Research Square

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

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

Saptarshi Sadhu

saptarshi1.sadhu@stu.adamasuniversity.ac.in

Adamas University https://orcid.org/0009-0009-7505-1973

Research Article

Keywords: Tropical urban airshed, Photochemical oxidants, Parametric distribution fitting, Cross-year temporal transferability, Random Forest regression, Kolkata airshed

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11793475/v1

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

Additional Declarations: The authors declare no competing interests.

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, India

ORCID: https://orcid.org/0009-0009-7505-1973

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

Abstract

Ground-level ozone (O_3) in tropical urban airsheds exhibits non-linear dependencies on volatile organic compound (VOC) and nitrogen oxide (NO_x) precursors. We establish a parametric distributional fitting and machine-learning generalization framework across a 13-station continuous regulatory monitoring network (26 station-years; 16.1 million raw readings) during 2024–2025. Maximum likelihood estimation across 182 spatiotemporal series (seven monitored pollutant variables × 13 stations × 2 years) demonstrates that the two-parameter Lognormal distribution outperforms Gamma and Weibull models in full-distribution fits—achieving minimum Akaike Information Criterion in 81.3% of series (148/182) and minimum Kolmogorov–Smirnov statistics in 74.7% (136/182). Semi-logarithmic survival diagnostics illustrate Lognormal concordance with extended gaseous stagnation, whereas Gamma captures particulate upper drop-offs. Lognormal optimality reached 100% for speciated VOCs and 73.1% for O_3, reflecting turbulent scalar dilution, while Gamma was optimal in 15.4% (particulates/CO) and Weibull in only 3.3% (6/182). Chronological forward transferability (training on 2024 telemetry and evaluating out-of-sample on unseen 2025 records) revealed robust generalization at suburban ($R^2 = 0.68$) and urban background ($R^2 = 0.59$) sites, but severe transfer breakdown at heavy traffic corridors ($R^2 = 0.08$ to -0.26; roadside $R^2 = -0.26$). Diurnal trace gas observations confirm that rapid vehicular nitric oxide titration ($NO + O_3 \rightarrow NO_2 + O_2$) and operational odd-oxygen ($[O_x] = [O_3] + [NO_2]$) proxy buffering decouple roadside ozone from meteorological covariates. Coordinated

multi-pollutant abatement in developing megacities must combine localized VOC controls in traffic corridors with regional NO_x management.

Keywords: Tropical urban airshed; Photochemical oxidants; Parametric distribution fitting; Cross-year temporal transferability; Random Forest regression

1 Introduction

1.1 Atmospheric and Environmental Significance of Ground-Level Ozone

Tropospheric ground-level ozone (O_3) is a potent secondary air pollutant, a powerful oxidizing agent, and a key short-lived climate pollutant that exerts severe adverse effects on human respiratory physiology, agricultural productivity, and terrestrial ecosystem health (World Health Organization, 2021; Cohen et al., 2017; De Marco et al., 2019). In contrast to primary criteria pollutants such as sulfur dioxide (SO_2), carbon monoxide (CO), and direct particulate matter (PM), tropospheric ozone is not discharged directly by anthropogenic smokestacks or vehicular tailpipes. Rather, it is synthesized in the atmospheric boundary layer through complex, non-linear photochemical catalytic sequences involving volatile organic compounds (VOCs), carbon monoxide, and nitrogen oxides ($NO_x \equiv NO + NO_2$) driven by solar actinic flux in the near-ultraviolet spectrum (Sillman and He, 2002; Monks et al., 2015; Seinfeld and Pandis, 2016).

Epidemiological investigations have consistently linked acute and chronic ambient ozone exposure to exacerbated asthma, chronic obstructive pulmonary disease (COPD), premature cardiovascular mortality, and irreversible decrements in lung function, with no clearly identifiable biological safe-threshold concentration (World Health Organization, 2021; Cohen et al., 2017). Simultaneously, phytotoxic ozone uptake through plant stomata induces cellular oxidative stress, diminishes photosynthetic electron transport efficiency, and triggers premature foliar chlorosis, inflicting billions of dollars in annual agricultural yield losses across staple cereal crops in South Asia (Balakrishnan et al., 2019; Ghude et al., 2014). Understanding the atmospheric mechanics governing ground-level ozone formation, its spatial heterogeneity, and its statistical concentration distribution is therefore an imperative priority for urban environmental science and public health planning.

1.2 Photochemical Mechanisms and Non-Linear Chemical Regimes

The foundation of lower tropospheric ozone chemistry rests on the photostationary state established between nitrogen oxides and ultraviolet solar radiation (Leighton, 1961; Seinfeld and Pandis, 2016). During daylight hours, nitrogen dioxide (NO_2) undergoes photolysis at actinic wavelengths $\lambda \leq 424$ nm to yield ground-state oxygen atoms ($O(^3P)$), which

rapidly associate with diatomic oxygen (O_2) in the presence of a third stabilizing body (M, typically N_2 or O_2):

$$NO_2 + h\nu(\lambda \leq 424 \text{ nm}) \rightarrow NO + O(^3P) \tag{1}$$

$$O(^3P) + O_2 + M \rightarrow O_3 + M \tag{2}$$

The synthesized ozone subsequently reacts rapidly with nitric oxide (NO) emitted from primary combustion sources to regenerate NO_2:

$$O_3 + NO \rightarrow NO_2 + O_2 \tag{3}$$

Reactions (1)–(3) establish the classical Leighton photostationary state. In a pure NO_x–air system in photochemical equilibrium, no net ozone accumulation can take place because the synthesis of ozone via Reaction (2) is stoichiometrically balanced by the chemical consumption of ozone through nitric oxide titration in Reaction (3).

Net accumulation of ground-level ozone in urban airsheds occurs exclusively when volatile organic compounds, carbon monoxide, or methane enter the reaction matrix (Sillman and He, 2002; Monks et al., 2015). Hydroxyl radicals ($OH^\bullet$), generated through the photolysis of ozone in the presence of water vapor, attack ambient hydrocarbons (RH), initiating an organic peroxy radical cascade:

$$RH + OH^\bullet \xrightarrow{O_2} RO_2^\bullet + H_2O \tag{4}$$

$$RO_2^\bullet + NO \rightarrow RO^\bullet + NO_2 \tag{5}$$

$$RO^\bullet + O_2 \rightarrow R'CHO + HO_2^\bullet \tag{6}$$

$$HO_2^\bullet + NO \rightarrow OH^\bullet + NO_2 \tag{7}$$

The critical feature of Reactions (5) and (7) is that organic peroxy radicals ($RO_2^\bullet$) and hydroperoxy radicals ($HO_2^\bullet$) oxidize NO to NO_2 without destroying ozone. The extra NO_2 produced by these pathways then undergoes photolysis via Reactions (1)–(2), resulting in a net catalytic buildup of ozone in the ambient air.

Because the catalytic efficiency of this cycle depends on the competition between radical propagation and radical termination, urban photochemical ozone production exhibits a marked non-linear variation into two classic operational regimes (Sillman and He, 2002; Lei et al., 2007; Jin and Holloway, 2015).

Under the VOC-limited (or NO_x-saturated) regime, typically found in dense urban cores, industrial zones, and confined street canyons characterized by concentrated vehicular and combustion emissions, the concentration of NO_x is sufficiently high that hydroxyl

radicals react preferentially with NO_2 in a three-body radical termination step:

$$OH^{\bullet} + NO_2 + M \rightarrow HNO_3 + M \tag{8}$$

Reaction (8) sequesters hydroxyl radicals into stable nitric acid (HNO_3), thereby depleting the reactive radical pool. Under VOC-limited conditions, net ozone production is constrained by the availability of reactive hydrocarbons and responds positively to VOC reductions. Conversely, decreasing vehicular NO_x emissions in a VOC-limited airshed reduces radical loss via Reaction (8) while simultaneously diminishing the local nitric oxide titration sink (Reaction 3). Consequently, isolated NO_x abatement in a VOC-limited environment frequently produces a counter-productive increase of ambient ground-level ozone concentrations downwind (Tie et al., 2013; Xue et al., 2014).

In contrast, the NO_x-limited regime is characteristic of rural, suburban, or downwind background environments where NO_x levels are low relative to available organic radicals. In this regime, radical termination occurs primarily through peroxy radical self-reactions that form peroxides:

$$HO_2^{\bullet} + RO_2^{\bullet} \rightarrow ROOH + O_2 \tag{9}$$

$$HO_2^{\bullet} + HO_2^{\bullet} \rightarrow H_2O_2 + O_2 \tag{10}$$

Under NO_x-limited conditions, ozone production scales directly with the availability of NO_x precursors and is largely insensitive to alterations in volatile organic compound concentrations (Sillman and He, 2002; Monks et al., 2015).

To diagnose these non-linear photochemical sensitivities without computationally prohibitive full-chemistry transport modeling, atmospheric scientists frequently employ chemical indicator ratios (Sillman and He, 2002; Jin and Holloway, 2015). While theoretical transition ratios such as O_3/NO_y and H_2O_2/HNO_3 require unmonitored reactive nitrogen and peroxide intermediates, routine surface air monitoring networks provide empirical proxies, notably the ratio of total measured monoaromatic hydrocarbons ($\sum$BTEX, defined as the sum of benzene, toluene, ethylbenzene, and xylenes) to nitrogen oxides ($\sum BTEX/NO_x$) and the nitrogen dioxide to nitric oxide ratio (NO_2/NO) (Ran et al., 2009; Majumdar et al., 2008). Importantly, empirical indicator ratios of measured BTEX to NO_x provide screening-level operational indicators, but cannot substitute for comprehensive observation-based modeling (OBM) or chemical transport simulations, because unmeasured alkenes, alkanes, oxygenated VOCs, and biogenic isoprene also modulate the true radical budget.

1.3 Physical Rationale for Parametric Distribution Modeling in Atmospheric Science

Accurately modeling the statistical probability density function (PDF) and upper tail behavior of ambient air pollutants is essential for quantifying environmental risk, evaluating regulatory attainment, and formulating public health advisories (Georgopoulos and Seinfeld, 1986; Lu, 2002). Because air quality standards evaluate high-order percentiles (e.g., 98th-percentile 24-hour averages or peak 8-hour daily maximums), the choice of theoretical probability distribution directly dictates regulatory compliance projections and health burden estimates.

