 5.49
TEY	MWh	135.75 ± 16.21	132.68 ± 15.30	132.17 ± 16.35	132.91 ± 13.60	133.99 ± 16.18
CDP	mbar	12.21 ± 1.15	12.00 ± 1.05	11.97 ± 1.13	12.02 ± 0.95	12.10 ± 1.14

Note: AT, ambient temperature; AP, ambient pressure; AH, ambient humidity; AFDP, air-filter differential pressure; GTEP, gas-turbine exhaust pressure; TIT, turbine inlet temperature; TAT, turbine after temperature; TEY, turbine energy yield; CDP, compressor discharge pressure. Variable names, numerical values, and units in this table follow the released dataset metadata. No numerical unit conversion was applied to CDP because the reported magnitude and published unit label cannot be reconciled unambiguously from the distributed CSV files alone. Likewise, no additional full-load filtering was applied by the present analysis. Dry/wet concentration basis, oxygen-reference correction, and other emission-reference conditions were not inferable from the numerical files and therefore were not reconstructed computationally.

The experimental protocol was divided into three chronological stages. Data from 2011 to 2013 formed the source-development domain,

$$\mathcal{S} = \{2011, 2012, 2013\}, \qquad n_s = 22,191. \tag{1}$$

The 2014 observations formed a configuration-selection domain,

$$\mathcal{C} = \{2014\}, \qquad n_c = 7158, \tag{2}$$

and the 2015 observations formed the locked final test domain,

$$\mathcal{T} = \{2015\}, \qquad n_t = 7384. \tag{3}$$

According to Equations (1)–(3), model hyperparameters, the primary regression algorithm, and high-emission thresholds were estimated using only $\mathcal{S}$. The labels from $\mathcal{C}$ were subsequently used to select among the alternative BASE, TAIL, SHIFT, and SHIFT–TAIL configurations. The same 2014 labels were also used in the additional calibration experiments described below; calibration parameters were frozen before application to 2015. Once selected, the configuration associated with each global-error budget was frozen before the labels from $\mathcal{T}$ were evaluated. Accordingly, all results reported for 2015 constitute selection-independent prospective estimates.

For shift-based configurations, the operational predictor matrix of each target period was treated as an unlabeled batch. The 2014 predictors were used to estimate source-to-2014 domain probabilities during configuration selection. After the weighting method, strength, and clipping parameter had been selected, the 2015 operational predictors were used to re-estimate source-to-2015 domain probabilities. The 2015 CO and NO_x labels were not used to construct weights, fit models, choose configurations, or select thresholds. This procedure represents unsupervised transductive covariate-shift adaptation in which the target covariates are available as a batch at prediction time. It therefore corresponds to a full-target-batch adaptation setting rather than to causal rolling deployment. No forecast

lead time was defined; consequently, the high-emission analyses evaluate contemporaneous prediction and identification rather than early-warning capability.

2.2. Regression Models, Hyperparameter Selection, and Matched Validation

Six regression algorithms were included in the definitive experiment: a mean dummy regressor, Ridge regression, Random Forest, Extra Trees, histogram-based gradient boosting regression (HistGBR), and XGBoost. Ridge regression was embedded in a pipeline that standardized every predictor using means and standard deviations estimated exclusively from each training partition. Tree-based models were fitted on the original predictor scale.

Hyperparameters were optimized independently for CO and NO_x using only the 2011–2013 source domain. The source-only tuning stage evaluated 24 randomly sampled parameter configurations with five forward temporal splits. Model-specific search spaces included the Ridge regularization coefficient; the number of trees, maximum depth, number or proportion of candidate predictors, minimum leaf size, and minimum split size for Random Forest and Extra Trees; the learning rate, number of iterations, maximum leaf nodes, minimum leaf size, and L_2 regularization coefficient for HistGBR; and the number of estimators, maximum depth, learning rate, row and column subsampling, minimum child weight, gamma, and L_1 and L_2 regularization coefficients for XGBoost.

Candidate configurations were ranked first by mean temporal root mean squared error (RMSE) and then by mean absolute error (MAE). The best source-only parameters obtained during this tuning stage were stored and reused unchanged throughout the definitive analysis. Reusing the same parameters ensured that subsequent performance differences arose from sample weighting and chronological evaluation rather than from a second optimization using future labels.

After hyperparameter selection, random and forward temporal validation were compared using exactly matched training and validation sample sizes. Five successive source-domain partitions were first generated using TimeSeriesSplit. For the temporal fold k, let $\mathcal{I}_k^{\mathrm{tr}}$ and $\mathcal{I}_k^{\mathrm{va}}$ denote its chronological training and validation indices. The corresponding temporal prefix was defined as

$$\mathcal{P}_k = \mathcal{I}_k^{\mathrm{tr}} \cup \mathcal{I}_k^{\mathrm{va}}. \tag{4}$$

As defined in Equation (4), the random comparator was restricted to exactly the same historical prefix as its temporal counterpart. For the matched random comparator, the observations in $\mathcal{P}_k$ were randomly permuted, after which the same numbers of training and validation observations as in the temporal fold were assigned to the random partitions. Thus,

$$\left|\mathcal{I}_{k,\mathrm{random}}^{\mathrm{tr}}\right| = |\mathcal{I}_k^{\mathrm{tr}}|, \qquad \left|\mathcal{I}_{k,\mathrm{random}}^{\mathrm{va}}\right| = |\mathcal{I}_k^{\mathrm{va}}|. \tag{5}$$

Equation (5) guarantees identical training and validation sample sizes within each matched comparison. Restricting the random comparison to the same temporal prefix prevented differences in sample size or historical coverage from being conflated with the effect of randomization. The matched comparison was repeated with seeds 17, 42, 137, 2026, and 7777. These seeds controlled the random partition and stochastic model initialization, while the forward temporal partitions remained fixed.

Because hyperparameters had already been selected using the source-domain temporal folds, the matched validation analysis was intended as a controlled post-tuning comparison of random and chronological resampling strategies rather than as a fully nested estimate of final prospective generalization. Independent prospective performance was instead assessed through the subsequent 2014 configuration-selection stage and the locked 2015 evaluation.

The primary regression algorithm for each pollutant was selected as the tuned model with the lowest mean RMSE under matched forward temporal validation. The selected

model was refitted using all 22,191 source observations before its unweighted and weighted variants were evaluated in 2014 and 2015. In the definitive experiment, Extra Trees was selected independently for both targets. The source-only optimal Extra Trees configuration for CO used 389 trees, a maximum depth of 12, a predictor-sampling proportion of 0.6, a minimum leaf size of 1, and a minimum split size of 13. For NO_x, it used 451 trees, a maximum depth of 26, a predictor-sampling proportion of 0.8, a minimum leaf size of 1, and a minimum split size of 3.

For measured values y_i and predictions $\hat{y}_i$, the principal global regression metrics were

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{7}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{8}$$

Equations (6)–(8) quantify average absolute error, squared-error sensitivity, and explained variation, respectively. Bias was defined as

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i), \tag{9}$$

so that negative values represent average underprediction. According to Equation (9), positive values indicate average overprediction. The 95th percentile of the absolute error was also calculated to characterize large individual errors.

Underprediction was decomposed into frequency and severity. Its frequency was measured using the underprediction rate,

$$\mathrm{UR} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i < y_i), \tag{10}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. The mean underprediction loss, which retains zero loss for observations that were correctly predicted or overpredicted, was

$$\mathrm{MUL} = \frac{1}{n}\sum_{i=1}^{n}\max(y_i - \hat{y}_i, 0). \tag{11}$$

The conditional underprediction magnitude was calculated only among underestimated observations:

$$\mathrm{CMU} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)\mathbf{1}(\hat{y}_i < y_i)}{\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i < y_i)}. \tag{12}$$

When at least one observation was underestimated, these quantities satisfy

$$\mathrm{MUL} = \mathrm{UR} \times \mathrm{CMU}. \tag{13}$$

Thus, according to Equation (13), MUL reflects both the frequency and conditional magnitude of underprediction.

To determine whether reductions in underprediction were obtained simply by shifting predictions upward, analogous overprediction metrics were also computed. The overprediction rate was

$$\mathrm{OR} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i > y_i), \tag{14}$$

the mean overprediction loss was

$$\mathrm{MOL} = \frac{1}{n}\sum_{i=1}^{n}\max(\hat{y}_i - y_i, 0), \tag{15}$$

and the conditional overprediction magnitude was

$$\mathrm{CMO} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)\mathbf{1}(\hat{y}_i > y_i)}{\sum_{i=1}^{n}\mathbf{1}(\hat{y}_i > y_i)}. \tag{16}$$

Conditional quantities in Equations (12) and (16) were treated as undefined when their corresponding denominator was zero; they were recorded as missing values rather than as zero.

This decomposition distinguishes frequent small underpredictions from infrequent but severe errors.

For illustration, consider three observations with measured values $(10, 12, 15)$ mg/m^3 and predictions $(8, 13, 11)$ mg/m^3. Underprediction occurs in the first and third observations, giving UR $= 2/3 = 66.7\%$. The positive underprediction losses are $(2, 0, 4)$ mg/m^3, so MUL $= (2+0+4)/3 = 2$ mg/m^3, whereas CMU $= (2+4)/2 = 3$ mg/m^3. Accordingly, MUL $= (2/3)(3) = 2$ mg/m^3. The example shows that UR measures frequency, MUL represents average loss over all observations, and CMU measures severity conditional on underprediction.

2.3. Distribution Shift and Tail-Risk Assessment

Univariate covariate shift was quantified separately between the source domain and each of the 2014 and 2015 domains. For every operational predictor, the analysis calculated the two-sample Kolmogorov–Smirnov statistic and its associated p-value, the first Wasserstein distance, the Wasserstein distance normalized by the source interquartile range, the standardized mean difference (SMD), and the population stability index (PSI).

The SMD for predictor j was defined as

$$\mathrm{SMD}_j = \frac{\bar{x}_{s,j} - \bar{x}_{t,j}}{\sqrt{\left(s_{s,j}^2 + s_{t,j}^2\right)/2}}, \tag{17}$$

where the subscripts s and t denote the source and target domains. Equation (17) expresses the source–target mean difference relative to their pooled dispersion. PSI was calculated using bins defined by source-domain deciles:

$$\mathrm{PSI}_j = \sum_{b=1}^{B}\left(p_{t,jb} - p_{s,jb}\right)\ln\left(\frac{p_{t,jb}}{p_{s,jb}}\right), \tag{18}$$

where $p_{s,jb}$ and $p_{t,jb}$ are the proportions of source and target observations in bin b. Proportions were lower-bounded by 10^{-6} before evaluating the logarithm.

Multivariate shift was evaluated by assigning domain label $D = 0$ to source observations and $D = 1$ to observations from the target year. Source-to-2014 and source-to-2015 classifiers were fitted separately. The primary weighting grid retained the prespecified histogram-based gradient boosting (HGB) classifier with a learning rate of 0.05, 180 iterations, 31 maximum leaf nodes, a minimum leaf size of 30, and L_2 regularization of 1.0. To evaluate the sensitivity of density-ratio estimation, the reviewer-oriented analysis additionally considered a standardized logistic-regression domain classifier and both randomly stratified and contiguous-block outer validation schemes.

For random validation, five-fold stratified out-of-fold probabilities were calculated. For blocked validation, the source and target sequences were each divided into five con-

tiguous blocks and corresponding source/target blocks were held out together. For source observation $\mathbf{x}_i$, the resulting probability was

$$p_i = \Pr(D = 1 \mid \mathbf{x}_i). \tag{19}$$

Equation (19) was estimated out of fold so that each propensity used for weight diagnostics or weighting was obtained from a domain model that had not been fitted on that observation.

A nested probability-calibration sensitivity analysis was also performed for HGB. Within each outer training fold, raw domain probabilities were generated by three-fold stratified cross-fitting and converted to logits. A logistic Platt calibrator was fitted only to these inner out-of-fold predictions and then applied to the held-out outer fold. Thus, the held-out domain labels were not used for probability calibration.

Domain separability was summarized using ROC–AUC, PR–AUC, Brier score, log loss, the mean and selected percentiles of p_i in each domain, and a domain-classifier ambiguity index

$$A_D = \frac{2}{n_s + n_t} \sum_{i=1}^{n_s+n_t} \min(p_i, 1 - p_i). \tag{20}$$

According to Equation (20), A_D ranges from 0 for maximally confident domain classification to 1 when all domain probabilities equal 0.5. Because A_D depends on the classifier, class priors, validation protocol, and probability calibration, it was treated as a classifier-dependent uncertainty diagnostic rather than as a universal coefficient of distributional overlap. Domain classification was used as a shift diagnostic and as an unlabeled source of sample weights; it was not treated as the final predictive task.

