Electronic Supplementary Material (Online Resource 1)

Parametric Distribution Fitting of Photochemical Oxidants, NO_x Titration Regimes, and Leakage-Free Machine Learning Generalization in a Tropical Urban-Industrial Airshed

Saptarshi Sadhu[1*]

[1]School of Engineering and Technology, Adamas University, Kolkata 700126, West Bengal, India

*Correspondence: saptarshi1.sadhu@stu.adamasuniversity.ac.in

Air Quality, Atmosphere & Health (Springer Nature) | Online Resource 1 (ESM_1)

Overview of Supplementary Information

This document provides supplementary analytical frameworks, extended mathematical derivations, instrument detection limit evaluations, comprehensive station-wise statistical distribution catalogs across all criteria pollutants, and machine learning counter-experiment diagnostics supporting the main manuscript.

- **Section S1:** Full CAAQMS Station Metadata, Analyzer Specifications, and Operating Protocols (Tables S1 and S2).
- **Section S2:** Analytical Limits of Detection (LOD) and Near-Zero Calibration Substitution Sensitivity Analysis.
- **Section S3:** Parametric Probability Distribution Goodness-of-Fit Catalog for Non-Ozone Pollutants ($PM_{2.5}$, PM_{10}, SO_2, CO) across 104 Spatiotemporal Series (Table S3).
- **Section S4:** Machine Learning Feature Importances, Counter-Experiment Metrics, and Co-Located Trace Gas Controls at Roadside Hotspots (Tables S4 and S5).

Section S1: CAAQMS Station Metadata and Analytical Instrument Specifications

Continuous ambient air quality monitoring telemetry across the Kolkata–Howrah conurbation is maintained by the Central Pollution Control Board (CPCB) and the West Bengal Pollution Control Board (WBPCB). Each station operates automated analyzers housed within environmentally controlled, temperature-stabilized shelters (22 ± 2°C). Intake sampling probes are positioned at standard breathing zone elevations (3–15 m above ground level) in conformity with USEPA Title 40 CFR Part 58 and Indian CPCB monitoring network siting criteria.

Table S 1. Comprehensive geographical coordinates, land-use zoning, and localized sampling micro-environments across all 13 CAAQMS stations in Kolkata and Howrah.

Station Name	Code	Latitude	Longitude	Sampling Environment	Inlet Height
Ballygunge	BG	22.5264°N	88.3692°E	Urban Residential / Roadside	6.5 m
Barrackpore (SVSPA)	SVSPA	22.7661°N	88.3644°E	Suburban / Industrial Fringe	8.0 m
Belur Math	BM	22.6362°N	88.3521°E	Semi-Urban / Religious Park	7.5 m
Bidhannagar	BN	22.5802°N	88.4331°E	Commercial Roadside Corridor	5.0 m
Botanical Garden	BOG	22.5473°N	88.2941°E	Historic Parkland / Riverine	4.5 m
Dasnagar	DN	22.5921°N	88.3053°E	Heavy Metallurgy & Foundry	8.5 m
Fort William	FW	22.5532°N	88.3444°E	Parkland / Military Cantonment	5.5 m
Ghusuri	GH	22.5824°N	88.3163°E	Heavy Industrial / Smelting	9.0 m
Jadavpur	JD	22.4991°N	88.3713°E	Institutional / Urban Mixed	7.0 m
Padmapukur	PP	22.5651°N	88.3072°E	Industrial / Waterworks	6.0 m
Rabindra Bharati Univ	RBU	22.6271°N	88.3784°E	Arterial Traffic Intersection	4.0 m
Rabindra Sarovar	RS	22.5113°N	88.3637°E	Urban Lake / Ecological Buffer	4.5 m
Victoria Memorial	VC	22.5448°N	88.3426°E	Urban Background Parkland	4.0 m

Table S 2. Analytical measurement principles, instrument models, detection limits, and calibration schedules for criteria air pollutants.