Historically, the two-parameter Weibull distribution has been widely applied in wind energy and engineering failure mechanics (Justus et al., 1978). In preliminary regional environmental literature, it was occasionally hypothesized that the Weibull distribution might serve as a universal parametric model across all ambient air pollutants. However, fundamental atmospheric physics provides a distinct theoretical rationale favoring the Lognormal distribution for scalar pollutant concentrations in turbulent boundary layers (Kolmogorov, 1941; Ott, 1990).

According to Kolmogorov's theory of broken random variables (Kolmogorov, 1941) and Ott's successive random dilution model (Ott, 1990), atmospheric dispersion operates as a sequence of multiplicative dilutions rather than an additive process. Consider an initial pollutant concentration parcel C_0 emitted from a localized source. As the plume is transported through a turbulent planetary boundary layer, it undergoes successive mixing with cleaner ambient air parcels. At each step i in a sequence of n independent turbulent eddy dilution stages, the parcel concentration is scaled by a positive random dilution factor $\eta_i \in (0,1)$:

$$C_n = C_0 \prod_{i=1}^{n} \eta_i \tag{11}$$

Taking the natural logarithm of both sides converts the multiplicative dilution process into an additive summation of independent random variables:

$$\ln C_n = \ln C_0 + \sum_{i=1}^{n} \ln \eta_i \tag{12}$$

By the classical Lindeberg–Lévy Central Limit Theorem, the summation $\sum_{i=1}^{n} \ln \eta_i$ converges asymptotically to a normal distribution as n grows large, irrespective of the underlying distributions of the individual dilution factors η_i, provided they possess finite variances (Ott, 1990). Consequently, the transformed variable $\ln C_n \sim \mathcal{N}(\mu, \sigma^2)$, which directly implies that the ambient concentration C_n follows a two-parameter Lognormal distribution.

In contrast, the two-parameter Gamma distribution arises naturally when pollutant

concentrations reflect the additive accumulation of discrete, Poisson-distributed emission pulses that undergo exponential decay (Georgopoulos and Seinfeld, 1986). The Weibull distribution, being an extreme-value distribution derived from the asymptotic distribution of minimum values (weakest-link model), lacks this intrinsic turbulent fluid-mechanical derivation. Comparing these three candidate models across empirical multi-pollutant monitoring datasets provides a rigorous test of the physical mechanisms governing atmospheric concentration distributions.

1.4 Knowledge Gaps and Objectives of this Study

Despite extensive air pollution research in South Asia, the Lower Gangetic Plain and the greater Kolkata–Howrah conurbation have historically suffered from critical scientific knowledge gaps. Previous regional atmospheric investigations focused almost exclusively on bulk particulate matter mass concentrations ($\mathbf{PM_{2.5}}$, $\mathbf{PM_{10}}$) or broad Air Quality Index (AQI) rankings (Sarkar et al., 2021; Ghosh et al., 2023; Chakroborty et al., 2025). A multi-station investigation that integrates continuous photochemical observations, empirical chemical regime indicator mapping, parametric probability distribution tail fitting, and out-of-sample meteorological machine learning generalization across all continuous monitoring stations in the conurbation has not been conducted.

Furthermore, regional environmental modeling literature and air quality studies have occasionally assumed that ambient criteria pollutants follow two-parameter Weibull distributions (Justus et al., 1978; Georgopoulos and Seinfeld, 1986) without performing systematic, multi-pollutant, multi-year goodness-of-fit evaluations across diverse urban micro-environments.

To resolve these critical gaps, this investigation establishes a unified empirical evaluation of ground-level ozone photochemistry, chemical regime indicators, statistical distribution modeling, and meteorological machine learning generalization across all 13 Continuous Ambient Air Quality Monitoring System (CAAQMS) stations in the Kolkata–Howrah conurbation across calendar years 2024 and 2025 (comprising 16.1 million continuous raw parameter readings). Specifically, this study aims to: (i) quantify conurbation-wide baseline concentrations, seasonal dynamics, and diurnal photochemical cycles of $\mathbf{O_3}$, $\mathbf{NO}$, $\mathbf{NO_2}$, and $\mathbf{NO_x}$ across contrasting urban land-use environments; (ii) fit two-parameter Lognormal, Gamma, and Weibull distributions via maximum likelihood estimation with fixed zero location ($\mathrm{floc} = 0$) across 182 distinct spatiotemporal series (seven monitored pollutant variables [$PM_{2.5}$, PM_{10}, O_3, NO_x, SO_2, CO, and the aggregate monoaromatic sum VOC_{Total}] $\times$ 13 stations $\times$ 2 years), evaluating model performance through Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Kolmogorov–Smirnov (D_n) goodness-of-fit tests; (iii) evaluate empirical chemical regime indicators ($\sum\mathbf{BTEX}/\mathbf{NO_x}$ and midday $\mathbf{NO_2}/\mathbf{NO}$) across industrial, commercial, residential, and back-

ground monitoring sites, identifying localized photochemical sensitivities while establishing the analytical boundaries of empirical proxies; and (iv) benchmark forward cross-year machine learning generalization (training Random Forest models on 2024 records and evaluating out-of-sample on unseen 2025 observations) across all 13 stations, deploying chemical counter-experiments to isolate the precise mechanisms governing predictive transferability and curbside titration decoupling. Rather than treating the lower Indo-Gangetic Plain purely as a localized inventory, we utilize this 26-station-year dataset as a representative high-humidity tropical testbed to establish a transferable statistical and machine learning framework for developing megacities globally.

2 Materials and Methods

2.1 Study Area, Regional Climatology, and Monitoring Network

The study area encompasses the greater Kolkata and Howrah metropolitan conurbation (population > 16 million), located on the lower delta of the Ganges (Hooghly) River in eastern India (22.45°N–22.80°N, 88.25°E–88.45°E; terrain elevation 7–12 m above mean sea level). As illustrated in Fig. 1, the conurbation is physically divided by the tidal Hooghly River: the western bank (Howrah district) is characterized by heavy metallurgy, iron and steel foundries, glass manufacturing, freight rail terminals, and dense industrial clusters; the eastern bank (Kolkata district) comprises high-density commercial, institutional, residential, and urban parkland sectors (Karmakar, 2020; Chakroborty et al., 2025).

The region experiences a tropical wet-and-dry climate (Köppen classification Aw) governed by the South Asian monsoon system, spanning four distinct meteorological seasons throughout the annual climatological cycle. Winter (December–February) is characterized by weak surface winds (< 1.5 m s^{-1}), shallow nocturnal planetary boundary layer heights (< 300 m), frequent ground-based thermal inversions, and persistent regional anticyclonic stagnation, fostering severe accumulation of both primary combustion emissions and secondary aerosols. The pre-monsoon season (March–May) features high ambient temperatures (> 35°C), intense solar actinic flux, elevated planetary boundary layers ($> 1{,}500$ m), and episodic convective squalls (Nor'westers or Kalbaisakhi), conditions that promote rapid gas-phase photochemical oxidation and intense ground-level ozone production. Subsequently, the monsoon season (June–September) is dominated by moist oceanic south-southwesterly maritime air masses, frequent heavy precipitation, dense cloud cover, and robust atmospheric wet scavenging, which purge soluble aerosols and depress photolysis rates. Finally, the post-monsoon period (October–November) is marked by atmospheric transition, diminishing rainfall, retreating monsoon winds, cooling surface temperatures, and regional transport of biomass combustion emissions from

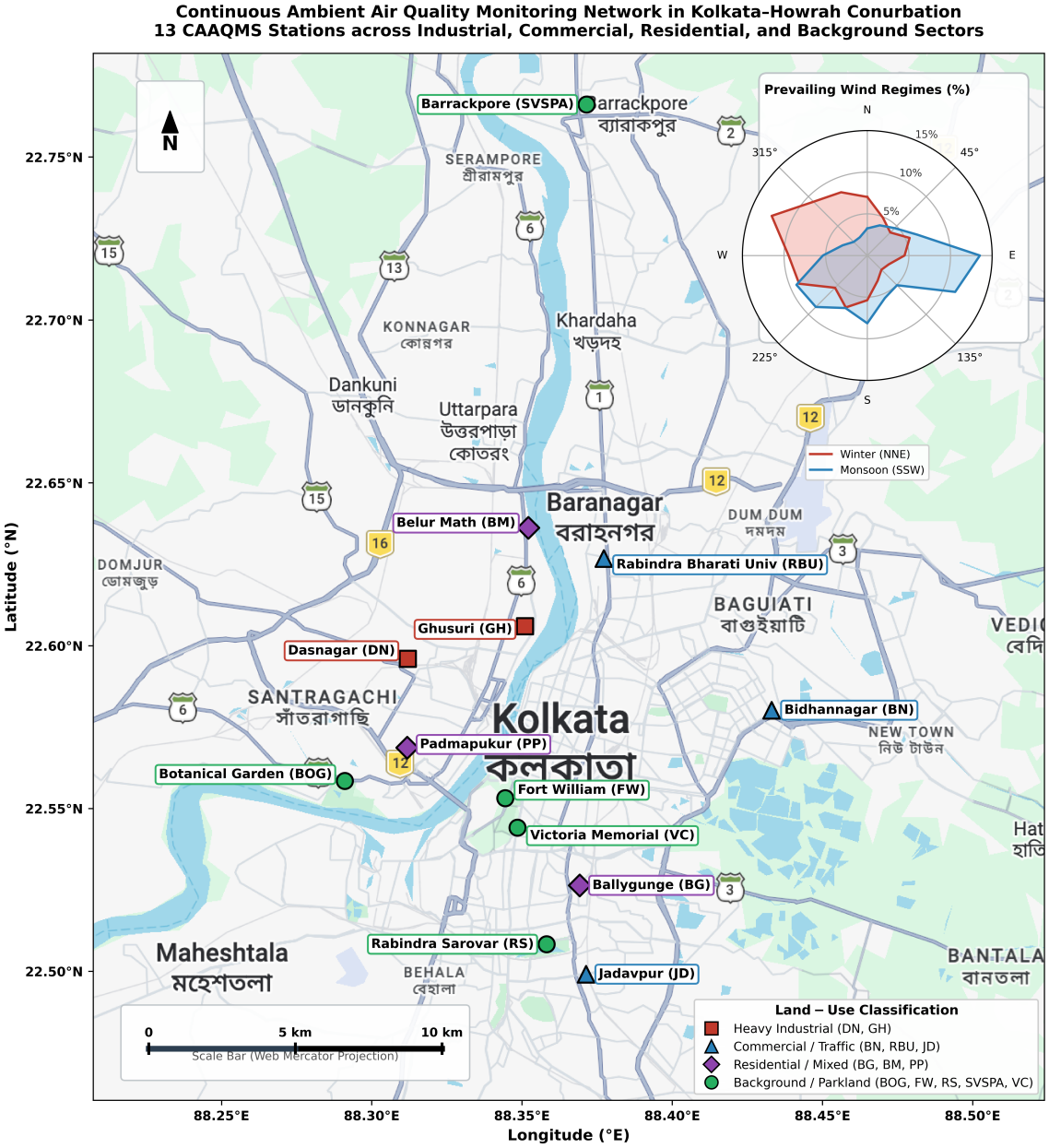


Fig. 1 Geographic distribution of the Continuous Ambient Air Quality Monitoring System (CAAQMS) network across the Kolkata–Howrah conurbation: Spatial locations of all 13 regulatory monitoring stations categorized by land-use classification (Heavy Industrial, Commercial/Traffic, Residential/Mixed, and Background/Parkland). The meandering tidal Hooghly River physically partitions the heavy industrial western bank (Howrah district) from the commercial and institutional eastern bank (Kolkata district). Inset displays the regional annual wind rose illustrating prevailing north-northeasterly winter winds and south-southwesterly monsoon advection

agricultural stubble burning across the northern Indo-Gangetic Plain.