Six domain-estimation sensitivity specifications were evaluated: HGB with random folds with and without nested Platt calibration, HGB with blocked folds with and without calibration, logistic regression with random folds, and logistic regression with blocked folds. The original HGB/random/uncalibrated specification was retained for the principal 49-configuration grid to preserve continuity with the prespecified analysis, while the alternatives were used to quantify sensitivity of domain discrimination, importance weights, covariate balance, and locked 2015 prediction performance.

Because changes in predictor distributions do not establish stability of the response mechanism, the covariate-shift assumption was examined further using year-dependent residual diagnostics. BASE predictions for 2014 and 2015 were averaged across the five seeds. Signed residuals were compared between years using Welch's test and the two-sample Kolmogorov–Smirnov test, absolute errors using the Mann–Whitney test, and residual dispersion using Levene's test. Conditional residual analyses additionally divided each operational variable and the BASE prediction into five reference-defined strata and compared residual behavior within corresponding 2014 and 2015 strata. These diagnostics were used to identify predictive miscalibration and year dependence; they were not interpreted as proof of a specific mechanism such as conditional shift.

The effect of weighting on covariate balance was evaluated in the locked 2015 domain. For each selected configuration and predictor, the unweighted source mean, weighted source mean, target mean, and absolute SMD before and after weighting were calculated. Weighted source variances used normalized sample weights with the finite-sample denominator

$$1 - \sum_{i=1}^{n_s} \omega_i^2, \qquad \omega_i = \frac{w_i}{\sum_{r=1}^{n_s} w_r}. \tag{21}$$

The normalization in Equation (21) was also used when assessing weighted source covariate balance.

Target observations were additionally stratified according to empirical source-domain support using propensity ranges. Two sensitivity definitions were used: overlap of the em-

pirical source and target propensity ranges and membership within the source 0.5th–99.5th percentile propensity interval. Locked 2015 regression performance was then reported separately for target observations classified as having adequate or insufficient source support.

High-emission thresholds were estimated exclusively from the source targets. For $q \in \{0.90, 0.95, 0.99\}$,

$$\tau_q = Q_q(\{y_i : i \in \mathcal{S}\}). \tag{22}$$

The thresholds in Equation (22) were therefore independent of both 2014 and 2015 target labels. These numerical thresholds were held fixed when counting and evaluating high-emission observations in 2014 and 2015. For target period $\mathcal{A}$, the corresponding tail subset was

$$\mathcal{A}_q = \{i \in \mathcal{A} : y_i \geq \tau_q\}, \qquad n_q = |\mathcal{A}_q|. \tag{23}$$

All global metrics were recalculated within $\mathcal{A}_q$. For example,

$$\mathrm{RMSE}_q = \sqrt{\frac{1}{n_q} \sum_{i \in \mathcal{A}_q} (y_i - \hat{y}_i)^2}, \tag{24}$$

and

$$\mathrm{MUL}_q = \frac{1}{n_q} \sum_{i \in \mathcal{A}_q} \max(y_i - \hat{y}_i, 0). \tag{25}$$

Equations (24) and (25) distinguish symmetric squared error from asymmetric underprediction loss within the same source-defined tail. Tail-specific UR and CMU were calculated analogously. Tail-specific OR, MOL, and CMO were also calculated to quantify the corresponding overprediction cost. Because the thresholds originated in the source domain, their prevalence was allowed to change naturally in 2014 and 2015.

To characterize whether the small tail subsets represented isolated rows or concentrated stretches of the record sequence, Q_{90}, Q_{95}, and Q_{99} observations were grouped into contiguous row clusters within each annual file. Because row-level timestamps were unavailable, these clusters were not interpreted as verified independent physical emission events.

2.4. Risk-Constrained Weighting, Calibration, and Alternative Baselines

Four regression variants were compared using the source-selected Extra Trees model: the unweighted baseline (BASE), tail weighting (TAIL), covariate-shift weighting (SHIFT), and the product of both components (SHIFT–TAIL). All regression models continued to use only labeled source-domain observations for fitting.

Tail weighting emphasized source observations at or above the source-defined Q_{95} threshold. Before stabilization, the tail component was

$$\tilde{w}_i^{\mathrm{tail}} = \begin{cases} 1 + \alpha, & y_i \geq \tau_{0.95}, \\ 1, & y_i < \tau_{0.95}, \end{cases} \tag{26}$$

where

$$\alpha \in \{0.25, 0.5, 1, 2, 4, 8\}. \tag{27}$$

Equation (26) therefore modifies model fitting according to source-label tail membership only; no 2014 or 2015 emission labels enter the weighting function.

Covariate-shift weights were calculated from the out-of-fold source-domain probabilities in Equation (19). Two formulations were evaluated. The density-ratio importance weight was

$$\widetilde{w}_i^{\text{imp}} = \frac{p_i}{1-p_i}\frac{n_s}{n_t}, \tag{28}$$

where the sample-size factor corrects for the source-to-target class ratio used by the domain classifier. A target-proximity formulation was also evaluated:

$$\widetilde{w}_i^{\text{prox}} = p_i. \tag{29}$$

The latter gives greater weight to source observations that resemble the target operating domain without using target-emission labels. The raw density ratios from Equation (28) and the raw proximity values from Equation (29) were retained before clipping so that the concentration of the unregularized weighting problem could be examined directly.

Every raw weight vector was transformed using a bounded mean-one projection. For upper bound c, the final weight was

$$w_i = \text{clip}\left(\lambda\widetilde{w}_i, \frac{1}{c}, c\right), \qquad \frac{1}{n_s}\sum_{i=1}^{n_s} w_i = 1, \tag{30}$$

where $\lambda > 0$ was obtained numerically. The pure TAIL configurations used $c = 10$, while the SHIFT and SHIFT–TAIL configurations evaluated

$$c \in \{3, 5, 10\}. \tag{31}$$

Equation (30) imposes both a lower bound of $1/c$ and an upper bound of c while preserving mean weight one.

For SHIFT–TAIL, the stabilized components were multiplied and projected again:

$$w_i^{\text{shift-tail}} = \Pi_c\left(w_i^{\text{shift}} w_i^{\text{tail}}\right), \tag{32}$$

where $\Pi_c(\cdot)$ denotes the bounded mean-one projection in Equation (30). This second projection ensured that the final product continued to satisfy the same normalization and clipping constraints.

Weight concentration was assessed using the effective sample size (ESS),

$$\text{ESS} = \frac{\left(\sum_{i=1}^{n_s} w_i\right)^2}{\sum_{i=1}^{n_s} w_i^2}, \tag{33}$$

which was also expressed as a percentage of n_s. ESS was calculated both for the normalized raw weight vector and after bounded projection. In addition to the minimum, median, 95th percentile, 99th percentile, maximum, mean, and standard deviation, the analysis recorded the fractions of raw weights lying below or above the eventual clipping bounds and the fractions of final weights located at those bounds. These quantities directly assessed whether an apparently acceptable post-clipping ESS masked severe concentration in the original density-ratio estimates.

The complete grid contained 49 configurations per pollutant: one BASE configuration, six TAIL configurations, six SHIFT configurations obtained from two shift formulations and three clipping bounds, and 36 SHIFT–TAIL combinations. Every configuration was evaluated on the 2014 selection period using the five prespecified seeds. Metrics were aggregated as means and standard deviations across seeds.

A configuration was considered Pareto-nondominated when no alternative achieved both a lower global RMSE and a lower Q_{99} RMSE, with at least one strict improvement. Final configurations were chosen using three illustrative global-error budgets. If RMSE_0 denotes the 2014 BASE value and $b \in \{0.02, 0.05, 0.10\}$, the feasible set is

$$\mathcal{F}_b = \{m : \mathrm{RMSE}_m \leq (1+b)\mathrm{RMSE}_0\}. \quad (34)$$

According to Equation (34), the budgets limit allowable global prediction error during configuration selection; they do not represent regulatory thresholds or guarantees of risk control in a future year. Within $\mathcal{F}_b$, the configuration with the lowest mean Q_{99} RMSE was selected. Ties were resolved using Q_{95} RMSE and then global RMSE. Q99 RMSE was retained as the prespecified primary objective because it penalizes large errors in the extreme-emission region while remaining directly comparable with the global-RMSE constraint. Because this symmetric criterion does not uniquely represent underprediction risk, Q99 MUL was additionally evaluated as an asymmetric sensitivity objective rather than assuming RMSE to be universally preferable. The 2%, 5%, and 10% values represent illustrative allowances for global RMSE deterioration and are not regulatory thresholds.

To evaluate whether the selected strategy depended on the definition of tail risk, the complete 2014 selection procedure was repeated using four primary objectives: Q_{90} RMSE, Q_{95} RMSE, Q_{99} RMSE, and Q_{99} MUL. The same global-RMSE feasibility budgets were retained. This sensitivity analysis separates changes caused by the tail definition from those caused by the choice between a symmetric squared-error objective and an asymmetric underprediction objective.

Because the 2014 extreme subsets were small and the supplied observations were ordered, configuration stability was evaluated using dependence-preserving moving-block reselection rather than observation-independent bootstrap alone. For each pollutant, observation-level 2014 predictions from all 49 configurations and each of the five seeds were retained. Each bootstrap replicate first sampled one of the five fitted-model seeds and then resampled the complete 2014 row sequence using moving blocks. Block lengths of 12, 24, 48, 96, 192, and 384 rows were evaluated in the definitive analysis; independent row resampling ($L = 1$) was retained only as a sensitivity comparator. A total of 2000 replicates were generated for every block length.

Within each replicate, global RMSE and the requested tail objective were recomputed for all configurations. Feasibility was re-established relative to that replicate's BASE RMSE according to Equation (34), after which the feasible configuration with the lowest bootstrap objective was selected. Replicates containing no observations from the requested tail were recorded as invalid for that objective rather than assigned an artificial performance value. Exact recovery was defined as the fraction of valid replicates in which the full-2014 selected configuration was selected again.

Only the configuration identity—variant, weighting formulation, α, and clipping bound—was transferred from 2014 to 2015. For SHIFT-based configurations, the source weights were recalculated from the unlabeled 2015 operational predictors using the locked formulation. The selected configuration was then fitted to all source observations and evaluated once on the 2015 labels. BASE and each selected configuration were repeated using the five prespecified seeds.

Cross-model confirmation examined whether the selected weighting effects depended on the regression architecture. For every pollutant and global-error budget, the exact configuration selected using 2014 was applied to the mean dummy regressor, Ridge, Random Forest, Extra Trees, HistGBR, and XGBoost. Each weighted model was compared with its own unweighted counterpart on the locked 2015 test set using the same five seeds.

Simple adaptation baselines were incorporated into the same chronological comparison rather than being evaluated only after the main selection. Two recency baselines trained the source-selected regression architecture using either 2013 alone or 2012–2013, while preserving the hyperparameters chosen exclusively within the original 2011–2013 development period. Both models were evaluated first on 2014 and subsequently on the locked

2015 period. Their 2014 global RMSE values were explicitly checked against the same 2%, 5%, and 10% feasibility budgets used for the weighting configurations. An extended 2014 selection analysis then admitted the original 49 configurations and the simple baselines to the same Q_{99}-RMSE selection framework.

The $q = 0.90$ and $q = 0.95$ quantile-regression baselines used histogram-based gradient boosting with quantile loss. Unlike the earlier fixed-parameter comparison, their hyperparameters were tuned using only the source domain. Twenty-four parameter combinations were sampled from grids for the number of boosting iterations, learning rate, maximum leaf nodes, minimum leaf size, and L_2 regularization. Five `TimeSeriesSplit` folds were used, and candidates were ranked first by mean pinball loss and then by mean RMSE. The selected quantile models were fitted to all source observations, evaluated in 2014 for budget feasibility, and then frozen for the 2015 comparison.