Pollutant	Measurement Principle	Analytical Instrument	Lower Detection Limit	Zero/Span Calibration
O_3	UV Absorption Photometry (253.7 nm)	Thermo 49i / Teledyne T400	$1.0\,\mu g\ m^{-3}$	Automated Weekly / Bi-weekly
$NO/NO_2/NO_x$	Gas-Phase Chemiluminescence	Thermo 42i / Teledyne T200	$0.5\,\mu g\ m^{-3}$	Automated Weekly
$PM_{2.5}$	Beta-Attenuation Mass Monitor (BAM)	Met One BAM-1020	$1.0\,\mu g\ m^{-3}$	Continuous Dynamic Foil Span
PM_{10}	Beta-Attenuation Mass Monitor (BAM)	Met One BAM-1020	$2.0\,\mu g\ m^{-3}$	Continuous Dynamic Foil Span
SO_2	Pulsed Ultraviolet Fluorescence	Thermo 43i / Teledyne T100	$1.0\,\mu g\ m^{-3}$	Automated Weekly
CO	Gas Filter Correlation NDIR	Thermo 48i / Teledyne T300	$0.05\,mg\ m^{-3}$	Automated Bi-weekly
$\sum$BTEX	Online Gas Chromatography (PID)	Syntech Spectras GC955	$0.1\,\mu g\ m^{-3}$	Certified Gas Cylinders

Section S2: Analytical Limit of Detection (LOD) and Near-Zero Substitution Sensitivity Analysis

Continuous air pollution monitoring networks occasionally record sensor readings near analytical zero or slightly negative baseline drifts during routine zero-air calibrations. In our data curation protocol for parametric MLE distribution fitting (which strictly requires $x > 0$), 218 non-positive observations (0.004%) across the 5,048,351 raw 15-minute observations were removed.

To evaluate whether this filtering could introduce sample selection bias or perturb the parametric distribution selection, a rigorous sensitivity test was performed. All values below the certified instrument Limit of Detection (LOD = $1.0\,\mu$g m^{-3} for O_3, $0.5\,\mu$g m^{-3} for NO, NO_2) were replaced with the robust Hornung and Reed threshold:

$$x_{\text{sub}} = \frac{\text{LOD}}{\sqrt{2}} \tag{1}$$

For ozone, $x_{\text{sub}} = 0.707\,\mu$g m^{-3}; for NO_x, $x_{\text{sub}} = 0.354\,\mu$g m^{-3}. Re-fitting the two-parameter Lognormal, Gamma, and Weibull distributions under MLE (floc = 0) yielded virtually identical parameter estimates:

- **Akaike Information Criterion (AIC) Optimality:** Exactly 148/182 series (81.3%) retained Lognormal optimality. No series switched its winning candidate model.
- **Kolmogorov–Smirnov (D_n) Optimality:** Exactly 136/182 series (74.7%) retained Lognormal optimality.
- **Parameter Deviations:** Maximum scale parameter shifts $|\Delta\sigma|/\sigma$ and $|\Delta\mu|/\mu$ were less than 0.002% across all series.

This sensitivity analysis confirms that the dominance of the Lognormal distribution is an authentic atmospheric signal rather than an artifact of baseline truncation.

Section S3: Distribution Tail Fitting Catalog for Non-Ozone Pollutants

While the main manuscript focuses on ground-level ozone and its photochemical precursors (NO_x and $\sum$BTEX), Table S3 provides the parametric distribution fitting metrics for particulate matter fractions ($PM_{2.5}$, PM_{10}), sulfur dioxide (SO_2), and carbon monoxide (CO) across representative station-years.

Table S 3. Station-wise parametric distribution model selection for non-ozone pollutants (representative 2024 series). Parameters include Lognormal (μ, σ), Gamma (α, β), Weibull (k, λ), ΔAIC relative to best-fit, and Kolmogorov–Smirnov statistic (D_n). Values extracted directly from the verified conurbation model ranking database.