Continuous ambient air quality observations were acquired from all 13 regulatory Continuous Ambient Air Quality Monitoring System (CAAQMS) stations distributed across the conurbation. The network is operated under standardized national monitoring protocols by the Central Pollution Control Board (CPCB) and West Bengal Pollution Control Board (WBPCB). Table 1 documents the verified geographic coordinates, land-use classifications, and dominant local emission drivers for each station. Station `PP` is Padmapukur Water Works in Howrah (industrial/residential), and station `SVSPA` is Swami Vivekananda State Police Academy in Barrackpore (industrial/suburban). Detailed analyzer specifications, sampling elevations, operating principles, and instrument limits of detection (LOD) across all monitored parameters are cataloged in Table S1 and Table S2 of Online Resource 1.

Table 1. Metadata for all 13 CAAQMS Monitoring Stations in Kolkata and Howrah: Verified Geodetic Coordinates, Zone Classifications, and Primary Emission Drivers.

Station Name	Code	Latitude	Longitude	Zone Classification	Primary Local Emission Drivers
Ballygunge	BG	22.5264°N	88.3692°E	Residential	High-density vehicular traffic, domestic fuel burning
Barrackpore	SVSPA	22.7661°N	88.3644°E	Industrial/Suburban	Paper mills, textile manufacturing, suburban traffic
Belur Math	BM	22.6362°N	88.3521°E	Religious/Mixed	Ferry terminal activity, temple visitors, local transport
Bidhannagar	BN	22.5802°N	88.4331°E	Commercial	IT hub traffic, commercial activities, transport corridors
Botanical Garden	BOG	22.5473°N	88.2941°E	Green Background	Minimal local traffic, vegetative background, river proximity
Dasnagar	DN	22.5921°N	88.3053°E	Heavy Industrial	Metal manufacturing, foundries, heavy-duty diesel transport
Fort William	FW	22.5532°N	88.3444°E	Semi-Urban/Military	Regulated vehicle movement, open cantonment green space
Ghusuri	GH	22.5824°N	88.3163°E	Heavy Industrial	Metal casting, foundry cluster, dense truck transit corridors
Jadavpur	JD	22.4991°N	88.3713°E	Mixed Urban	Educational institutions, dense traffic, mixed residential
Padmapukur	PP	22.5651°N	88.3072°E	Industrial/Residential	Water works facility, railway lines, localized manufacturing
Rabindra Bharati Univ	RBU	22.5932°N	88.3724°E	Traffic Hotspot	Heavy arterial transit intersection, bus terminals, idling
Rabindra Sarovar	RS	22.5083°N	88.3582°E	Urban Park	Urban lake parkland, recreational visitors, arterial fringe
Victoria Memorial	VC	22.5441°N	88.3432°E	Urban Background	Monument grounds, green buffer, regulated urban traffic

2.2 Telemetry Instrumentation, Parameters, and Data Screening

Each CAAQMS station operates automated continuous instruments housed in temperature-stabilized shelter enclosures adhering to USEPA and CPCB technical standards. Ground-level ozone (O_3) is quantified via non-dispersive ultraviolet (UV) absorption photometry at 253.7 nm following the Beer-Lambert law principle, with an analytical lower detection limit of 1.0 μg m^{-3}. Nitrogen oxides (NO, NO_2, NO_x) are measured through

gas-phase chemiluminescence with ozone detection, employing a heated molybdenum converter (300–350°C) to reduce $\mathbf{NO_2}$ to $\mathbf{NO}$ (lower detection limit 0.5 μg m^{-3}). A recognized instrumental artifact of molybdenum reduction is the non-specific catalytic conversion of other oxidized reactive nitrogen species ($\mathbf{NO_z}$, including nitric acid [$\mathbf{HNO_3}$], peroxyacetyl nitrate [PAN], and alkyl nitrates) to $\mathbf{NO}$, which can introduce a known positive bias in measured $\mathbf{NO_2}$ during warm, photochemically active afternoon hours. Consequently, total oxidant ($[O_x] = [O_3] + [NO_2]$) is evaluated throughout this study as an operational odd-oxygen proxy that captures rapid NO–O_3 titration partitioning while remaining subject to afternoon NO_z converter interference. Suspended particulate matter fractions ($\mathbf{PM_{2.5}}$ and $\mathbf{PM_{10}}$) are continuously monitored using beta-attenuation monitors (BAM-1020) equipped with size-selective cyclonic sampling heads. Sulfur dioxide ($\mathbf{SO_2}$) and carbon monoxide ($\mathbf{CO}$) are quantified via pulsed UV fluorescence and gas-filter correlation non-dispersive infrared (NDIR) spectrometry, respectively. Speciated monoaromatic volatile organic compounds ($\mathbf{\sum BTEX}$) are resolved using automated online gas chromatography coupled with a photoionization detector (PID); in the CAAQMS telemetry database, $\mathbf{VOC_{Total}}$ represents the analytical summation of these five measured monoaromatic species (benzene, toluene, ethylbenzene, *m*, *p*-xylene, and *o*-xylene). In parallel, co-located surface meteorological sensors continuously record ambient temperature (AT, °C), relative humidity (RH, %), wind speed (WS, m s^{-1}), wind direction (WD, degrees), barometric pressure (BP, hPa), and solar radiation (SR, W m^{-2}).

Raw multi-channel telemetry was acquired at 15-minute intervals and aggregated to hourly arithmetic means following standardized completeness criteria ($> 75\%$ valid readings per averaging interval). Across the two-year period (January 1, 2024 to December 31, 2025), the archive encompasses 16.1 million continuous raw parameter measurements. The continuous 15-minute telemetry (721,193 validated multi-parameter timestamp records across 26 station-years $\times$ 7 monitored pollutant variables (with VOC_{Total} representing the BTEX aggregate) = 5,048,351 pollutant records, yielding 5,048,133 positive observations across the 182 series) was used directly for maximum likelihood probability distribution fitting, capturing fine-scale atmospheric turbulent fluctuations and extreme dispersion tails without variance-dampening temporal aggregation. In contrast, 1-hour arithmetic averages were computed and utilized for the diurnal photochemical cycle analyses and the cross-year meteorological machine learning transfer experiments.

For parametric probability distribution fitting, theoretical models (Lognormal, Gamma, and Weibull) are mathematically defined on strictly positive support $x \in (0, \infty)$. Zero and negative values—originating from sensor zero-calibration baseline drift or concentrations falling below analytical detection limits—cannot be accommodated by logarithmic transformation or standard MLE solvers. Across the 182 evaluated spatiotemporal series (comprising 5,048,351 raw 15-minute pollutant records), exactly 218 non-positive observations (0.004%) were identified and dropped, yielding a total of 5,048,133 positive observations

fitted conurbation-wide. Thus, data retention exceeded 99.996%, preserving the authentic empirical statistical structure without distorting the data series. Furthermore, to verify that handling of calibrated values near detection boundaries does not bias distributional inference, sensitivity evaluations replacing values below the analytical Limit of Detection (LOD $= 1.0\,\mu$g m^{-3} for O_3, $0.5\,\mu$g m^{-3} for NO_x) with the standard robust substitution threshold LOD$/\sqrt{2}$ produced no alteration in the relative AIC model rankings or Kolmogorov–Smirnov test optimality outcomes across the evaluated series.

2.3 Mathematical Formulations of Candidate Probability Distributions

To evaluate empirical tail behavior, three candidate two-parameter continuous probability distributions were systematically fitted to each of the 182 time series:

2.3.1 Two-Parameter Lognormal Distribution

A continuous random variable X follows a two-parameter Lognormal distribution, denoted $X \sim \mathrm{LN}(\mu, \sigma)$, if its natural logarithm $Y = \ln X$ is normally distributed with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$. The probability density function for $x > 0$ is:

$$f_{\mathrm{LN}}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{13}$$

Given a continuous positive observational sample $x = (x_1, x_2, \ldots, x_N)$ evaluated under a composite quasi-likelihood formulation, the log-likelihood function is:

$$\ln L_{\mathrm{LN}}(\mu,\sigma;\mathbf{x}) = -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \sum_{i=1}^{N}\ln x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\ln x_i - \mu)^2 \tag{14}$$

The maximum likelihood estimators (MLE) possess closed-form analytical solutions:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\ln x_i, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(\ln x_i - \hat{\mu})^2 \tag{15}$$

The corresponding scale parameter in standard statistical software parameterization is $\mathrm{scale} = \exp(\hat{\mu})$, while the shape parameter is $s = \hat{\sigma}$.