A direct calibration experiment was added to evaluate whether the large prospective bias, particularly for NO_x, could be corrected without using 2015 labels. For each seed, BASE and the configurations selected in 2014 generated predictions for the complete 2014 calibration period. Three mappings were then considered: no correction, a constant-offset correction, and linear calibration. For the offset correction,

$$\hat{y}^{\,\mathrm{cal}} = \hat{y} + \frac{1}{n_c} \sum_{i \in \mathcal{C}} (y_i - \hat{y}_i), \tag{35}$$

whereas linear calibration estimated

$$y_i = a + b\hat{y}_i + \varepsilon_i, \qquad i \in \mathcal{C}, \tag{36}$$

by ordinary least squares and applied the frozen mapping

$$\hat{y}^{\mathrm{cal}}_{2015} = a + b\hat{y}_{2015}. \tag{37}$$

Equations (35)–(37) use 2014 labels only; no 2015 emission measurements enter estimation of the calibration parameters. Calibration was evaluated for BASE, each primary locked configuration, and a tail-only comparator under the corresponding budget. Global and source-defined tail metrics were recalculated after calibration to determine whether improved overall calibration also improved high-emission prediction. A labeled update using a subset of the 2015 observations was not performed because this would have converted part of the locked test period into adaptation data and changed the prospective evaluation protocol. Instead, recent-history retraining and calibration based exclusively on the preceding 2014 labeled period were included as lower-complexity adaptation comparators while preserving the complete 2015 label set for final evaluation. A secondary classification experiment evaluated direct detection of source-defined Q_{90}, Q_{95}, and Q_{99} events. For each pollutant and threshold,

$$z_i = \mathbf{1}(y_i \geq \tau_q). \tag{38}$$

Equation (38) defines contemporaneous high-emission class membership; it does not define a future-event forecast. The candidate classifiers were balanced logistic regression, class-weighted Random Forest, and weighted XGBoost. Logistic regression used standardized predictors, balanced class weights, a regularization parameter $C = 1$, and a maximum of 3000 iterations. Random Forest used 250 trees, a maximum depth of 12, a minimum leaf size of 2, a predictor-sampling proportion of 0.8, and balanced class weights. XGBoost used 350 estimators, a maximum depth of 4, a learning rate of 0.04, row and column subsampling proportions of 0.8, and

$$\text{scale_pos_weight} = \frac{n_-}{n_+}. \tag{39}$$

Classifier selection used two forward-year validation experiments: training with 2011 and validating with 2012, followed by training with 2011–2012 and validating with 2013. The validation probabilities and labels from both experiments were concatenated. Models were ranked by PR–AUC and then recall. For each classifier, the decision threshold was selected from $\{0.01, 0.02, \ldots, 0.99\}$ by maximizing the F_2 score,

$$F_2 = \frac{5 \cdot \text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}}, \tag{40}$$

which assigns greater importance to recall. The selected threshold in Equation (40) was frozen before the 2015 evaluation. The selected classifier was refitted using all source observations and evaluated on the locked 2015 events. For reproducibility and false-alarm assessment, locked 2015 results were also retained for every candidate classifier, not only the selected classifier. Final classification metrics included ROC–AUC, PR–AUC, balanced accuracy, recall, specificity, precision, F_1, Matthews correlation coefficient, Brier score, prevalence, and the confusion matrix.

2.5. Statistical Inference, Interpretability, and Reproducibility

Uncertainty in the locked 2015 comparisons was assessed using a paired hierarchical seed–moving-block bootstrap. This procedure incorporated both model-initialization variability and dependence among adjacent rows. Predictions from BASE and each configuration selected in 2014 were retained for all five prespecified seeds.

For bootstrap replicate r, one seed was first sampled with replacement from

$$\mathcal{R} = \{17, 42, 137, 2026, 7777\}. \tag{41}$$

Equation (41) defines the fitted-model uncertainty propagated within the hierarchical bootstrap. Sampling one seed within each bootstrap replicate treats model initialization as a random model-fitting factor and propagates seed-to-seed variability together with row-level dependence. The five-seed means reported in the performance tables therefore serve as descriptive point estimates, whereas the bootstrap intervals characterize uncertainty for a randomly realized fitted model under the prespecified training procedure.

Moving blocks were then sampled from the complete 2015 row sequence until 7384 indices had been generated. Following the residual-dependence diagnostics and the reviewer request for longer-block sensitivity, block lengths were extended to

$$L \in \{12, 24, 48, 96, 192, 384\} \text{ rows}. \tag{42}$$

Because exact timestamps were unavailable, these values represent numbers of consecutive rows rather than verified physical durations. Autocorrelation functions were calculated from lag 1 through lag 384 for the mean BASE residual sequence. In addition, ACFs were calculated for the paired squared-loss differences and paired underprediction-loss differences between BASE and each selected configuration. These diagnostics were used to assess whether the block-length sensitivity range captured dependence in the quantities entering the paired comparison.

Tail masks were applied after resampling the complete sequence. This preserved the relative positions and temporal clustering of high-emission observations instead of resampling the tail observations as an isolated series. Bootstrap comparisons were performed for the complete test set and the source-defined Q_{90}, Q_{95}, and Q_{99} subsets. The evaluated statistics were MAE, RMSE, MUL, UR, CMU, MOL, OR, and CMO.

For metric M, the paired bootstrap difference was

$$\Delta_r = M_r^{\text{selected}} - M_r^{\text{BASE}}. \tag{43}$$

Therefore, negative differences favor the configuration selected in 2014 for error, underprediction, and overprediction-loss metrics. A total of 2000 bootstrap replicates were generated for every pollutant, budget, block length, evaluation subset, and metric. Replicates containing no observations from the requested tail were excluded for that tail metric and counted explicitly. Likewise, CMU and CMO were treated as undefined when a resample contained no underpredicted or overpredicted observations, respectively, instead of assigning zero. The number of valid and invalid replicates was retained for every analysis.

The 2.5th and 97.5th percentiles formed the 95% confidence interval. The two-sided empirical p-value used a finite-replicate correction:

$$p = \min\left[1,\, 2\min\left(\frac{1+\sum_{r=1}^{B}\mathbf{1}(\Delta_r \le 0)}{B+1}, \frac{1+\sum_{r=1}^{B}\mathbf{1}(\Delta_r \ge 0)}{B+1}\right)\right], \tag{44}$$

where $B = 2000$ in the definitive analysis. Since configuration selection used 2014 and inference used only the locked 2015 outcomes, the resulting comparisons remained independent of the selection labels. Bootstrap p-values were treated as complementary descriptive evidence rather than as a family of independent confirmatory tests. Interpretation emphasized effect magnitude, 95% bootstrap confidence intervals, and consistency across the prespecified block lengths; no multiplicity-adjusted significance threshold was applied.

Interpretability was evaluated for the unweighted source-trained Extra Trees models. For each pollutant, the selected model was refitted using all 2011–2013 observations with seed 17. Model-agnostic permutation importance was calculated on the locked 2015 test set using RMSE as the scoring criterion and 15 permutations per predictor. Positive importance indicates deterioration in prospective RMSE after permuting the corresponding predictor.

Tree SHAP values were calculated for a reproducible random sample of 1500 observations from the 2015 test set. Feature importance was summarized as the mean absolute SHAP value. SHAP and external permutation importance were interpreted separately: SHAP characterized the dependence encoded by the fitted model, whereas permutation importance quantified the contribution of each predictor to performance in the shifted target domain. Rather than reporting ranks alone, the definitive analysis retained importance magnitudes and uncertainty. Permutation importance was summarized by its mean, standard deviation, and empirical 95% interval across the 15 repeats. For Tree SHAP, uncertainty in mean absolute SHAP magnitude was estimated by bootstrap resampling of the 1500 evaluated observations. Rank agreement was used to identify variables that remained influential under both interpretations.

Additional reviewer-oriented diagnostic plots were generated from seed-averaged predictions for visualization only. These included measured-versus-predicted plots for 2015, residual sequences with source-defined tail observations indicated, and prediction errors stratified by source-defined quintile ranges of TEY, TIT, and GTEP. Statistical inference continued to use the seed-aware procedures described above rather than these seed-averaged visualization values.

The analyses were executed using the full 36,733-observation database in Python 3.10.11 on Windows 10. The definitive environment included NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, XGBoost 3.2.0, and SHAP 0.49.1. Matplotlib 3.10.8 used the noninteractive Agg backend to maintain compatibility with parallel model fitting.

The definitive v1.4 analysis used five temporal folds, five prespecified seeds, 24 randomized regression candidates per tunable model, 24 randomized candidates for each quantile-regression objective, 49 weighting configurations per pollutant, three global-error budgets, four configuration-selection objectives (Q_{90} RMSE, Q_{95} RMSE, Q_{99} RMSE,

and Q_{99} MUL), and three source-defined high-emission thresholds. The 2014 configuration-reselection analysis used 2000 bootstrap replicates for independent-row resampling and for moving-block lengths of 12, 24, 48, 96, 192, and 384 rows. Locked 2015 inference used the same six moving-block lengths and 2000 replicates per pollutant, budget, subset, metric, and block length.

The reproducibility package stored the input-file SHA-256 hash, execution arguments, software versions, source-only hyperparameters, fold-level predictions, complete 2014 configuration predictions, primary and sensitivity selections, simple-baseline results, calibration parameters and predictions, raw and clipped weight diagnostics, effective sample sizes, clipping-bound saturation, alternative domain-classifier results, covariate-balance statistics, support-stratified performance, tail-cluster diagnostics, locked 2015 observation-level predictions, complete classification probabilities and confusion matrices, bootstrap outputs, residual and paired-loss ACFs, interpretability uncertainty, and figure manifests. A persistent task-level checkpoint allowed interrupted runs to resume without repeating completed analyses.

The protocol was designed as a reliability assessment on a public multiyear benchmark. Consequently, the reported predictive results do not constitute PEMS certification, approval under a specific regulatory performance specification, or evidence that the fitted models can replace certified reference or continuous emission monitoring systems.

3. Results

3.1. Temporal Shift, Domain Separation, and Prospective Validation

The analysis included 36,733 complete observations: 22,191 from 2011–2013 for model development, 7158 from 2014 for configuration selection, and 7384 from 2015 for locked evaluation. The target distributions changed substantially across years (Figure 1). Mean CO was 2.214 mg/m^3 during 2011–2013, decreased to 2.081 mg/m^3 in 2014, and increased to 3.130 mg/m^3 in 2015. Mean NO_x decreased from 68.777 mg/m^3 during development to 60.067 mg/m^3 in 2014 and 59.891 mg/m^3 in 2015. The locked test year therefore combined a higher CO level with a sustained reduction in NO_x, representing two different forms of temporal target shift.

The annual predictor summaries in Table 1 showed concurrent changes in ambient and operating conditions. Mean ambient humidity decreased from 82.14 ± 13.00% in 2014 to 68.65 ± 13.54% in 2015, while mean AFDP decreased from 3.91 ± 0.71 to 3.60 ± 0.61 mbar. Mean GTEP increased from 25.74 ± 3.68 to 26.13 ± 4.47 mbar, and mean TIT declined from 1084.73 ± 16.13 °C in 2011 to 1078.97 ± 19.76 °C in 2015. Formal shift diagnostics were consistent with these descriptive changes. In 2014, the largest population stability index (PSI) values occurred for TIT (1.380), TEY (0.702), CDP (0.422), and AFDP (0.302). In 2015, TEY reached a PSI of 1.444, followed by AH (1.196), TIT (1.015), and AFDP (0.820).

Domain-classification sensitivity analysis confirmed strong source–target distinguishability, while also showing that its estimated magnitude depended on the validation protocol and probability treatment. Using the raw histogram-based gradient boosting (HGB) domain classifier with random stratified folds, mean ROC–AUC values were 0.99975 for 2014 and 0.99978 for 2015, with corresponding PR–AUC values of 0.99919 and 0.99937. When contiguous-block validation was used instead, the mean ROC–AUC decreased to 0.9567 for 2014 and 0.9829 for 2015. A logistic domain classifier produced blocked ROC–AUC values of 0.9837 and 0.9974, respectively. Probability calibration changed the probability-based ambiguity diagnostic but did not eliminate the strong discrimination between periods. Because this diagnostic depends on the classifier, probability calibration, and class priors, it is hereafter referred to as the domain-classifier ambiguity index rather than as a distribution-overlap coefficient. These results indicate severe temporal domain separation

while also showing that near-perfect discrimination under randomly mixed folds partly reflects local dependence among observations.