Station	Species	LN (μ, σ)	GM (α, β)	WB (k, λ)	Best Model	KS (D_n)	ΔAIC_{WB}	Physical Mechanism
Ballygunge	$PM_{2.5}$	$(3.62, 0.82)$	$(1.73, 29.45)$	$(1.32, 55.71)$	Lognormal	0.0555	+2,555.1	Multiplicative Boundary Mixing
Ballygunge	PM_{10}	$(4.32, 0.74)$	$(2.08, 47.14)$	$(1.47, 108.89)$	Lognormal	0.0528	+2,401.6	Regional Boundary Mixing
Ballygunge	SO_2	$(1.86, 0.79)$	$(1.83, 4.71)$	$(1.34, 9.47)$	Lognormal	0.0310	+2,843.9	Dispersed Fugitive Plumes
Ballygunge	CO	$(-0.69, 0.79)$	$(1.89, 0.35)$	$(1.36, 0.73)$	Lognormal	0.0411	+1,819.3	Atmospheric Dilution
Dasnagar	$PM_{2.5}$	$(4.58, 0.58)$	$(3.36, 33.76)$	$(1.94, 128.46)$	Gamma	0.0195	+246.0	Additive Foundry Combustion
Dasnagar	PM_{10}	$(5.10, 0.55)$	$(3.73, 50.78)$	$(2.07, 214.48)$	Gamma	0.0283	+226.4	Mixed Fugitive & Dilution
Dasnagar	SO_2	$(2.41, 0.95)$	$(1.32, 12.85)$	$(1.12, 17.79)$	Lognormal	0.0669	+589.3	Smelting Plume Dispersion
Dasnagar	CO	$(0.32, 0.62)$	$(3.36, 0.48)$	$(1.95, 1.81)$	Gamma	0.0421	+133.6	Additive Cupola Emissions
Victoria	$PM_{2.5}$	$(3.58, 0.75)$	$(2.00, 23.57)$	$(1.44, 52.34)$	Lognormal	0.0480	+2,547.7	Diluted Parkland Background
Victoria	PM_{10}	$(4.47, 0.60)$	$(2.97, 35.15)$	$(1.75, 118.11)$	Lognormal	0.0624	+3,701.5	Urban Background Dilution
Victoria	SO_2	$(1.91, 0.53)$	$(3.16, 2.51)$	$(1.60, 8.97)$	Lognormal	0.1519	+12,804.2	Secondary Regional Transport
Victoria	CO	$(-0.36, 0.78)$	$(2.12, 0.42)$	$(1.51, 1.00)$	Gamma	0.0273	+744.8	Homogeneous Background

Section S4: Machine Learning Diagnostics and Chemical Counter-Experiment Metrics

To mathematically demonstrate that localized vehicular nitric oxide titration is the direct cause of machine learning transferability breakdown at roadside stations, paired diagnostic models were evaluated under identical hyperparameter conditions (Random Forest: 100 trees, maximum tree depth 15, minimum leaf samples 4).

Table S 4. Diagnostic machine learning performance across roadside monitors: Comparison of Meteorology-Only vs. Co-Located Trace Gas Chemical Controls.

Station	Evaluation Protocol & Predictor Suite	Test R^2	RMSE (μg m^{-3})	MAE (μg m^{-3})
Bidhannagar (BN)	Forward Transfer (2024 → 2025): Met Only	-0.2577	13.83	9.15
	Forward Transfer (2024 → 2025): Met + NO + NO_2	-0.0764	12.92	8.65
	Forward Transfer (2024 → 2025): Met + Full Chemical Suite	+0.0195	12.28	8.12
	5-Fold Cross-Validation: Met Only	0.6808	8.54	5.12
	5-Fold Cross-Validation: Met + NO + NO_2	**0.7800**	**6.98**	**4.21**
Ballygunge (BG)	Forward Transfer (2024 → 2025): Met Only	-0.1722	15.60	11.23
	5-Fold Cross-Validation: Met Only	0.6412	9.82	6.45
	5-Fold Cross-Validation: Met + NO + NO_2	**0.7548**	**8.11**	**5.23**
Rabindra Bharati (RBU)	Forward Transfer (2024 → 2025): Met Only	0.0755	29.29	19.18
	5-Fold Cross-Validation: Met Only	0.7024	14.21	9.84
	5-Fold Cross-Validation: Met + NO + NO_2	**0.7915**	**11.64**	**7.82**

Table S 5. Gini feature importances (%) for Random Forest ozone regression models at contrasting micro-environments.

Feature Descriptor	Barrackpore (Suburban)	Victoria (Background)	Bidhannagar (Curbside)	Dasnagar (Industrial)
Solar Radiation (SR)	38.4%	35.2%	14.2%	26.8%
Ambient Temperature (AT)	24.1%	22.8%	11.8%	19.4%
Hour of Day (Diurnal Cycle)	18.2%	19.5%	12.5%	16.2%
Relative Humidity (RH)	11.5%	12.1%	9.4%	12.5%
Wind Speed (WS)	4.6%	6.2%	7.1%	14.8%
Wind Direction (WD)	3.2%	4.2%	5.0%	10.3%
Co-located NO (Chemical Sink)	–	–	**28.5%**	–
Co-located NO_2 (Odd-Oxygen Component)	–	–	**11.5%**	–