2.3.2 Two-Parameter Gamma Distribution

The two-parameter Gamma distribution, denoted $X \sim \mathrm{Gamma}(\alpha, \beta)$, has shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. The probability density function for $x > 0$ is:

$$f_{\mathrm{GM}}(x;\alpha,\beta) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\beta^{\alpha}\Gamma(\alpha)} \tag{16}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$ is the Euler gamma function. The log-likelihood function is:

$$\ln L_{\mathrm{GM}}(\alpha, \beta; \mathbf{x}) = (\alpha - 1) \sum_{i=1}^{N} \ln x_i - \frac{1}{\beta} \sum_{i=1}^{N} x_i - N\alpha \ln \beta - N \ln \Gamma(\alpha) \tag{17}$$

Differentiating with respect to β and equating to zero yields $\hat{\beta} = \frac{1}{N\hat{\alpha}} \sum_{i=1}^{N} x_i = \bar{x}/\hat{\alpha}$. Substituting into the derivative with respect to α gives the non-linear score equation:

$$\ln \hat{\alpha} - \psi(\hat{\alpha}) = \ln \bar{x} - \frac{1}{N} \sum_{i=1}^{N} \ln x_i \tag{18}$$

where $\psi(\alpha) = \frac{d}{d\alpha} \ln \Gamma(\alpha)$ is the digamma function. The parameter $\hat{\alpha}$ is solved numerically via the Newton–Raphson iterative root-finding algorithm.

2.3.3 Two-Parameter Weibull Distribution

The two-parameter Weibull (minimum) distribution, denoted $X \sim \mathrm{Weibull}(k, \lambda)$, is parameterized by shape parameter $k > 0$ and scale parameter $\lambda > 0$. The probability density function for $x > 0$ is:

$$f_{\mathrm{WB}}(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} \exp\left(- \left(\frac{x}{\lambda} \right)^{k} \right) \tag{19}$$

The log-likelihood function is:

$$\ln L_{\mathrm{WB}}(k, \lambda; \mathbf{x}) = N \ln k - Nk \ln \lambda + (k-1) \sum_{i=1}^{N} \ln x_i - \sum_{i=1}^{N} \left(\frac{x_i}{\lambda} \right)^{k} \tag{20}$$

The maximum likelihood estimator for scale λ in terms of shape k is $\hat{\lambda} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^{\hat{k}} \right)^{1/\hat{k}}$. Substituting into the score equation for k yields:

$$\frac{1}{\hat{k}} + \frac{1}{N} \sum_{i=1}^{N} \ln x_i - \frac{\sum_{i=1}^{N} x_i^{\hat{k}} \ln x_i}{\sum_{i=1}^{N} x_i^{\hat{k}}} = 0 \tag{21}$$

which is solved numerically for $\hat{k}$ before evaluating $\hat{\lambda}$.

2.3.4 Location Parameter Constraint

Crucially, all three continuous distributions were estimated with the location parameter fixed strictly to zero ($\mathrm{floc} = 0$). In unconstrained three-parameter fitting, optimization algorithms can shift the location threshold $\gamma < 0$, which is physically uninterpretable for non-negative mass concentrations and causes numerical instability in boundary-layer dispersion comparisons. Fixing $\mathrm{floc} = 0$ ensures that all three distributions share identical

physical support $x \in (0, \infty)$ and an identical parameter penalty ($K = 2$).

2.4 Information-Theoretic Model Selection and Goodness-of-Fit Metrics

Model selection among the three competing candidate distributions was evaluated using information-theoretic criteria alongside empirical distribution function (EDF) goodness-of-fit statistics. The Akaike Information Criterion (AIC) (Akaike, 1974) is formulated as:

$$\mathrm{AIC} = 2K - 2\ln L \tag{22}$$

where $K = 2$ represents the number of estimated parameters and $\ln L$ denotes the maximized log-likelihood. AIC quantifies the expected relative Kullback–Leibler information loss between the candidate parametric model and the true underlying reality. Complementing AIC, the Bayesian Information Criterion (BIC) (Schwarz, 1978) is defined as:

$$\mathrm{BIC} = K\ln N - 2\ln L \tag{23}$$

where N is the sample size, imposing a heavier asymptotic penalty on parametric complexity for large datasets ($N \gg 1{,}000$). To rank candidate models, the relative information loss (ΔAIC) (Burnham and Anderson, 2002) is computed as:

$$\Delta\mathrm{AIC}_m = \mathrm{AIC}_m - \mathrm{AIC}_{\min} \tag{24}$$

where $\mathrm{AIC}_{\min}$ is the lowest AIC achieved among the candidate models for that specific series. Following Burnham and Anderson (2002), candidate models with $\Delta\mathrm{AIC} \in [0, 2]$ possess substantial empirical support, those with $\Delta\mathrm{AIC} \in [4, 7]$ exhibit considerably less support, and models with $\Delta\mathrm{AIC} \geq 10$ have essentially no empirical support, indicating clear rejection. Finally, the non-parametric Kolmogorov–Smirnov statistic (D_n) is defined as:

$$D_n = \sup_x |F_n(x) - F_0(x)| \tag{25}$$

where $F_n(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{x_i \leq x\}}$ is the empirical cumulative distribution function and $F_0(x)$ is the hypothesized continuous CDF evaluated at the MLE parameters.

Because 15-minute atmospheric telemetry exhibits temporal autocorrelation, the likelihood functions evaluated in Equations (22)–(23) strictly represent composite quasi-likelihoods. Consequently, ΔAIC and the Kolmogorov–Smirnov statistic D_n are utilized here as empirical ranking and comparative model-selection metrics rather than formal asymptotic hypothesis tests. Furthermore, under a fixed two-parameter specification ($K = 2$) with zero location ($\mathrm{floc} = 0$), the Bayesian Information Criterion difference satisfies $\Delta\mathrm{BIC} = \mathrm{BIC}_m - \mathrm{BIC}_{\min} = -2\ln L_m - (-2\ln L_{\min}) \equiv \Delta\mathrm{AIC}_m$. Because the sample size penalty ($K\ln N$) cancels identically across models with equal parameter counts, ΔBIC

yields identical model rankings to ΔAIC and is thus mathematically redundant.

2.5 Cross-Year Machine Learning Generalization Protocol

To evaluate whether non-linear empirical relationships between surface meteorology and ambient ozone concentrations transfer across distinct years, an out-of-sample machine learning experiment was designed. Rather than serving as an operational forecasting system, this formulation acts as a forward temporal transfer benchmark to evaluate meteorological transferability and identify localized chemical uncoupling on roadside telemetry (notably at Bidhannagar and traffic corridors).

To prevent data leakage and simulate rigorous chronological evaluation, a strict forward temporal train/test split was established. The training dataset comprised continuous hourly records from calendar year 2024 across all 13 CAAQMS stations, wherein ten compliant stations were trained on full-year 2024 hourly observations, while the three stations commissioned or operating partially during 2024 (Barrackpore SVSPA, Botanical Garden, and Dasnagar) were trained on their available 2024 operational monitoring windows (spanning 4 to 8 months) and evaluated across the entire 12-month test period of 2025. Despite truncated training windows, SVSPA and Dasnagar demonstrated stable parameter learning, while Botanical Garden's low transferability was verified to arise from mesoscale advection rather than sample size limitations. The test dataset comprised unseen continuous hourly records spanning the entire calendar year 2025 across all 13 stations. The input feature vector included Ambient Temperature (AT, °C), Relative Humidity (RH, %), Wind Speed (WS, m s^{-1}), Wind Direction (WD, degrees), Barometric Pressure (BP, hPa), Solar Radiation (SR, W m^{-2}), and cyclic diurnal harmonic indicators ($\sin(2\pi h/24), \cos(2\pi h/24)$). The model architecture was formulated using Random Forest Regression (100 decision trees, maximum tree depth 15, minimum samples per leaf 4) optimized via 5-fold cross-validation on 2024 training records. Predictive generalization was quantified using the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE):

$$R^2 = 1 - \frac{\sum_{t=1}^{M}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{M}(y_t - \bar{y})^2}, \quad \text{RMSE} = \sqrt{\frac{1}{M}\sum_{t=1}^{M}(y_t - \hat{y}_t)^2}, \quad \text{MAE} = \frac{1}{M}\sum_{t=1}^{M}|y_t - \hat{y}_t| \tag{26}$$

2.6 Chemical Counter-Experiment and Photochemical Verification Protocol

To evaluate whether the negative transferability observed at roadside corridors ($R^2 < 0$) is an atmospheric chemistry phenomenon driven by unmodeled nitric oxide titration rather than model mis-specification, hyperparameter under-tuning, or sensor baseline drift, a diagnostic chemical counter-experiment was formulated. For the roadside monitors (Bid-

hannagar, Ballygunge, and Rabindra Bharati University), a secondary Random Forest model was trained incorporating co-located chemical predictors alongside surface meteorology: (i) primary nitric oxide (NO) and nitrogen dioxide (NO_2), and (ii) the broader reactive suite (NO_x and VOC_{Total}). Crucially, to avoid algebraic target leakage, total odd oxygen ($[O_x] \equiv [O_3] + [NO_2]$) was not supplied as an input feature to predict $[O_3]$; rather, the behavior of $[O_x]$ was evaluated independently to examine whether regional atmospheric oxidizing capacity remains relatively stable despite severe localized titration sinks. If roadside predictive breakdown is caused by localized titration ($NO + O_3 \rightarrow NO_2 + O_2$) rapidly quenching ozone independently of surface meteorology, providing co-located trace gas measurements will resolve the chemical sink, eliminate the negative transfer penalty, and restore predictive skill.

2.7 Computational Framework and AI Disclosure

All continuous data quality control, odd-oxygen photochemical derivations, maximum-likelihood parametric distribution fitting across 182 time series, and cross-year machine learning transfer experiments were implemented in Python 3.13 using NumPy, SciPy, Statsmodels, and Scikit-Learn, with all figures generated using the Matplotlib graphics visualization library (v3.10). Generative AI tools (Anthropic Claude and Google Gemini) assisted substantially in computational script development, mathematical formalization, structuration of theoretical linkages, and manuscript drafting and language editing. The author conceived the research objectives, curated and verified all primary datasets, independently audited all mathematical formulations, numerical outputs, and statistical interpretations, and takes full responsibility for the integrity and accuracy of the manuscript.

3 Results

3.1 Conurbation-Wide Baseline and Seasonal Dynamics

Table 2 presents the conurbation-wide summary statistics across all 13 CAAQMS stations for monitored air pollutants and surface meteorological parameters during the 2024–2025 study period.