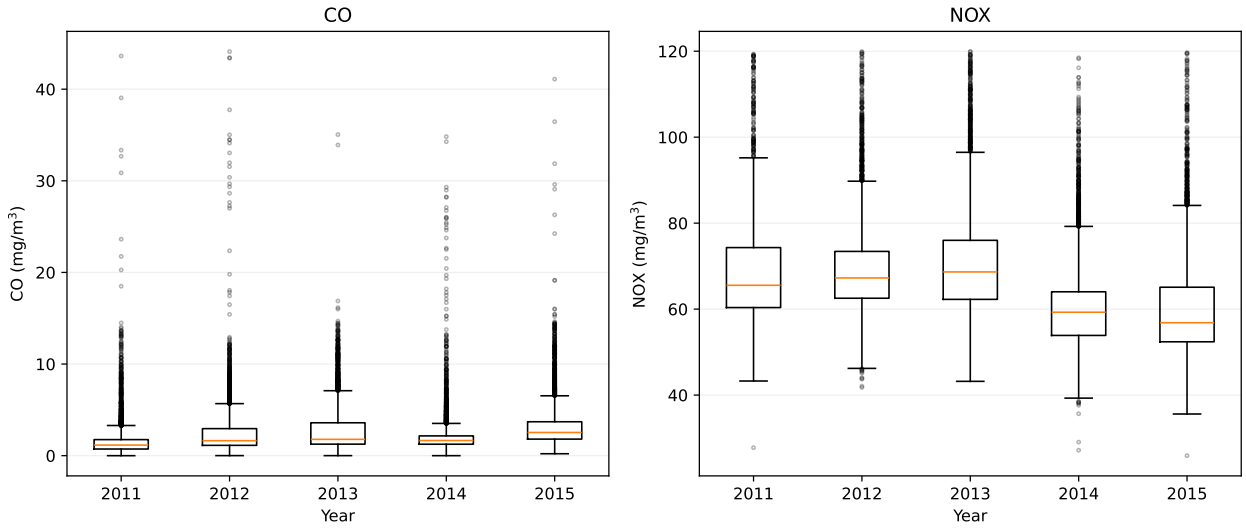


Figure 1. Annual distributions of CO and NO_x emissions. Boxes represent the interquartile range, horizontal lines indicate the median, and whiskers summarize the central distribution, while the individual extreme observations are retained because upper-tail behavior is a primary focus of the study. The 2011–2013 observations formed the development domain, 2014 was used for configuration selection, and 2015 was retained as the locked final test period. An upper-tail empirical distribution view is provided in the Supplementary Material.

Matched-size validation likewise confirmed that temporal ordering materially affected estimated predictive performance. Across all models, seeds, and folds, mean CO RMSE increased from 1.341 mg/m^3 under random within-prefix validation to 1.654 mg/m^3 under forward temporal validation. For NO_x, the corresponding values increased from 5.476 to 7.377 mg/m^3. Identical training and validation sizes were used within each matched pair, so the comparison isolates the effect of temporal ordering rather than differences in sample size or historical coverage (see Table 2 and Figure 2).

Extra Trees achieved the lowest forward temporal RMSE for both pollutants and remained the primary regression architecture for the subsequent adaptation experiments. For Extra Trees, random within-prefix RMSE was 1.054 versus 1.406 mg/m^3 under temporal validation for CO, corresponding to a 25.0% lower random estimate. For NO_x, the corresponding values were 3.838 and 5.818 mg/m^3, a 34.0% lower random estimate. The temporally validated R^2 values were 0.627 for CO and 0.643 for NO_x.

The unweighted Extra Trees model obtained an RMSE of 1.367 mg/m^3 and an R^2 of 0.536 in the 2014 selection period. On the locked 2015 test, CO RMSE was 1.484 mg/m^3, MAE was 0.869 mg/m^3, and R^2 was 0.559. The mean prediction error changed from a positive bias of 0.529 mg/m^3 in 2014 to a negative bias of -0.343 mg/m^3 in 2015, while the underprediction rate increased from 25.5% to 68.6%.

For NO_x, the 2014 baseline RMSE was 10.782 mg/m^3 with an R^2 of -0.169 and a positive bias of 8.529 mg/m^3. In the locked 2015 test, RMSE increased to 13.246 mg/m^3, R^2 decreased to -0.416, and mean bias increased to 11.968 mg/m^3. Thus, the 2015 NO_x baseline was strongly miscalibrated in absolute concentration despite preservation of the source-only model-selection protocol.

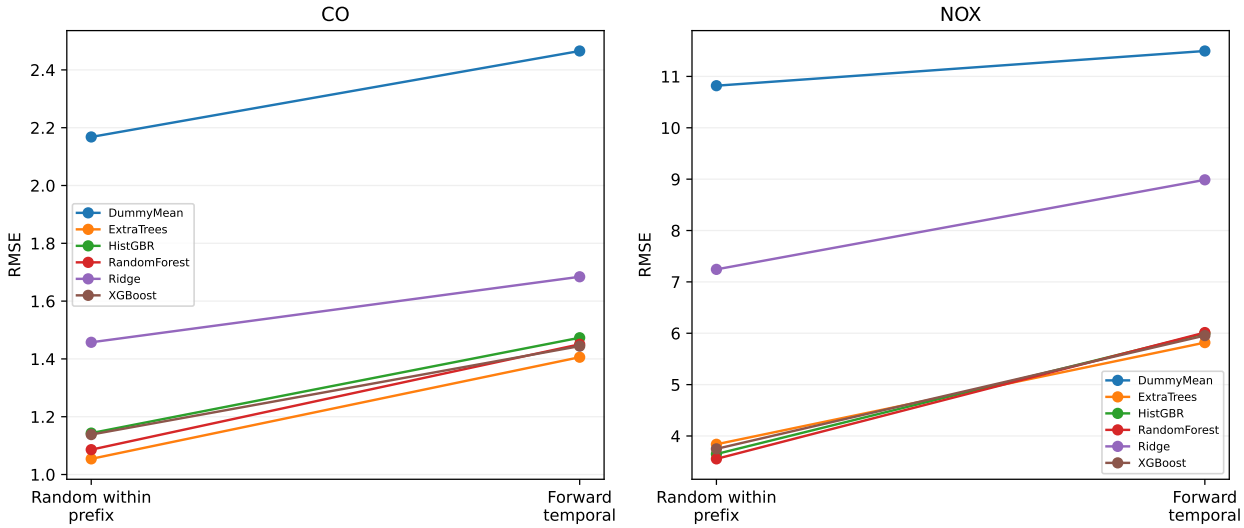


Figure 2. Model -specific matched-size comparison between random within-prefix validation and forward temporal validation in the 2011–2013 development domain. Paired results are shown separately for each regression architecture rather than pooled across models. Training and validation sample sizes were identical within each matched comparison.

Table 2. Mean matched-size cross-validation performance within the 2011–2013 development domain.

Target	Model	Random RMSE	Temporal RMSE	Temporal R^2
CO	Dummy mean	2.168	2.465	−0.078
CO	Ridge	1.457	1.684	0.484
CO	Random Forest	1.086	1.451	0.606
CO	Extra Trees	1.054	1.406	0.627
CO	HistGBR	1.143	1.473	0.595
CO	XGBoost	1.138	1.444	0.608
NO_x	Dummy mean	10.819	11.495	−0.217
NO_x	Ridge	7.244	8.987	0.217
NO_x	Random Forest	3.554	6.013	0.620
NO_x	Extra Trees	3.838	5.818	0.643
NO_x	HistGBR	3.649	5.997	0.624
NO_x	XGBoost	3.749	5.955	0.628

Residual diagnostics provided additional evidence of year-dependent predictive behavior. Comparing 2014 and 2015, the signed-residual distributions differed for both CO and NO_x according to Welch and Kolmogorov–Smirnov tests ($p < 0.001$), and the absolute-error distributions also differed ($p < 0.001$). Residual variance differed significantly for CO (Levene $p < 0.001$), whereas the corresponding NO_x variance test was not significant at the 0.05 level ($p = 0.054$). Conditional residual analyses across quantile strata of the nine predictors and the baseline prediction showed nominal year-to-year residual differences in all 50 evaluated CO strata and in 48 of 50 NO_x strata. These analyses document substantial year-dependent predictive miscalibration; they do not, by themselves, identify whether the underlying mechanism is covariate shift, conditional shift, or model misspecification. Measured-versus-predicted plots, residual-sequence plots, and operating-stratum analyses are provided in the Supplementary Materials.

3.2. Tail-Risk Selection, Objective Sensitivity, and Source-Domain Support

High-emission subsets were defined using thresholds estimated exclusively from the 2011–2013 source observations. The CO thresholds were 4.815, 6.168, and 10.717 mg/m^3 for Q_{90}, Q_{95}, and Q_{99}, respectively; the corresponding NO_x thresholds were 82.133, 87.133,

and 107.452 mg/m^3. In 2014, the CO Q_{90}, Q_{95}, and Q_{99} subsets contained 354, 173, and 56 observations. In 2015, their counts increased to 987, 528, and 123, corresponding to prevalences of 13.37%, 7.15%, and 1.67%. For NO_x, the 2014 Q_{90}, Q_{95}, and Q_{99} subsets contained 256, 150, and only 15 observations, respectively, while the locked 2015 subsets contained 350, 210, and 41 observations, corresponding to prevalences of 4.74%, 2.84%, and 0.56%.

For the CO baseline, locked 2015 RMSE increased from 1.484 mg/m^3 globally to 2.948 at Q_{90}, 3.688 at Q_{95}, and 6.790 mg/m^3 at Q_{99}. At Q_{99}, mean underprediction loss was 2.686 mg/m^3, the underprediction rate was 73.8%, and the conditional magnitude among underestimated observations was 3.640 mg/m^3.

For NO_x, the large positive global bias resulted in a global underprediction rate of only 1.20%, but this behavior reversed in the extreme tail. At Q_{99}, 82.4% of the 41 high-emission observations were underpredicted, with a mean underprediction loss of 8.170 mg/m^3 and a conditional underprediction magnitude of 9.922 mg/m^3. Thus, systematic global overprediction coexisted with frequent underprediction of the highest observed NO_x concentrations.

A total of 49 BASE, TAIL, SHIFT, and SHIFT–TAIL configurations were evaluated for each pollutant on the 2014 selection period using five prespecified random seeds. Selection minimized source-defined Q_{99} RMSE subject to allowable global-RMSE increases of at most 2%, 5%, or 10% relative to BASE. These percentages represent prespecified selection allowances rather than regulatory thresholds. The selected configurations were frozen before the 2015 emission labels were evaluated.

For CO, TAIL weighting with $\alpha = 8$ minimized the prespecified Q_{99} RMSE under all three global-error budgets. Its 2014 global RMSE was 1.347 mg/m^3 compared with 1.367 mg/m^3 for BASE, while Q_{99} RMSE decreased from 7.602 to 7.338 mg/m^3. For NO_x, increasing risk budgets selected progressively stronger tail weighting: TAIL $\alpha = 1$ under the 2% budget, TAIL $\alpha = 4$ under the 5% budget, and TAIL $\alpha = 8$ under the 10% budget. Their 2014 Q_{99} RMSE values were 20.169, 19.870, and 19.512 mg/m^3, respectively, compared with 20.431 mg/m^3 for BASE. No SHIFT or SHIFT–TAIL configuration minimized the primary Q_{99}-RMSE criterion, as shown in Figure 3.

The preferred configuration depended on the risk objective. For CO, Q_{90}-, Q_{95}-, and Q_{99}-RMSE selection all chose TAIL $\alpha = 8$ under the three budgets, whereas using Q_{99} mean underprediction loss (MUL) as the primary asymmetric objective selected TAIL $\alpha = 4$. For NO_x, the Q_{90}-RMSE objective selected SHIFT importance weighting with a clipping threshold of 5, and the Q_{95}-RMSE objective selected SHIFT–TAIL weighting with $\alpha = 4$ and clipping threshold 3. In contrast, both Q_{99} RMSE and Q_{99} MUL selected TAIL $\alpha = 1$, 4, and 8 under the 2%, 5%, and 10% budgets, respectively. Thus, the selected adaptation mechanism was not invariant to the tail definition or loss function.