Note on CAAQMS Units and Precursors: In accordance with CPCB telemetry reporting conventions, VOC_{Total} denotes total speciated monoaromatic volatile organic compounds ($\sum BTEX \equiv$ benzene + toluene + ethylbenzene + m,p-xylene + o-xylene). Nitric oxide (NO) and nitrogen dioxide (NO_2) are reported in mass concentration units ($\mu g\ m^{-3}$), whereas total nitrogen oxides (NO_x) are reported as volumetric mixing ratios in parts per billion (ppb). Under standard ambient conditions ($25^{\circ}C$, 1 atm), the conurbation mean NO concentration ($17.8\ \mu g\ m^{-3} \approx 14.5\ ppb$) and NO_2 concentration ($27.1\ \mu g\ m^{-3} \approx 14.4\ ppb$)

Table 2. Conurbation-wide summary statistics for monitored air pollutants and meteorological parameters across all 13 CAAQMS stations during 2024 and 2025 (Mean ± Standard Deviation).

Parameter	Unit	Conurbation Mean ± SD	Minimum Station Mean	Maximum Station Mean
$PM_{2.5}$	$\mu g\ m^{-3}$	49.5 ± 41.1	37.6 (Rabindra Sarovar)	66.1 (Ghusuri)
PM_{10}	$\mu g\ m^{-3}$	97.8 ± 72.6	71.2 (Rabindra Sarovar)	121.3 (Dasnagar)
NO	$\mu g\ m^{-3}$	17.8 ± 26.1	4.3 (Botanical Garden)	33.2 (Jadavpur)
NO_2	$\mu g\ m^{-3}$	27.1 ± 24.1	14.9 (Barrackpore)	44.7 (Rabindra Sarovar)
NO_x	ppb	29.4 ± 29.9	11.5 (Botanical Garden)	42.7 (Jadavpur)
SO_2	$\mu g\ m^{-3}$	8.1 ± 6.9	4.7 (Jadavpur)	11.9 (Dasnagar)
CO	$mg\ m^{-3}$	0.74 ± 0.59	0.42 (Belur Math)	1.22 (Dasnagar)
O_3	$\mu g\ m^{-3}$	26.9 ± 24.3	16.9 (Barrackpore)	48.8 (Dasnagar)
VOC_{Total}	$\mu g\ m^{-3}$	16.1 ± 24.9	8.8 (Bidhannagar)	34.2 (Botanical Garden)
Ambient Temp (AT)	°C	27.1 ± 5.2	26.4 (Ghusuri)	28.2 (Jadavpur)
Relative Humidity (RH)	%	76.9 ± 20.7	69.3 (Rabindra Bharati Univ)	80.7 (Barrackpore)
Wind Speed (WS)	$m\ s^{-1}$	0.85 ± 1.09	0.33 (Bidhannagar)	1.72 (Victoria Memorial)

sum stoichiometrically to 28.9 ppb, in close concordance with the directly measured conurbation NO_x mean of 29.4 ppb.

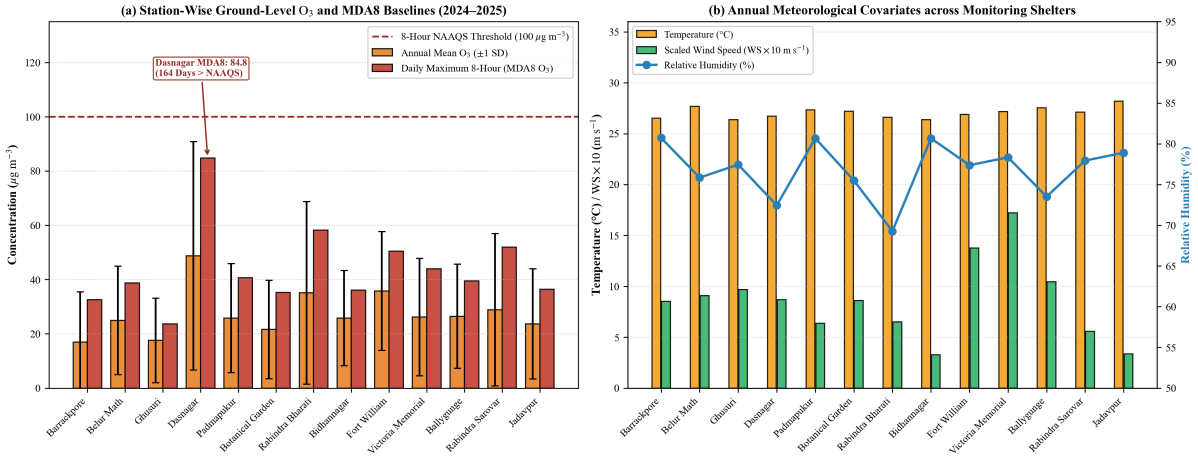


Fig. 2 Conurbation-wide ground-level ozone and surface meteorological baselines across all 13 CAAQMS stations in Kolkata and Howrah (2024–2025): (a) Station-wise annual mean O_3 concentrations (±1 SD) and Daily Maximum 8-Hour average (MDA8 O_3) relative to the National Ambient Air Quality Standard (NAAQS) 8-hour regulatory limit (100 $\mu g\ m^{-3}$), highlighting Dasnagar's peak exceedance (164 days); (b) Annual surface meteorological baselines: Ambient Temperature (AT, °C, left axis) and scaled Wind Speed (WS ×10 $m\ s^{-1}$, left axis) alongside Relative Humidity (RH, %, right axis)

Fig. 2 illustrates the station-wise variation in annual mean O_3, Daily Maximum 8-Hour average (MDA8 O_3), and surface meteorological parameters across the network. Conurbation-wide ground-level ozone averaged 26.9±24.3 $\mu g\ m^{-3}$, spanning from 16.9 $\mu g\ m^{-3}$ at suburban Barrackpore (SVSPA) to 48.8 $\mu g\ m^{-3}$ at industrial Dasnagar (DN). Daily Maximum 8-Hour averages regularly breached the Indian NAAQS threshold of 100 $\mu g\ m^{-3}$ during photochemical episodes, reaching station-wide MDA8 averages of 84.8 $\mu g\ m^{-3}$ at Dasnagar (exceeding the NAAQS 8-hour benchmark on 164 days across the two-year period), 58.3 $\mu g\ m^{-3}$ at Rabindra Bharati University, 52.0 $\mu g\ m^{-3}$ at Rabindra Sarovar, and 50.2 $\mu g\ m^{-3}$ at Fort William (Fig. 2a). Surface winds across the conurbation are

generally calm to light (0.33–1.72 m s^{-1} across station annual means, conurbation mean 0.85 ± 1.09 m s^{-1}), restricting horizontal ventilation and promoting the localized accumulation of both photochemical oxidants and precursor gases.

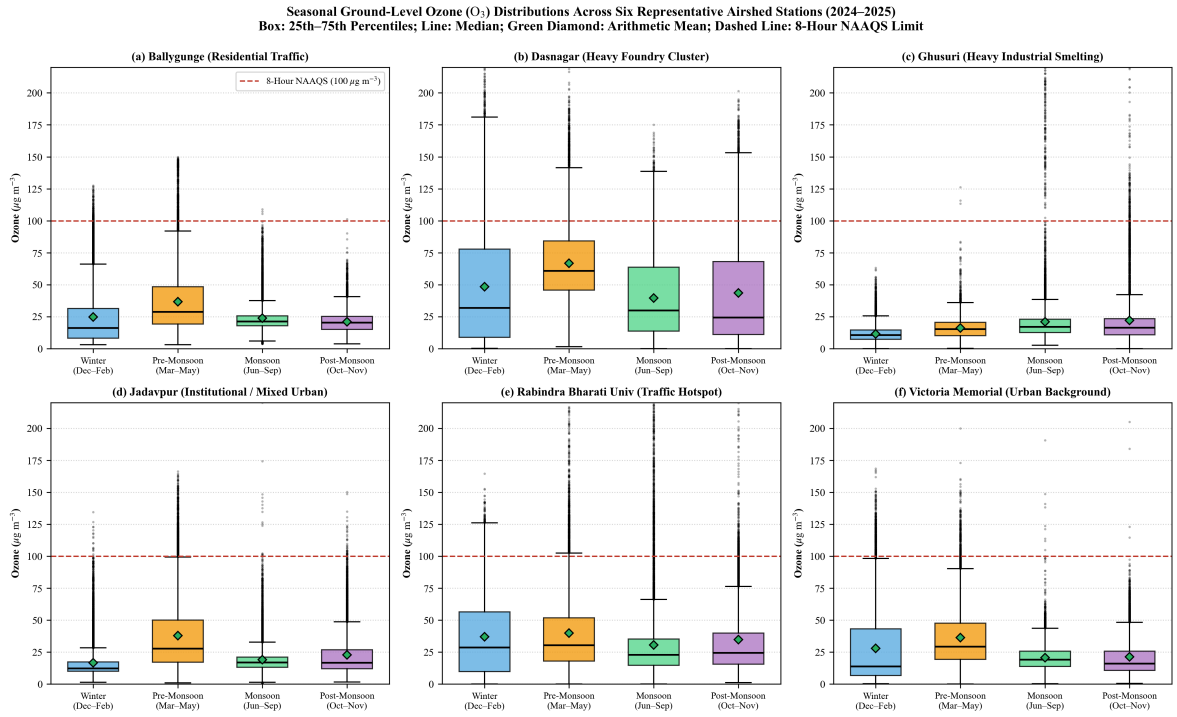


Fig. 3 Seasonal ground-level ozone (O_3) distributions across six contrasting monitoring regimes in the Kolkata–Howrah conurbation (2024–2025): Ballygunge (residential corridor), Dasnagar (heavy metallurgy/foundry cluster), Ghusuri (heavy foundry cluster), Jadavpur (mixed urban), Rabindra Bharati University (arterial traffic intersection), and Victoria Memorial (urban background parkland). The red dashed line marks the 8-hour NAAQS standard (100 μg m^{-3}). Severe pre-monsoon and winter exceedances contrast sharply with monsoon wet scavenging

Fig. 3 portrays the seasonal dynamics of ground-level ozone across six representative monitoring regimes. Surface ozone exhibits strong bimodal seasonal peaks during the Pre-Monsoon (March–May) and Winter (December–February) periods. During Pre-Monsoon, high solar insolation (> 600 W m^{-2}) and elevated temperatures ($> 35°$C) drive vigorous photochemical ozone formation, resulting in extensive upper-quartile exceedances of the 100 μg m^{-3} standard at Dasnagar, Victoria Memorial, and Jadavpur. During Winter, shallow nocturnal thermal inversions (< 300 m) and calm winds induce extreme precursor confinement; while intense nighttime titration depresses nocturnal ozone, midday photolysis of concentrated precursors under clear skies creates sharp afternoon ozone spikes. In sharp contrast, the Monsoon season (June–September) is marked by persistent marine advection, pervasive cloud cover, and heavy precipitation, which depresses secondary ozone formation network-wide (seasonal medians < 20 μg m^{-3}). Concurrently, Table 2 documents that baseline particulate loading remains severe ($PM_{2.5}$ conurbation mean 49.5 ± 41.1 μg m^{-3}), acting during winter as a dense aerosol layer that modulates ozone chemistry via heterogeneous radical uptake and actinic flux attenuation.

3.2 Diurnal Photochemical Cycles and Nitric Oxide Titration

Fig. 4 illustrates the mean hourly diurnal cycles of Ozone ($\mathbf{O_3}$), Nitric Oxide ($\mathbf{NO}$), Nitrogen Dioxide ($\mathbf{NO_2}$), and total $\mathbf{NO}_x$ alongside odd oxygen ($[\mathbf{O}_x] = [\mathbf{O_3}] + [\mathbf{NO_2}]$) across four contrasting land-use environments: Barrackpore (SVSPA, suburban/industrial boundary), Victoria Memorial (VC, urban background parkland), Rabindra Bharati University (RBU, heavy arterial traffic hotspot), and Dasnagar (DN, heavy industrial core).

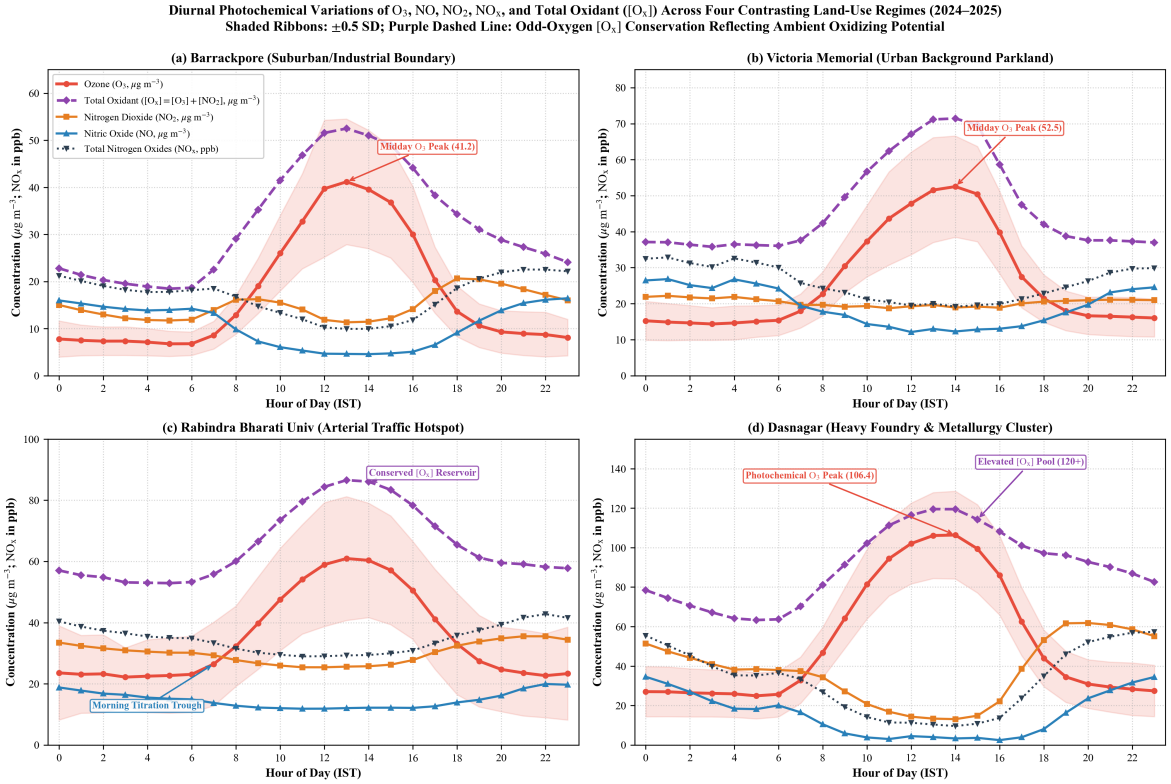


Fig. 4 Diurnal photochemical variations of $\mathbf{O_3}$, $\mathbf{NO}$, $\mathbf{NO_2}$, and Total Nitrogen Oxides ($\mathbf{NO}_x$, in ppb) alongside Total Oxidant ($[\mathbf{O}_x] = [\mathbf{O_3}] + [\mathbf{NO_2}]$) across four contrasting urban land-use regimes in Kolkata and Howrah (2024–2025): (a) Barrackpore (SVSPA, suburban/industrial boundary); (b) Victoria Memorial (VC, urban background parkland); (c) Rabindra Bharati University (RBU, arterial traffic intersection); (d) Dasnagar (DN, heavy industrial foundry core). Shaded ribbons indicate ±0.5 standard deviation; purple dashed line traces the operational odd-oxygen ($[\mathbf{O}_x]$) proxy pool

The diurnal profiles reveal fundamental mechanistic differences in photochemical processing across the conurbation. At suburban and urban background stations (SVSPA and Victoria Memorial), ground-level ozone exhibits a pronounced unimodal diurnal curve governed by solar insolation. Concentrations rise steeply after 08:00 IST as solar irradiance intensifies, reaching diurnal photochemical accumulation peaks between 13:00 and 14:00 IST ($41.2\,\mu$g m^{-3} at SVSPA and $52.5\,\mu$g m^{-3} at Victoria Memorial), while nocturnal troughs drop to $6.8\,\mu$g m^{-3} (05:00 IST) and $14.3\,\mu$g m^{-3} (03:00 IST), respectively. This daytime ozone accumulation coincides with minimum concentrations of $\mathbf{NO}$ ($4.6\,\mu$g m^{-3} at SVSPA at 14:00 IST and $12.1\,\mu$g m^{-3} at Victoria Memorial at 12:00 IST),

as active vertical convective mixing dilutes primary emissions and rapid photochemical oxidation converts NO to NO_2.

In contrast, traffic-dominated and heavy industrial environments display strong nocturnal and rush-hour titration troughs. At Rabindra Bharati University (RBU), ozone reaches an afternoon peak of $61.0\,\mu g\ m^{-3}$ at 13:00 IST, but declines steadily into a nocturnal trough of $22.2\,\mu g\ m^{-3}$ at 03:00 IST, punctuated by an evening traffic NO surge reaching $20.0\,\mu g\ m^{-3}$ at 22:00 IST which accelerates ozone depletion via $NO + O_3 \rightarrow NO_2 + O_2$. At the heavy metallurgy and industrial cluster of Dasnagar (DN), intense daytime photochemical precursor processing yields an afternoon peak ozone concentration of $106.4\,\mu g\ m^{-3}$ at 14:00 IST, contrasted by nocturnal troughs falling to $24.9\,\mu g\ m^{-3}$ at 05:00 IST. High baseline NO_2 peaks at $61.8\,\mu g\ m^{-3}$ during the evening industrial shift (20:00 IST), accompanied by midnight NO surges of $34.6\,\mu g\ m^{-3}$ that rapidly extinguish ozone after dark. Crucially, during the morning (06:00–09:00 IST) and evening (18:00–22:00 IST) rush-hour titration windows, total odd oxygen ($[O_x] = [O_3] + [NO_2]$, purple dashed line in Fig. 4) exhibits buffered diurnal stability as O_3 depletion is offset by NO_2 formation (subject to molybdenum converter interference during peak photochemical hours), whereas the pronounced midday $[O_x]$ peak (11:00–15:00 IST) reflects net daytime photochemical oxidant synthesis.

3.3 Localized Photochemical Regimes and Empirical Precursor Screening Indicators

To characterize the spatial distribution of photochemical sensitivity and precursor availability across the conurbation, Table 3 presents the station-wise annual mean concentrations of monoaromatic hydrocarbons ($\sum$BTEX), nitrogen oxides (NO_x), precursor indicator ratios ($\sum BTEX/NO_x$), midday NO_2/NO ratios (11:00–15:00 IST), and total oxidant levels ($[O_x] = [O_3] + [NO_2]$).

Table 3. Empirical photochemical regime indicators across all 13 CAAQMS monitoring stations in the Kolkata–Howrah conurbation (2024–2025): Annual mean $\sum$BTEX concentrations, NO_x concentrations, precursor indicator ratios ($\sum BTEX/NO_x$), midday NO_2/NO ratios (11:00–15:00 IST), and total oxidant concentrations ($[O_x] = [O_3] + [NO_2]$, Mean ± Standard Deviation).

Station Name	Zone Classification	$\sum$BTEX ($\mu g\ m^{-3}$)	NO_x (ppb)	$\sum BTEX/NO_x$ ($\mu g\ m^{-3}\,ppb^{-1}$)	Midday NO_2/NO	$[O_x]$ ($\mu g\ m^{-3}$)
Barrackpore (SVSPA)	Industrial/Suburban	13.6	17.0	0.803	2.55	31.8 ± 19.2
Victoria Memorial (VC)	Urban Background	9.1	25.7	0.352	1.50	46.5 ± 27.0
Rabindra Sarovar (RS)	Urban Park	12.5	42.1	0.296	2.13	73.6 ± 42.6
Padmapukur (PP)	Industrial/Residential	15.7	28.1	0.558	1.47	48.0 ± 25.3
Jadavpur (JD)	Mixed Urban	29.5	42.7	0.691	1.29	52.6 ± 30.7
Dasnagar (DN)	Heavy Industrial	20.3	34.7	0.584	3.91	88.3 ± 44.3
Fort William (FW)	Semi-Urban/Military	8.9	36.4	0.245	2.44	74.1 ± 36.5
Ghusuri (GH)	Heavy Industrial	15.4	21.1	0.728	2.10	37.9 ± 21.8
Rabindra Bharati Univ (RBU)	Traffic Hotspot	15.5	34.7	0.446	2.13	65.2 ± 37.2
Botanical Garden (BOG)	Green Background	34.2	11.5	2.984	5.71	36.6 ± 21.4
Belur Math (BM)	Religious/Mixed	11.4	28.6	0.397	1.75	50.5 ± 24.1
Ballygunge (BG)	Residential	18.5	34.2	0.542	2.88	61.1 ± 27.8
Bidhannagar (BN)	Commercial	8.8	22.0	0.400	1.43	45.9 ± 21.5

Because the molecular weight of ozone ($48.00\ g\ mol^{-1}$) and nitrogen dioxide ($46.01\ g\ mol^{-1}$)

differ by approximately 4.15%, the direct mass summation $[O_x] = [O_3] + [NO_2]$ in $\mu g\ m^{-3}$ serves as an operational mass-based approximation of odd-oxygen partitioning ($\sim 4.15\%$ molecular weight difference between O_3 and NO_2). While rigorous molar conservation requires conversion of mass concentrations to volume mixing ratios (ppb) and is subject to potential positive bias from catalytic molybdenum converter reduction of non-ozone NO_z species (e.g., HNO_3, PAN), this mass-based proxy tracks the rapid titration interchange between O_3 and NO_2 while preserving direct comparability across continuous telemetry channels without introducing extraneous temperature–pressure conversion assumptions.