Moving-block reselection of the 2014 data further showed that exact configuration stability was limited when temporal dependence was preserved. The analysis used 2000 replicates and block lengths of 12, 24, 48, 96, 192, and 384 rows, in addition to observation-independent resampling as a sensitivity analysis. Under the primary Q_{99}-RMSE objective and a block length of 384 rows, exact recovery of CO TAIL $\alpha = 8$ was 31.0%, 28.8%, and 28.8% under the 2%, 5%, and 10% budgets, respectively. For NO_x, recovery frequencies were 7.0% for TAIL $\alpha = 1$, 46.1% for TAIL $\alpha = 4$, and 81.1% for TAIL $\alpha = 8$. Observation-independent resampling yielded higher NO_x recovery frequencies, particularly for the 10% configuration, demonstrating the importance of accounting for dependence during selection-stability assessment, as shown in Table 3.

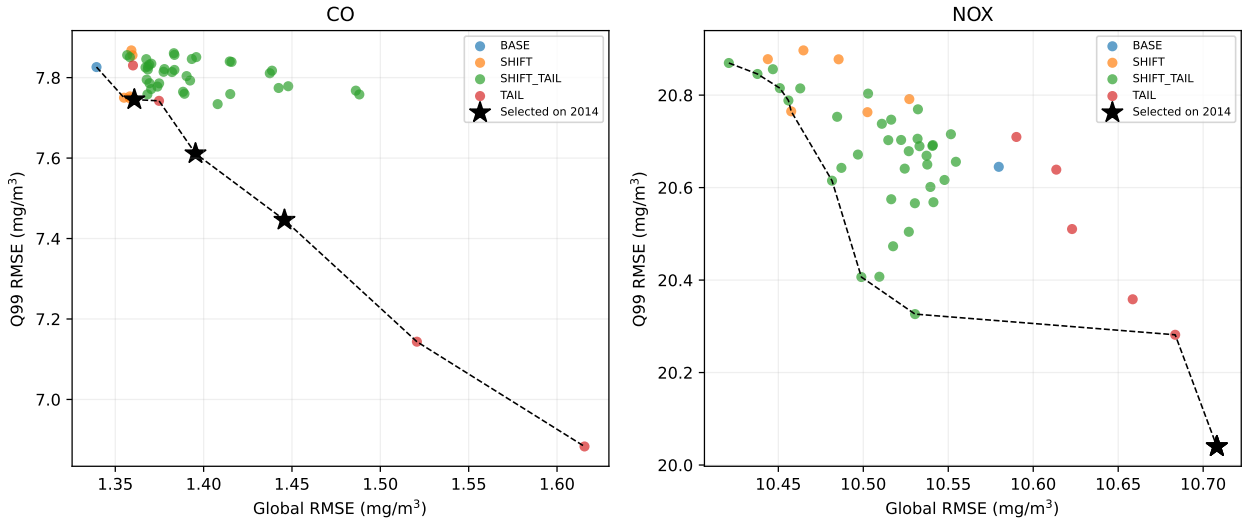


Figure 3. Configuration selection using only the 2014 period. Each point represents mean global and source-defined Q_{99} RMSE across five seeds. Dashed lines identify empirical Pareto fronts, and black stars indicate configurations selected under the 2%, 5%, and 10% allowable global-RMSE increases. The three CO stars overlap because TAIL weighting with $\alpha = 8$ was selected under all three budgets.

Table 3. Exact recovery of the primary 2014 Q_{99}-RMSE selections. IID denotes observation-independent bootstrap resampling; block-384 denotes moving-block resampling with 384-row blocks.

Target	Budget	Selected Configuration	IID Recovery (%)	Block-384 Recovery (%)
CO	2%	TAIL, $\alpha = 8$	34.65	30.95
CO	5%	TAIL, $\alpha = 8$	33.05	28.80
CO	10%	TAIL, $\alpha = 8$	33.05	28.75
NO_x	2%	TAIL, $\alpha = 1$	24.35	6.98
NO_x	5%	TAIL, $\alpha = 4$	67.75	46.08
NO_x	10%	TAIL, $\alpha = 8$	96.70	81.06

The small NO_x Q_{99} samples were also examined along the supplied record sequence. The 15 Q_{99} observations in 2014 occurred in nine contiguous row clusters, with a maximum cluster length of four rows; the 41 observations in 2015 occurred in 24 clusters, with a maximum length of five rows. For CO, the corresponding Q_{99} counts of 56 and 123 were distributed across 30 and 60 contiguous clusters. Because individual timestamps are unavailable, these clusters describe concentration along the supplied record order and are not interpreted as independent physical emission events.

Importance-weight diagnostics confirmed that density-ratio estimation was poorly conditioned under the severe domain separation. Across the alternative domain classifiers, mean raw effective sample size (ESS) for 2015 ranged from approximately 0.57% to 4.42% of the source sample, with a minimum seed-level ESS of 0.0068%; normalized raw weights reached values above 18,000 in the most extreme case. Clipping increased ESS substantially, to approximately 13.0–58.8% depending on the classifier and clipping threshold, but between 54.3% and 94.6% of weights could be affected by the lower bound and up to 4.0% by the upper bound. Mean absolute standardized covariate difference was 0.263 before weighting and reached values as low as approximately 0.184 after weighting. Prediction performance also varied with the density-ratio estimator and clipping choice. These results confirm that clipping improves numerical stability and balance but does not eliminate the underlying common-support limitation. Complete predictor-level covariate-balance results and locked 2015 performance stratified by empirical source-domain support are provided in the reproducibility materials.

3.3. Locked 2015 Performance, Simple Adaptation Baselines, and Calibration

Table 4 and Figure 4 show the independent 2015 performance of the configurations selected with the primary 2014 Q_{99}-RMSE criterion. For CO, the same TAIL $\alpha = 8$ configuration was selected under all three budgets. Global RMSE decreased modestly from 1.484 to 1.472 mg/m^3 (−0.87%), while Q_{99} RMSE changed from 6.790 to 6.784 mg/m^3 (−0.10%) and Q_{99} MUL decreased from 2.686 to 2.632 mg/m^3 (−2.0%). The Q_{99} underprediction rate changed only from 73.8% to 73.5%. Global mean overprediction loss decreased from 0.263 to 0.252 mg/m^3, whereas Q_{99} mean overprediction loss increased from 0.392 to 0.417 mg/m^3.

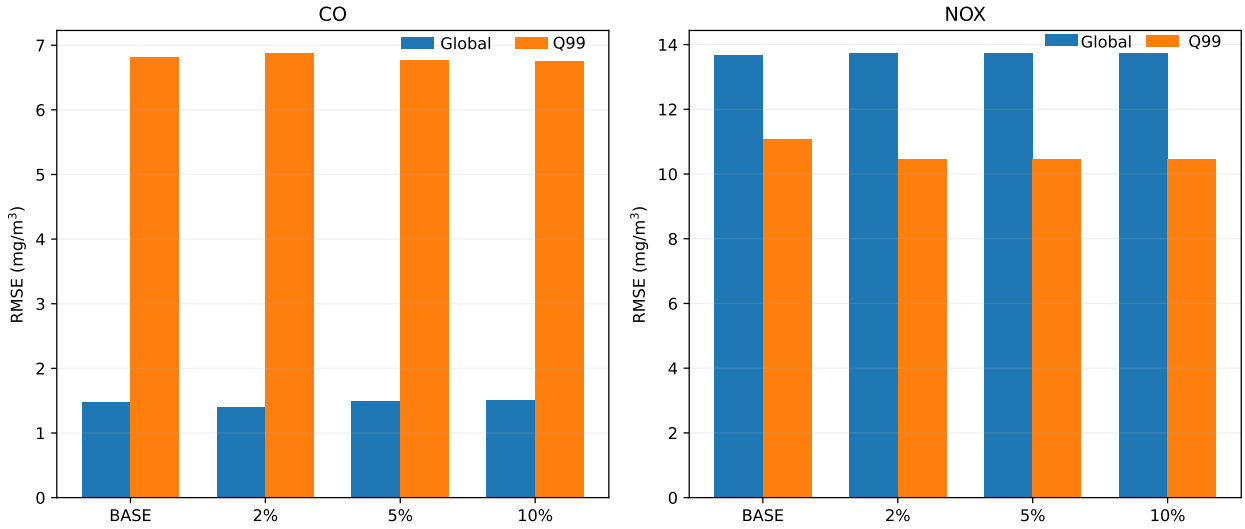


Figure 4. Global and source-defined Q_{99} RMSE on the locked 2015 test set. BASE denotes the unweighted Extra Trees model. The budget labels identify configurations selected using only 2014 under the corresponding maximum allowable increase in global RMSE. Duplicate selections are represented by the same configuration.

Table 4. Locked 2015 performance of configurations selected using 2014. Percentage changes are relative to the corresponding unweighted Extra Trees baseline. Negative values represent reductions. MUL denotes mean underprediction loss.

Target	Locked Configuration	Budget	Global RMSE	Δ	Q_{95} MUL	Δ	Q_{99} RMSE	Δ	Q_{99} MUL	Δ
				(%)		(%)		(%)		(%)
CO	BASE	—	1.484	0.0	1.082	0.0	6.790	0.0	2.686	0.0
CO	TAIL, $\alpha = 8$	2%/5%/10%	1.472	−0.9	1.059	−2.1	6.784	−0.1	2.632	−2.0
NO_x	BASE	—	13.246	0.0	2.151	0.0	11.054	0.0	8.170	0.0
NO_x	TAIL, $\alpha = 1$	2%	13.462	+1.6	2.071	−3.7	10.794	−2.3	7.952	−2.7
NO_x	TAIL, $\alpha = 4$	5%	13.774	+4.0	2.000	−7.1	10.538	−4.7	7.740	−5.3
NO_x	TAIL, $\alpha = 8$	10%	14.018	+5.8	1.973	−8.3	10.444	−5.5	7.662	−6.2

For NO_x, stronger tail weighting produced progressively larger improvements in the extreme tail together with increasing global error. Under the 2%, 5%, and 10% configurations, global RMSE increased by 1.63%, 3.99%, and 5.83%, respectively, while Q_{99} RMSE decreased by 2.35%, 4.66%, and 5.51%. Corresponding Q_{99} MUL reductions were 2.67%, 5.26%, and 6.22%. Global mean overprediction loss increased from 12.039 mg/m^3 for BASE to 12.268, 12.582, and 12.817 mg/m^3, respectively. The reduction in high-emission underprediction was therefore accompanied by a measurable increase in overprediction elsewhere in the response distribution.

The underprediction frequency and conditional severity of the unique locked 2015 configurations are shown in Figure 5.

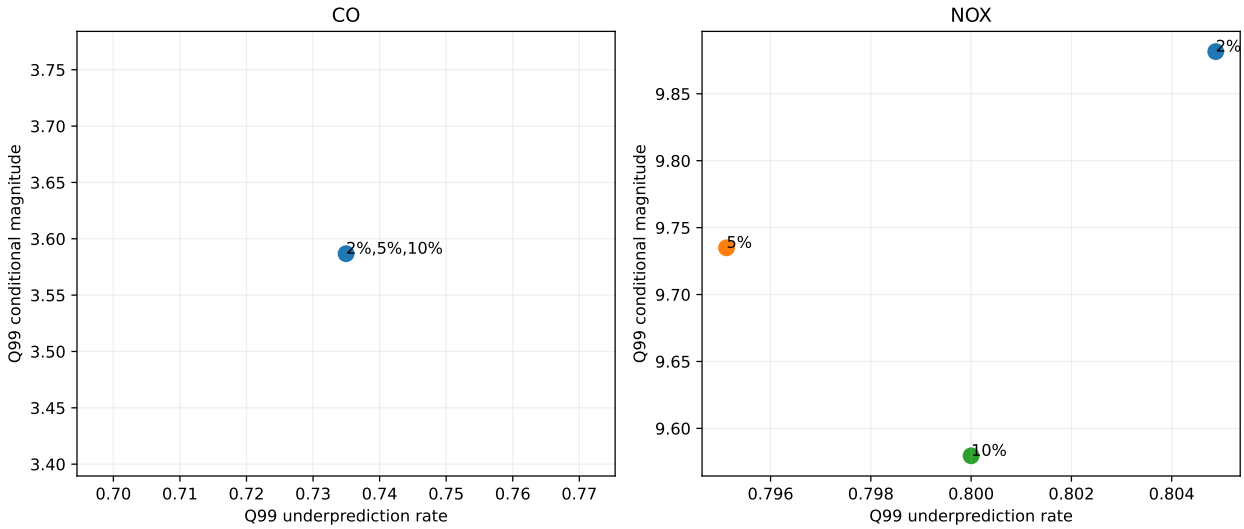


Figure 5. Underprediction frequency and conditional severity for the unique locked 2015 configurations. Repeated budget selections of the same fitted configuration are combined to avoid presenting duplicate models as independent validations.

Simple temporal baselines provided an important comparator to weighting-based adaptation. For CO, training Extra Trees only on 2013 produced the lowest global 2015 RMSE among the simple baselines (1.389 mg/m^3), but this model was feasible in 2014 only under the 10% global-error allowance and yielded a Q_{99} MUL of 3.008 mg/m^3. Training on 2012–2013 produced a 2014 RMSE of 1.433 mg/m^3 and was feasible under the 5% and 10% budgets. In 2015 it achieved a global RMSE of 1.467 mg/m^3, a Q_{99} RMSE of 6.630 mg/m^3, and a Q_{99} MUL of 2.255 mg/m^3. When this baseline was admitted to the same extended 2014 selection framework, RECENT-2012–2013 was selected for CO under the 5% and 10% budgets, whereas TAIL $\alpha = 8$ remained selected under the 2% budget.

For NO_x, the recent-2013 and recent-2012–2013 models were feasible under the global-error budgets in 2014 but produced worse Q_{99} performance in 2015 than the selected TAIL configurations. Quantile-regression baselines achieved substantially smaller Q_{99} errors but were not feasible under any of the prespecified 2%, 5%, or 10% global-error allowances because of their considerably larger 2014 global RMSE. The comparative results therefore show that simpler recency adaptation is competitive for CO, whereas the global-versus-tail trade-off is materially different for NO_x, as shown in Table 5.

Direct calibration experiments further separated global predictive calibration from extreme-tail performance. Calibration parameters were estimated only from the 2014 labels and were then frozen before application to 2015. For NO_x, a constant-offset correction reduced 2015 global RMSE from 13.246 to 6.637 mg/m^3, while linear calibration reduced it to 6.333 mg/m^3 and increased R^2 from -0.416 to 0.676. However, the same linear correction increased Q_{99} RMSE from 11.054 to 24.463 mg/m^3 and increased Q_{99} MUL from 8.170 to 23.559 mg/m^3; all 41 Q_{99} observations were underpredicted after calibration. Combining linear calibration with the selected TAIL configurations retained the strong global-RMSE improvement but further increased Q_{99} error. By contrast, neither constant-offset nor linear calibration improved the CO baseline globally. The NO_x calibration results are summarized in Table 6.

Table 5. Simple temporal and quantile-regression baselines. "Feasible" indicates the global-RMSE budgets satisfied on the 2014 selection period.

Target	Baseline	2014 RMSE	Feasible Budgets	2015 RMSE	2015 Q_{99} RMSE	2015 Q_{99} MUL
CO	Multiyear BASE	1.367	reference	1.484	6.790	2.686
CO	Recent 2013	1.477	10%	1.389	6.883	3.008
CO	Recent 2012–2013	1.433	5%, 10%	1.467	6.630	2.255
CO	Quantile Q_{90}	1.908	none	1.558	6.605	2.593
CO	Quantile Q_{95}	2.510	none	1.994	6.506	2.242
NO_x	Multiyear BASE	10.782	reference	13.246	11.054	8.170
NO_x	Recent 2013	10.280	2%, 5%, 10%	14.223	13.392	10.196
NO_x	Recent 2012–2013	10.956	2%, 5%, 10%	14.304	11.318	8.989
NO_x	Quantile Q_{90}	16.970	none	19.946	7.159	3.138
NO_x	Quantile Q_{95}	18.758	none	22.078	7.134	1.646

Table 6. NO_x calibration fitted using only 2014 labels and frozen for the locked 2015 evaluation. Linear-calibration combinations with the three selected TAIL configurations are included to show the global-versus-tail trade-off.

Configuration	Calibration	Global RMSE	R^2	Bias	Q_{99} RMSE	Q_{99} MUL
BASE	None	13.246	−0.416	11.968	11.054	8.170
BASE	Offset	6.637	0.644	3.439	18.199	16.519
BASE	Linear	6.333	0.676	2.814	24.463	23.559
TAIL, $\alpha = 1$	Linear	6.330	—	—	24.846	23.997
TAIL, $\alpha = 4$	Linear	6.328	—	—	25.700	24.939
TAIL, $\alpha = 8$	Linear	6.347	—	—	26.283	25.569

3.4. Statistical Robustness, Cross-Model Confirmation, and Secondary Analyses

Paired seed–moving-block bootstrap inference used 2000 replicates and block lengths of 12, 24, 48, 96, 192, and 384 rows. Resamples containing no observations in the requested tail, or no observations required for a conditional metric, were treated as undefined for that metric rather than assigned a zero value. Figure 6 shows Q_{99} RMSE differences at a representative block length of 192 rows.

For CO, the small Q_{99} RMSE and MUL differences for TAIL $\alpha = 8$ were not statistically resolved at any examined block length. At 192 rows, the Q_{99} RMSE difference was -0.007 mg/m^3 (95% CI, $[-0.121, 0.080]$), and the Q_{99} MUL difference was -0.057 mg/m^3 (95% CI, $[-0.191, 0.099]$). Thus, the point estimates favored TAIL only slightly and remained compatible with no tail improvement after accounting for dependence.

For NO_x, statistical support depended on the allowable global-error budget. The 2% configuration did not show a statistically resolved Q_{99} RMSE improvement at any block length. The 5% configuration became significant for block lengths of 96 rows and longer; at 192 rows its mean Q_{99} RMSE difference was -0.460 mg/m^3 (95% CI, $[-1.017, -0.038]$; $p = 0.034$). The 10% configuration produced negative Q_{99} RMSE confidence intervals at every examined block length; at 192 rows the difference was -0.555 mg/m^3 (95% CI, $[-1.273, -0.131]$; $p = 0.018$). In contrast, the Q_{99} MUL differences remained statistically unresolved for all three NO_x budgets.

Residual and paired-loss autocorrelation supported the expanded block-length sensitivity analysis. BASE residual autocorrelation at lag 1 was 0.436 for CO and 0.789 for NO_x. At lag 96 it was 0.107 and 0.470, respectively, and at lag 384 it decreased to 0.005 for CO but remained 0.337 for NO_x. For the NO_x 5% comparison, squared-loss-difference autocorrelation decreased from approximately 0.743 at lag 1 to 0.102 at lag 384. The longer-block analyses therefore remained particularly relevant for NO_x.

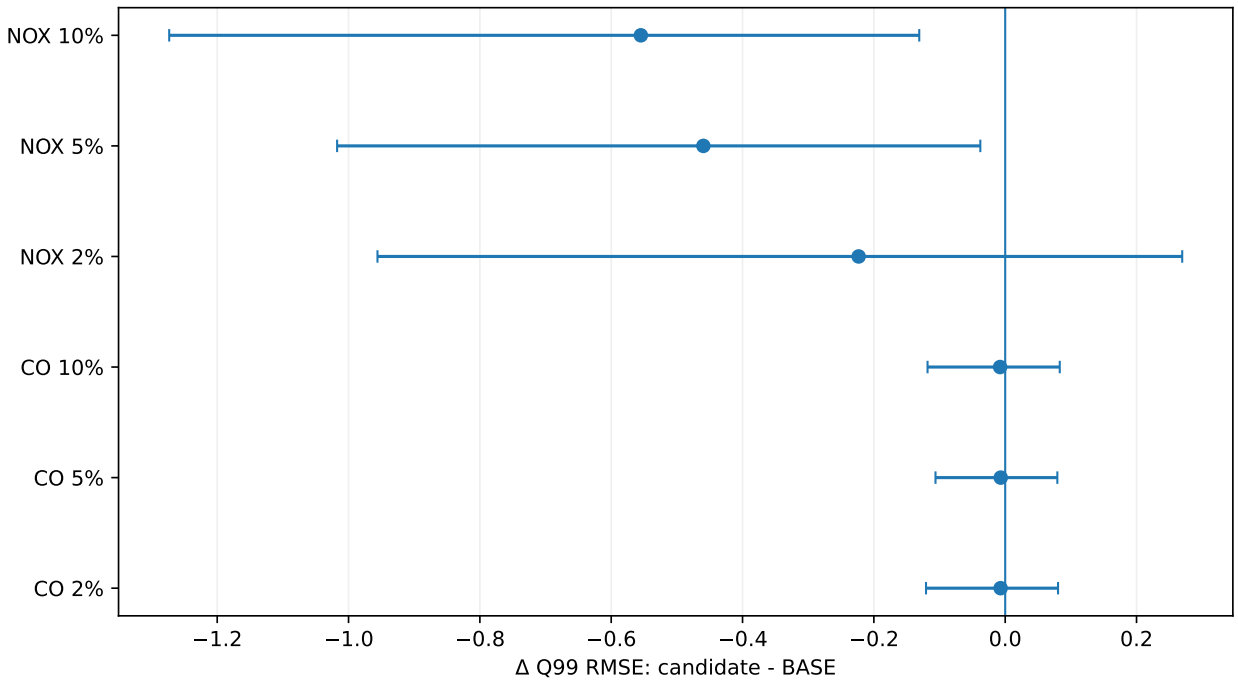


Figure 6. Paired seed–moving-block bootstrap differences in source-defined Q_{99} RMSE on the locked 2015 test set using a block length of 192 rows and 2000 replicates. Points represent mean differences between the configuration selected in 2014 and BASE; horizontal lines represent 95% confidence intervals. Negative values favor the selected configuration.

Cross-model confirmation showed that the effect of weighting depended on regression architecture. For CO, TAIL $\alpha = 8$ reduced Q_{99} MUL across the evaluated regressors, although changes in Q_{99} RMSE and global RMSE were not uniformly favorable. For example, Random Forest and Ridge showed lower Q_{99} underprediction loss but higher Q_{99} RMSE after tail weighting. For NO_x, the strongest TAIL configuration ($\alpha = 8$) reduced Q_{99} RMSE across all confirmatory regression architectures, while global error generally increased. These results support a recurring tail-versus-global trade-off rather than uniform superiority of a single weighting strategy.

The secondary high-emission classification experiments quantified the associated event-detection and false-alarm burden. On the locked 2015 test, selected CO classifiers achieved ROC–AUC values of 0.943, 0.868, and 0.896 for Q_{90}, Q_{95}, and Q_{99}, respectively. At Q_{99}, recall was 52.0% and precision was 88.9%, corresponding to 64 true positives, 8 false positives, and 59 false negatives. For NO_x, ROC–AUC values were 0.958, 0.978, and 0.951. The Q_{99} classifier detected 39 of the 41 source-defined high-emission observations (95.1% recall), but generated 398 false positives, yielding a precision of 8.9%. Its PR–AUC was 0.103 compared with a positive-class prevalence of only 0.56%. Complete confusion matrices and decision thresholds are provided in the Supplementary Material.

Predictor-importance magnitudes and uncertainty are summarized in Figure 7. For CO, external permutation importance ranked TIT first, followed by TAT, GTEP, and CDP; mean absolute Tree-SHAP magnitude likewise ranked TIT and TAT first and second, followed by TEY. For NO_x, ambient temperature was the dominant predictor under both methods. Permutation importance next ranked TAT, AFDP, and TIT, whereas Tree SHAP ranked TIT, AH, and TAT after ambient temperature. Thus, agreement was strongest for the leading variables, whereas lower-order contributions remained attribution-method-dependent.

Overall, the expanded analyses confirmed three distinct features of prospective PEMS evaluation. First, randomly mixed validation substantially underestimated future-period error relative to matched forward temporal validation. Second, severe source–target sepa-

ration made unconstrained importance weighting poorly conditioned, and simple recency adaptation was competitive for CO under moderate global-error allowances. Third, global calibration and extreme-tail performance represented different objectives: NO_x calibration substantially improved global RMSE while markedly worsening Q_{99} underprediction, whereas stronger tail weighting improved Q_{99} RMSE at the cost of global error. The moving-block analyses further showed that statistical support and configuration stability weakened when temporal dependence and the small effective number of extreme observations were explicitly considered.