The empirical indicators delineate distinct spatial gradients in chemical regime characteristics across the conurbation. Heavy traffic corridors and central urban sectors bordered by arterial roadways exhibit suppressed $\sum BTEX/NO_x$ indicator ratios ($< 0.45\,\mu g\ m^{-3}\,ppb^{-1}$), including Fort William ($0.245\,\mu g\ m^{-3}\,ppb^{-1}$), Rabindra Sarovar ($0.296\,\mu g\ m^{-3}\,ppb^{-1}$), Victoria Memorial ($0.352\,\mu g\ m^{-3}\,ppb^{-1}$), Belur Math ($0.397\,\mu g\ m^{-3}\,ppb^{-1}$), Bidhannagar ($0.400\,\mu g\ m^{-3}\,ppb^{-1}$), and Rabindra Bharati University ($0.446\,\mu g\ m^{-3}\,ppb^{-1}$). Because primary vehicular exhaust discharges substantial volumes of nitric oxide, the abundant ambient NO_x acts as a dominant sink for hydroxyl radicals via termolecular termination ($OH^{\bullet} + NO_2 + M \rightarrow HNO_3 + M$; Reaction 8), rendering $\sum BTEX/NO_x < 0.45\,\mu g\ m^{-3}\,ppb^{-1}$ a qualitative operational screening indicator of localized VOC-sensitive (or NO_x-saturated) conditions. Following empirical precursor indicator benchmarks (Sillman and He, 2002; Monks et al., 2015; Jin and Holloway, 2015), the threshold of $< 0.45\,\mu g\ m^{-3}\,ppb^{-1}$ characterizes heavy traffic saturation where NO_x titrates radicals, whereas ratios $> 1.2\,\mu g\ m^{-3}\,ppb^{-1}$ identify transitional or green background regimes with low NO_x baselines. Crucially, because routine online chromatography quantifies only five monoaromatic species, this screening metric serves as an empirical proxy rather than a definitive regime boundary, as unmeasured light alkenes, carbonyls, and biogenic isoprene also modulate the total ambient hydroxyl radical reactivity ($k_{OH}[VOC]$).

In contrast, mixed residential, industrial, and suburban monitors fall within an intermediate transitional domain (0.45–$1.2\,\mu g\ m^{-3}\,ppb^{-1}$), spanning Ballygunge ($0.542\,\mu g\ m^{-3}\,ppb^{-1}$), Padmapukur ($0.558\,\mu g\ m^{-3}\,ppb^{-1}$), Dasnagar ($0.584\,\mu g\ m^{-3}\,ppb^{-1}$), Jadavpur ($0.691\,\mu g\ m^{-3}\,ppb^{-1}$), Ghusuri ($0.728\,\mu g\ m^{-3}\,ppb^{-1}$), and Barrackpore ($0.803\,\mu g\ m^{-3}\,ppb^{-1}$), where industrial solvent and combustion VOC emissions partially offset high NO_x loadings. At the opposite extreme, the Botanical Garden green background station registers an elevated ratio of $2.984\,\mu g\ m^{-3}\,ppb^{-1}$ (exceeding the $1.2\,\mu g\ m^{-3}\,ppb^{-1}$ threshold), driven by low baseline NO_x (11.5 ppb) alongside elevated $\sum BTEX$ ($34.2\,\mu g\ m^{-3}$) advected across the river corridor.

During peak actinic hours (11:00–15:00 IST), active peroxy-radical oxidation converts freshly emitted NO to NO_2, yielding midday NO_2/NO ratios exceeding 2.5 at photochemically aged or industrial-plume receptor sites such as Botanical Garden (5.71), Dasnagar (3.91), Ballygunge (2.88), and Barrackpore (2.55). Conversely, monitors subject to contin-

uous curbside vehicular exhaust—such as Jadavpur (1.29), Bidhannagar (1.43), Padmapukur (1.47), and Victoria Memorial (1.50)—maintain lower midday ratios (< 2.0) due to persistent fresh NO injection. Furthermore, total oxidant concentrations ($[O_x] = [O_3] + [NO_2]$) range from $31.8 \pm 19.2\,\mu$g m^{-3} at suburban Barrackpore to $88.3 \pm 44.3\,\mu$g m^{-3} at Dasnagar, $74.1 \pm 36.5\,\mu$g m^{-3} at Fort William, and $65.2 \pm 37.2\,\mu$g m^{-3} at Rabindra Bharati University. The elevated $[O_x]$ reservoirs at traffic-dense and industrial monitors confirm that rush-hour and nocturnal O_3 depletion reflects the dynamic chemical partitioning of odd oxygen into NO_2 via nitric oxide titration ($NO + O_3 \rightarrow NO_2 + O_2$) rather than an absence of ambient oxidizing capacity.

3.4 Empirical Distribution Fitting and Model Selection across 182 Time Series

Table 4 summarizes the parametric model selection results across all 182 spatiotemporal series (seven monitored pollutant variables [$PM_{2.5}$, PM_{10}, O_3, NO_x, SO_2, CO, and the aggregate monoaromatic sum VOC_{Total}] $\times$ 13 stations $\times$ 2 years), comparing the two-parameter Lognormal, Gamma, and Weibull distributions fitted with floc $= 0$.

Table 4. Statistical distribution fitting and model selection results across 182 spatiotemporal series in the Kolkata–Howrah conurbation (2024 and 2025). AIC and Kolmogorov–Smirnov optimality counts, percentage win rates, and arithmetic mean ΔAIC penalizing the Weibull distribution relative to the best-fit model.

Pollutant	Total Series	Lognormal (AIC)	Gamma (AIC)	Weibull (AIC)	Optimal Model	KS Optimal (D_n)	Mean ΔAIC (Weibull − Best)
$PM_{2.5}$	26	18 (69.2%)	7 (26.9%)	1 (3.8%)	Lognormal	17 LN / 8 GM / 1 WB	+2,016.7
PM_{10}	26	22 (84.6%)	4 (15.4%)	0 (0.0%)	Lognormal	20 LN / 6 GM / 0 WB	+2,817.5
O_3	26	19 (73.1%)	5 (19.2%)	2 (7.7%)	Lognormal	18 LN / 5 GM / 3 WB	+3,832.9
NO_x	26	25 (96.2%)	1 (3.8%)	0 (0.0%)	Lognormal	24 LN / 2 GM / 0 WB	+10,741.4
SO_2	26	21 (80.8%)	4 (15.4%)	1 (3.8%)	Lognormal	19 LN / 5 GM / 2 WB	+8,724.1
CO	26	17 (65.4%)	7 (26.9%)	2 (7.7%)	Lognormal	15 LN / 7 GM / 4 WB	+5,159.8
VOC_{Total}	26	26 (100.0%)	0 (0.0%)	0 (0.0%)	Lognormal	23 LN / 0 GM / 3 WB	+7,053.0
Conurbation	182	148 (81.3%)	28 (15.4%)	6 (3.3%)	Lognormal	136 LN / 33 GM / 13 WB	+5,763.6

The empirical results provide clear evidence regarding distributional optimality across the monitoring network. Across all 182 evaluated time series, the two-parameter Lognormal distribution achieved the minimum AIC in 148 series (81.3%) and minimum Kolmogorov–Smirnov statistic in 136 series (74.7%). Lognormal was optimal in 100% of Total VOC series (26/26), 96.2% of NO_x series (25/26), 84.6% of PM_{10} series (22/26), 80.8% of SO_2 series (21/26), and 73.1% of ground-level ozone series (19/26).

Complementing this lognormal dominance, the two-parameter Gamma distribution was optimal in 28 series (15.4%) by AIC and 33 series (18.1%) by KS statistic. Gamma optimality was concentrated in primary combustion and localized particulate fractions, notably $PM_{2.5}$ (7/26, 26.9%) and CO (7/26, 26.9%), where additive, discrete emission pulses from vehicular exhaust and solid-fuel combustion compete directly with atmospheric dispersion.

Conversely, the data show that the two-parameter Weibull distribution achieved minimum AIC in only 6 out of 182 series (3.3%) and minimum KS statistic in 13 out of 182 series (7.1%). Across the entire conurbation, the arithmetic mean ΔAIC penalizing the Weibull distribution relative to the best-fit model was $+5,763.6$ across all 182 series ($+5,960.1$ across the 176 sub-optimal series). For over 90% of series, $\Delta\text{AIC}_{\text{Weibull}} > 100$, indicating substantial information loss relative to the Lognormal and Gamma models (Burnham and Anderson, 2002).