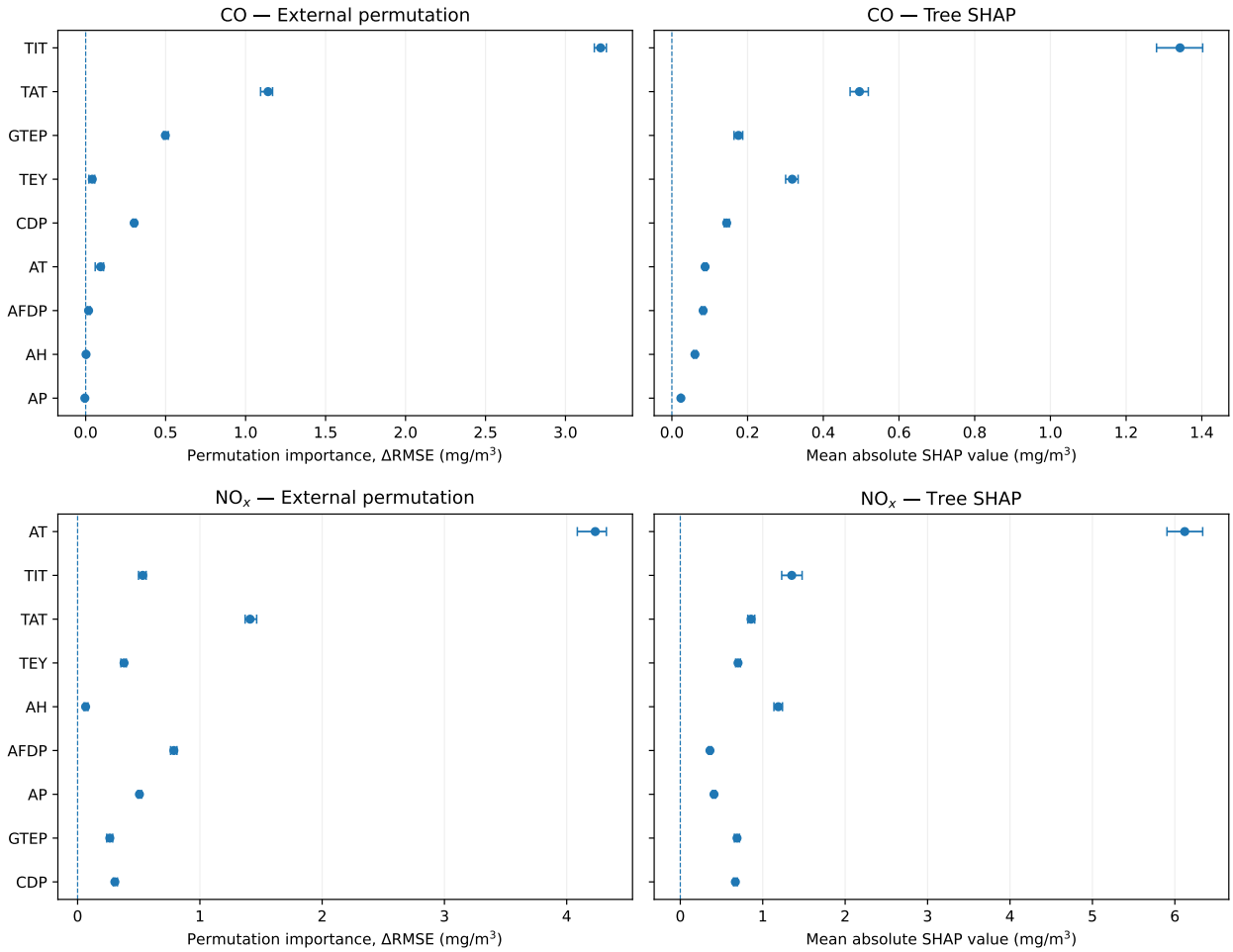


Figure 7. Predictor-importance magnitudes and uncertainty for the locked 2015 evaluation. External permutation importance is expressed as the increase in RMSE after predictor permutation, whereas Tree SHAP importance is expressed as the mean absolute SHAP value. Points represent importance magnitudes and horizontal bars represent 95% uncertainty intervals. Results are shown separately for CO and NO_x.

4. Discussion

4.1. Temporal Validation, Distribution Shift, and Empirical Support

The results show that validation design materially affects the apparent reliability of gas-turbine PEMS. For the selected Extra Trees models, matched random within-prefix validation produced RMSE estimates approximately 25.0% lower for CO and 34.0% lower for NO_x than matched forward temporal validation. Because the paired protocols used the same historical prefixes and identical train–validation sample sizes, the difference is primarily attributable to temporal mixing rather than sample-size effects. This finding is consistent with recommendations that validation partitions for structured data should

reflect the intended deployment setting [11], and complements prior PEMS studies focused mainly on predictive accuracy under conventional validation [2–6,9].

Prospective degradation occurred together with substantial changes in both emissions and operating variables. Domain classifiers strongly separated the source period from 2014 and 2015, although discrimination decreased when contiguous-block rather than randomly stratified validation was used. Residual distributions also changed between 2014 and 2015 across both pollutants and many operating-variable strata. These results establish year-dependent predictive miscalibration and limited empirical support, but they do not uniquely identify the underlying mechanism: covariate shift, conditional change, model misspecification, and support limitations can produce overlapping signatures. Covariate shift should therefore be interpreted here as one candidate adaptation setting rather than as a verified description of the temporal change [12].

The importance-weight diagnostics reinforce this limitation. Across the alternative domain classifiers, the mean raw effective sample size for the 2015 comparison represented only approximately 0.57–4.42% of the source sample, and some normalized raw weights exceeded 18,000. Clipping improved numerical stability and covariate balance, but a substantial fraction of observations remained affected by the clipping bounds. Alternative classifiers and probability calibration changed the estimated weights without removing the common-support problem. Thus, clipping and ESS monitoring can regularize importance weighting, but they cannot reconstruct target operating conditions that are absent or poorly represented in the source domain [13–16].

4.2. Pollutant-Specific Adaptation and Tail-Risk Trade-Offs

For CO, the primary Q_{99}-RMSE criterion selected TAIL weighting with $\alpha = 8$ under all three global-error allowances. On the locked 2015 period, however, the resulting changes were small: global RMSE decreased by approximately 0.9%, Q_{99} RMSE by approximately 0.1%, and Q_{99} mean underprediction loss by approximately 2.0%. None of the Q_{99} differences was statistically resolved in the moving-block bootstrap analyses. Configuration identity was also unstable under dependence-preserving 2014 reselection, indicating that several neighboring configurations provided similar empirical trade-offs. A simpler recent-history model trained on 2012–2013 was feasible under the 5% and 10% allowances and was preferred when admitted to the same extended selection framework. These results suggest that additional adaptation complexity does not necessarily improve prospective CO performance.

NO_x exhibited a different problem: the locked baseline was strongly miscalibrated, with substantial positive bias and poor global fit. Historical-label calibration corrected much of this global error without using 2015 labels; linear calibration reduced global RMSE from 13.246 to 6.333 mg/m^3 and increased R^2 from -0.416 to 0.676. However, the same correction increased Q_{99} RMSE from 11.054 to 24.463 mg/m^3 and worsened high-emission underprediction. Conversely, progressively stronger TAIL weighting reduced Q_{99} RMSE by approximately 2.4–5.5% while increasing global RMSE by approximately 1.6–5.8% and increasing overprediction elsewhere. Bootstrap support was budget-dependent: the 2% configuration was not statistically resolved, the 5% configuration was supported only at longer block lengths, and the 10% configuration showed a consistent reduction in Q_{99} RMSE across the examined dependence scales. The corresponding Q_{99} mean-underprediction-loss differences nevertheless remained statistically uncertain. Global calibration and high-emission performance are therefore distinct objectives rather than interchangeable measures of reliability.

Sparse tails further constrain the strength of configuration-level conclusions. Only 15 NO_x observations exceeded the source-defined Q_{99} threshold in 2014, and exact moving-block recovery of the selected configuration ranged from approximately 7% to 81% across the three global-error allowances. Sensitivity analyses also showed that the preferred strategy changed with the tail definition and loss function: broader Q_{90} and Q_{95} objectives could favor SHIFT or SHIFT–TAIL, whereas the primary Q_{99} criterion favored pure TAIL weighting. No SHIFT-based configuration was selected under the primary Q_{99} objective for either pollutant. Together with the severe support limitations, this result argues against treating any single adaptation mechanism as universally preferable.

4.3. Operational Interpretation

The asymmetric analyses emphasize why global bias alone is insufficient for deployment-oriented assessment. The baseline NO_x model overpredicted the 2015 distribution on average while still underpredicting most actual Q_{99} observations. Tail-oriented weighting reduced some high-emission errors but increased overprediction elsewhere, so a statistically favorable tail metric should not be interpreted automatically as improved operational reliability without specifying the relative costs of missed high-emission conditions and false alarms [17,18]. The supplementary classification experiments illustrate the same issue: the NO_x Q_{99} classifier achieved high recall but low precision because of the very low prevalence of extreme observations. Because the dataset contains no verified forecast lead time, these results represent contemporaneous high-emission identification rather than early-warning prediction.

Interpretability results were comparatively stable for the dominant predictors. TIT was the leading CO variable under both external permutation importance and Tree SHAP, while ambient temperature dominated NO_x under both approaches. Lower-order rankings differed more substantially, consistent with correlated predictors, interactions, and the distinction between model-internal attribution and external performance degradation. These quantities therefore describe predictive dependence in the fitted models rather than isolated causal effects of turbine variables.

From a deployment perspective, the results support objective-specific rather than method-specific adaptation. Recent-history training was competitive for CO, tail weighting provided the clearest extreme-region squared-error reduction for NO_x, and historical-label calibration substantially improved global NO_x accuracy while degrading the extreme tail. The compared strategies also differ in their deployment requirements. Recent-history retraining uses previously available labeled data, SHIFT additionally requires an unlabeled target-period predictor batch and domain-model estimation, and historical calibration requires a recent labeled reference period; a true target-domain update would further require new reference emission measurements. Thus, recent-history retraining and historical calibration are computationally simpler, whereas SHIFT avoids target emission labels at the cost of additional target-covariate access and density-ratio estimation. The 2%, 5%, and 10% global-RMSE allowances should therefore be interpreted only as methodological decision parameters for comparing trade-offs, not as regulatory limits or guarantees. Regulatory use would additionally require site-specific qualification, reference-method testing, oxygen-reference corrections, permit-specific limits, and the applicable procedures under frameworks such as U.S. EPA Performance Specification 16 and the European industrial-emissions directives [20–22].

4.4. Limitations and Future Work

Several limitations restrict the scope of these conclusions. The data originate from a single gas turbine, and the public files do not provide independent load-state labels, row-level timestamps, or sufficient metadata to reconstruct dry/wet emission basis, oxygen-reference correction, or all measurement reference conditions. The published CDP unit was therefore retained without reinterpretation. Because elapsed time between adjacent rows is unknown, moving-block lengths represent dependence-sensitivity parameters rather than physical durations; residual dependence in NO_x remained appreciable even at the longest examined block. Tail-specific inference is additionally limited by the small numbers of Q_{99} observations, and the SHIFT procedure is transductive because the target predictor batch is assumed to be available when the density ratio is estimated.

Future work should therefore evaluate multiple turbines, fuels, operating regimes, and timestamped sequences that permit genuinely sequential adaptation. Larger labeled calibration and selection periods would support more precise tail-risk estimation and adaptive recalibration, while broader density-ratio estimators and explicit common-support rejection rules could improve robustness under severe domain separation. Methods that jointly account for underprediction and overprediction costs, together with adaptive conformal prediction, could provide a more complete uncertainty-aware framework [19]. Plant-level studies should ultimately connect statistical tail metrics to application-specific outcomes such as missed high-emission conditions, false-alarm burden, reference-method compliance, and intervention costs.

5. Conclusions

This study developed a reliability-oriented framework for CO and NO_x predictive emission monitoring under multiyear distribution shift. For the selected Extra Trees models, matched random within-prefix validation produced RMSE estimates approximately 25.0% lower for CO and 34.0% lower for NO_x than matched forward temporal validation, reinforcing the importance of chronological assessment. Selection on 2014 followed by locked evaluation on 2015 identified pollutant-specific trade-offs: for CO, the primary Q99-RMSE criterion selected tail weighting with $\alpha = 8$ under all three global-error budgets. On the locked 2015 period, this configuration produced only modest changes, reducing global RMSE by approximately 0.9%, Q99 RMSE by approximately 0.1%, and Q99 mean underprediction loss by approximately 2.0%. The Q99 differences were not statistically resolved in the moving-block bootstrap analyses. A recent-history model trained on 2012–2013 was also competitive and was preferred in the extended selection framework under the 5% and 10% global-error allowances. For NO_x, increasing tail-weight intensity reduced Q99 RMSE by approximately 2.4%, 4.7%, and 5.5% under the 2%, 5%, and 10% budgets, respectively, while global RMSE increased by approximately 1.6%, 4.0%, and 5.8%. Bootstrap support depended on the selected budget and dependence scale: the 10% configuration showed a consistent reduction in Q99 RMSE across the examined block lengths, whereas the corresponding reductions in Q99 mean underprediction loss remained statistically uncertain. Moving-block reselection also showed that exact configuration stability was more limited than suggested by observation-independent resampling, particularly for CO and for the more restrictive NO_x budgets. Sensitivity analyses using Q90, Q95, Q99, and an asymmetric underprediction objective further confirmed that the preferred adaptation can depend on the tail definition and loss function used for selection.

The calibration analysis further showed that global predictive calibration and high-emission performance are distinct objectives. For NO_x, linear calibration estimated exclusively from 2014 reduced locked 2015 global RMSE from 13.246 to 6.333 mg/m^3 and increased R^2 from -0.416 to 0.676, but simultaneously increased Q99 RMSE from 11.054 to

24.463 mg/m^3. Thus, a substantial improvement in overall calibration did not translate into improved performance for the highest-emission observations. Overall, the results establish global accuracy and high-emission reliability as complementary deployment criteria. They also indicate that no single adaptation strategy was uniformly preferable across pollutants, risk objectives, and allowable global-error budgets. Temporal validation, domain-support diagnostics, asymmetric tail metrics, selection-stability analysis, and locked risk-constrained evaluation together provide a reproducible basis for comparing PEMS adaptations under evolving operating conditions. The severe source–target separation and sensitivity of raw importance weights additionally indicate that covariate weighting should be applied only with explicit support and effective-sample-size diagnostics. The SHIFT formulation evaluated here is a batch-transductive adaptation procedure and should not be interpreted as an online or early-warning system. Broader validation across turbines, fuels, load regimes, and labeled target periods, together with timestamped sequential evaluation, larger extreme-emission samples, and adaptation strategies that jointly consider global error, underprediction, and overprediction, is the next step toward transferable and uncertainty-aware deployment.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/eng7100538/s1: Figure S1: Upper-tail empirical cumulative distribution functions of CO and NO_x emissions across 2011–2015; Figure S2: Measured versus predicted CO and NO_x concentrations on the locked 2015 test set; Figure S3: Residual sequences for the locked 2015 predictions, highlighting high-emission observations; Figure S4: Prediction RMSE stratified by source-defined TEY quintiles for CO and NO_x; Figure S5: Autocorrelation functions of baseline residuals and paired loss differences. Additional materials include analysis code, detailed result tables, observation-level predictions, statistical outputs, and reproducibility metadata.

Author Contributions: Conceptualization, O.R.-A. and L.A.I.C.; methodology, O.R.-A. and L.A.I.C.; software, O.R.-A. and C.A.G.-G.; validation, B.S.D.-B. and J.M.Z.P.; formal analysis, O.R.-A., B.S.D.-B. and J.M.Z.P.; investigation, O.R.-A., C.A.G.-G. and L.A.I.C.; resources, O.R.-A., L.A.I.C. and J.R.-R.; data curation, O.R.-A., B.S.D.-B. and C.A.G.-G.; writing—original draft preparation, O.R.-A.; writing—review and editing, O.R.-A., L.A.I.C. and J.R.-R.; visualization, O.R.-A. and J.M.Z.P.; supervision, J.R.-R.; project administration, O.R.-A. and J.R.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals.

Informed Consent Statement: Not applicable. This study did not involve humans.

Data Availability Statement: The dataset used in this study is publicly available from the UCI Machine Learning Repository in the Gas Turbine CO and NO_x Emission Data Set [7]. The analysis code, processed result tables, observation-level predictions, statistical inference outputs, diagnostic analyses, and supplementary reproducibility materials generated for this study are publicly available in Zenodo at https://doi.org/10.5281/zenodo.22968246.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.5 Pro (OpenAI) for language editing. The authors reviewed and edited all generated content and take full responsibility for the content of this publication.

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

Abbreviations and Symbols

The following abbreviations and symbols are used in this manuscript:

General, modeling, and operational abbreviations

PEMS	Predictive emission monitoring system
CO	Carbon monoxide
NO_x	Nitrogen oxides
UCI	University of California, Irvine Machine Learning Repository
CSV	Comma-separated values
BASE	Conventional unweighted baseline learning configuration
TAIL	Tail-weighted learning configuration
SHIFT	Covariate-shift-weighted learning configuration
SHIFT–TAIL	Combined covariate-shift and tail-weighted learning configuration
HGB	Histogram-based gradient boosting
HistGBR	Histogram-based gradient boosting regressor
XGBoost	Extreme gradient boosting
SHAP	SHapley Additive exPlanations
IID	Independent and identically distributed
ACF	Autocorrelation function
SHA-256	Secure hash algorithm with a 256-bit digest

Operational predictors

AT	Ambient temperature
AP	Ambient pressure
AH	Ambient humidity
AFDP	Air-filter differential pressure
GTEP	Gas-turbine exhaust pressure
TIT	Turbine inlet temperature
TAT	Turbine after temperature
TEY	Turbine energy yield
CDP	Compressor discharge pressure

Regression, risk, and diagnostic metrics

MAE	Mean absolute error
RMSE	Root mean squared error
R^2	Coefficient of determination
UR	Underprediction rate
MUL	Mean underprediction loss
CMU	Conditional underprediction magnitude
OR	Overprediction rate
MOL	Mean overprediction loss
CMO	Conditional overprediction magnitude
SMD	Standardized mean difference
PSI	Population stability index
ESS	Effective sample size
ROC–AUC	Area under the receiver operating characteristic curve
PR–AUC	Area under the precision–recall curve
F1	F1 score, harmonic mean of precision and recall
F2	F2 score, recall-weighted harmonic mean of precision and recall

Tail-risk and experimental symbols

Q_{90}	Source-defined 90th-percentile emission threshold
Q_{95}	Source-defined 95th-percentile emission threshold
Q_{99}	Source-defined 99th-percentile emission threshold
S	Source-development domain comprising observations from 2011 to 2013
C	Configuration-selection and calibration domain comprising observations from 2014
T	Locked final test domain comprising observations from 2015
n_s	Number of observations in the source-development domain
n_c	Number of observations in the configuration-selection domain
n_t	Number of observations in the locked test domain
y_i	Measured emission concentration for observation i
$\hat{y}_i$	Predicted emission concentration for observation i
$\bar{y}$	Mean measured emission concentration
q	Quantile level used to define a high-emission region
τ_q	Source-defined emission threshold corresponding to quantile q
A_q	Subset of observations whose measured emissions are at or above τ_q
n_q	Number of observations in the source-defined tail subset A_q

Distribution-shift and weighting symbols

D	Domain indicator, with $D = 0$ for source observations and $D = 1$ for target observations
p_i	Estimated probability that observation i belongs to the target domain, $\Pr(D = 1 \mid x_i)$
A_D	Domain-classifier ambiguity index
$\tilde{w}_i^{\text{imp}}$	Raw density-ratio importance weight for source observation i
$\tilde{w}_i^{\text{prox}}$	Raw target-proximity weight for source observation i
w_i	Final normalized and bounded sample weight for observation i
w_i^{tail}	Tail-weight component for observation i
w_i^{shift}	Covariate-shift-weight component for observation i
$w_i^{\text{shift}-\text{tail}}$	Combined covariate-shift and tail weight for observation i
α	Tail-weighting strength parameter
c	Upper clipping bound for stabilized sample weights
$1/c$	Lower clipping bound for stabilized sample weights
λ	Scaling factor used in the bounded mean-one weight normalization
$\Pi_c(\cdot)$	Bounded mean-one projection operator with clipping bound c
F_b	Set of configurations feasible under global-RMSE allowance b
b	Maximum allowable proportional increase in global RMSE during configuration selection

Statistical-inference symbols

L	Moving-block length expressed in consecutive rows
B	Number of bootstrap replicates
R	Set of prespecified random seeds used in hierarchical bootstrap inference
r	Bootstrap replicate index
M	Generic performance metric evaluated in the paired bootstrap
Δ_r	Paired bootstrap difference between the selected configuration and BASE in replicate r
CI	Confidence interval
p	Empirical two-sided bootstrap p-value, when used in statistical inference
L_1	L1 regularization penalty
L_2	L2 regularization penalty

References

1. Kaya, H.; Tüfekci, P.; Uzun, E. Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 4783–4796. [CrossRef]
2. Pachauri, N. An emission predictive system for CO and NOx from gas turbine based on ensemble machine learning approach. *Fuel* **2024**, *366*, 131421. [CrossRef]
3. Li, J.; Han, J.; Niu, D.; Jiang, X.Z. Fast and accurate gas turbine emission prediction based on a light and enhanced Transformer model. *Fuel* **2024**, *376*, 132750. [CrossRef]
4. Si, M.; Tarnoczi, T.J.; Wiens, B.M.; Du, K. Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library – Keras: A Case Study on a Cogeneration Unit. *IEEE Access* **2019**, *7*, 113463–113475. [CrossRef]
5. Si, M.; Du, K. Development of a predictive emissions model using a gradient boosting machine learning method. *Environ. Technol. Innov.* **2020**, *20*, 101028. [CrossRef]
6. Potts, R.; Hackney, R.; Leontidis, G. Tabular Machine Learning Methods for Predicting Gas Turbine Emissions. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1055–1075. [CrossRef]
7. UCI Machine Learning Repository. *Gas Turbine CO and NOx Emission Data Set*; University of California: Irvine, CA, USA, 2019. [CrossRef]
8. Kochueva, O.; Nikolskii, K. Data Analysis and Symbolic Regression Models for Predicting CO and NOx Emissions from Gas Turbines. *Computation* **2021**, *9*, 139. [CrossRef]
9. dos Santos Coelho, L.; Hultmann Ayala, H.V.; Cocco Mariani, V. CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering. *Fuel* **2024**, *355*, 129366. [CrossRef]
10. Wood, D.A. Long-term atmospheric pollutant emissions from a combined cycle gas turbine: Trend monitoring and prediction applying machine learning. *Fuel* **2023**, *343*, 127722. [CrossRef]
11. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
12. Moreno-Torres, J.G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N.V.; Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **2012**, *45*, 521–530. [CrossRef]
13. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **2000**, *90*, 227–244. [CrossRef]
14. Sugiyama, M.; Krauledat, M.; Müller, K.R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.* **2007**, *8*, 985–1005.
15. Bickel, S.; Brückner, M.; Scheffer, T. Discriminative Learning Under Covariate Shift. *J. Mach. Learn. Res.* **2009**, *10*, 2137–2155.
16. Kouw, W.M.; Loog, M. A Review of Domain Adaptation without Target Labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 766–785. [CrossRef] [PubMed]
17. Branco, P.; Torgo, L.; Ribeiro, R.P. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Comput. Surv.* **2016**, *49*, 1–50. [CrossRef]
18. Qi, D.; Majda, A.J. Using machine learning to predict extreme events in complex systems. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 52–59. [CrossRef] [PubMed]
19. Moe, D.A.; Rossi, A.; Pulga, L.; Barriere, L.; Allegorico, C. Efficient emission monitoring in gas turbines: A conformal adaptive learning approach. In *Proceedings of the ECAI 2025: 28th European Conference on Artificial Intelligence 25–30 October 2025, Bologna, Italy—Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025) Proceedings*; SAGE Publications: Thousand Oaks, CA, USA, 2026; pp. 5288–5295. [CrossRef]
20. United States Environmental Protection Agency. *Performance Specification 16: Specifications and Test Procedures for Predictive Emission Monitoring Systems in Stationary Sources*. 40 CFR Part 60, Appendix B, 2023. Available online: https://www.epa.gov/emc/performance-specification-16-predictive-emissions-monitoring-systems (accessed on 1 October 2026).
21. European Parliament and Council of the European Union. *Directive 2010/75/EU on Industrial Emissions (Integrated Pollution Prevention and Control)*; European Parliament and Council of the European Union: Brussels, Belgium, 2010.
22. European Parliament and Council of the European Union. Directive (EU) 2024/1785 of the European Parliament and of the Council of 24 April 2024 Amending Directive 2010/75/EU on Industrial Emissions and Council Directive 1999/31/EC on the Landfill of Waste. *Off. J. Eur. Union L* **2024**, *2024*, 1785.

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.