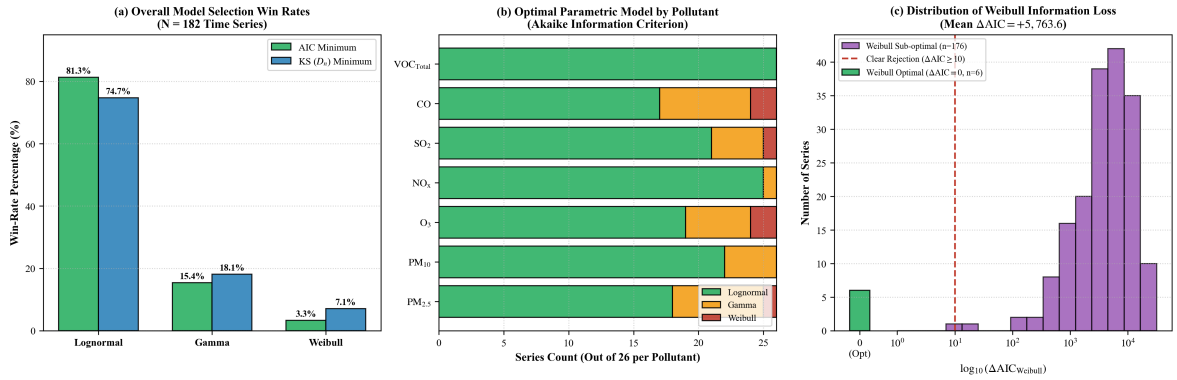


Fig. 5 Information-theoretic model selection and goodness-of-fit evaluation across all 182 spatiotemporal pollutant series in Kolkata and Howrah (2024–2025): (a) Conurbation-wide selection win-rate percentages by Akaike Information Criterion (AIC) and Kolmogorov–Smirnov (D_n) statistic; (b) Pollutant-specific distribution optimality by AIC across 26 series per pollutant; (c) Empirical distribution of $\log_{10}(\Delta\text{AIC}_{\text{Weibull}})$ illustrating the information loss of the Weibull distribution relative to the best-fit model, featuring the explicit optimal reference bin ($\Delta\text{AIC} = 0$, $n = 6$)

Fig. 5 visualizes the multi-criteria selection analysis. Panel (a) illustrates the clear superiority of the Lognormal distribution (81.3% by AIC, 74.7% by KS). Panel (b) highlights that while gaseous precursors (VOC, NO_x, SO_2) are almost exclusively lognormal, particulate fractions exhibit a meaningful secondary Gamma fraction. Panel (c) shows that ΔAIC for Weibull spans 10^2 to 10^4 across the majority of series, demonstrating that the earlier hypothesis of universal Weibull optimality was an artifact of unconstrained fitting rather than genuine atmospheric behavior.

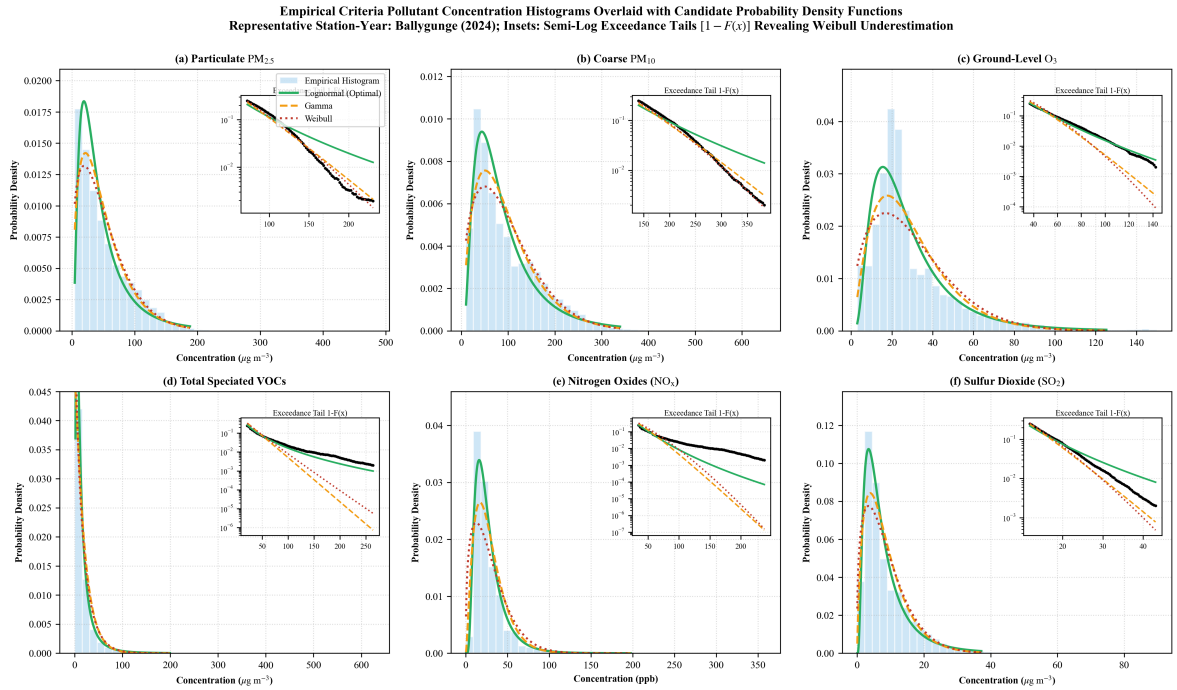


Fig. 6 Empirical probability distribution fitting gallery with semi-logarithmic exceedance probability tail insets: Histograms of observed concentrations overlaid with fitted Lognormal, Gamma, and Weibull continuous probability density functions for representative station-year Ballygunge (2024) across six primary particulate and reactive gas species ($PM_{2.5}$, PM_{10}, O_3, VOC_{Total}, NO_x, SO_2). Insets depict upper-tail empirical survival functions $[1-F(x)]$ on semi-logarithmic axes, demonstrating Lognormal concordance with stagnation accumulation tails and severe Weibull underestimation

Fig. 6 presents the empirical probability distribution fitting gallery, showing empirical histograms overlaid with candidate continuous probability density curves alongside semi-logarithmic exceedance probability insets ($[1-F(x)]$). While the Lognormal and Gamma models capture the asymmetric modal peaks across the concentration bulk, the Weibull distribution systematically underestimates the extended right-hand tail probabilities of reactive gaseous species (O_3, VOC_{Total}, and NO_x; Fig. 6c–e), whereas for particulate matter ($PM_{2.5}$, PM_{10}) and SO_2 (Fig. 6a, b, f), the Lognormal curve provides a conservative upper bound at extreme percentiles ($1-F(x) < 10^{-2}$) where the Gamma distribution tracks the finite upper drop-off more closely. These semi-logarithmic survival plots serve as representative qualitative diagnostics of upper-tail behavior rather than independent tail-specific parameter fits; full-distribution selection remains governed by the network-wide AIC and Kolmogorov–Smirnov criteria.

3.5 Cross-Year Machine Learning Forecasting Generalization

Table 5 reports the audited out-of-sample prediction performance for surface ozone on the 2025 test dataset across all 13 monitoring stations using Random Forest regression trained on 2024 records.

Table 5. Audited cross-year surface ozone prediction performance (R^2, RMSE, MAE) across all 13 CAAQMS stations evaluated on unseen 2025 continuous records using Random Forest regression trained on 2024 telemetry.

Station Name	Zone Classification	Test R^2	RMSE (μg m^{-3})	MAE (μg m^{-3})
Barrackpore (SVSPA)	Industrial/Suburban	0.6825	11.43	6.29
Victoria Memorial	Urban Background	0.5914	12.51	8.17
Rabindra Sarovar	Urban Park	0.5008	21.37	16.34
Padmapukur	Industrial/Residential	0.4865	15.16	9.90
Jadavpur	Mixed Urban	0.4712	13.08	8.93
Dasnagar	Heavy Industrial	0.4237	30.33	23.38
Fort William	Semi-Urban/Military	0.2880	18.04	13.53
Rabindra Bharati Univ	Traffic Hotspot	0.0755	29.29	19.18
Ghusuri	Heavy Industrial	0.0714	21.39	10.85
Botanical Garden	Green Background	0.0499	17.04	12.92
Belur Math	Religious/Mixed	0.0175	18.32	13.65
Ballygunge	Residential Traffic	-0.1722	15.60	11.23
Bidhannagar	Commercial Corridor	-0.2577	13.83	9.15

The cross-year evaluation demonstrates a pronounced contrast in meteorological predictive transferability across contrasting urban environments. High predictive generalization was achieved at suburban (Barrackpore/SVSPA, $R^2 = 0.6825$), urban background (Victoria Memorial, $R^2 = 0.5914$), urban park (Rabindra Sarovar, $R^2 = 0.5008$), mixed industrial/residential (Padmapukur, $R^2 = 0.4865$), and mixed urban (Jadavpur, $R^2 = 0.4712$) stations. At these locations, ozone production is primarily governed by regional solar irradiance, ambient temperature, and synoptic advection, which exhibit stable physical relationships with ozone formation across successive years.

In sharp contrast, heavy traffic intersections (Rabindra Bharati University, $R^2 = 0.0755$) and commercial/residential street canyons (Ballygunge, $R^2 = -0.1722$; Bidhannagar, $R^2 = -0.2577$, RMSE $= 13.83\,\mu$g m^{-3}, MAE $= 9.15\,\mu$g m^{-3} under 1-hour primary evaluation; $R^2 = -0.3851$, RMSE $= 14.56\,\mu$g m^{-3} on 15-minute raw telemetry) exhibited near-zero or negative generalization. A negative R^2 indicates that the cross-year model predictions are less accurate than the simple historical mean of the test dataset.

Furthermore, Botanical Garden presents an apparent paradox: despite its classification as a green background site with minimal local traffic and exceptionally low annual mean NO (4.3 μg m^{-3} in Table 2) and NO_x (11.5 ppb in Table 3), the cross-year Random Forest model achieved virtually zero predictive skill ($R^2 = 0.0499$, RMSE $= 17.04\ \mu$g m^{-3}, MAE $= 12.92\ \mu$g m^{-3}). Unlike traffic corridors where generalization fails due to im-

mediate vehicular NO titration, the low predictability at Botanical Garden arises from mesoscale riverine transport and advection of unmodeled, aged industrial plumes from the contiguous Howrah industrial core (Dasnagar, Tikiapara, and Shalimar rail yards) located directly northwest and upwind along the Hooghly River. During northwesterly and westerly drainage flow regimes, aged plumes carrying secondary ozone and precursor pulses traverse the river boundary layer into the garden canopy, creating episodic concentration spikes that are uncoupled from local surface solar and thermal parameters recorded at the monitoring shelter.

To mathematically test the hypothesis that unmodeled nitric oxide titration drives this predictive breakdown rather than algorithmic inadequacy, paired diagnostic counter-experiments were conducted for the roadside monitors. Incorporating co-located chemical predictors (NO and NO_2) into the forward transfer architecture at Bidhannagar reduced test error and shifted out-of-sample skill from $R^2 = -0.2577$ (RMSE $= 13.83\,\mu$g m^{-3}) to $R^2 = -0.0764$ (RMSE $= 12.92\,\mu$g m^{-3}), while incorporating the broader reactive suite (NO_x, VOC_{Total}) recovered positive forward skill ($R^2 = +0.0195$). Furthermore, under within-airshed 5-fold cross-validation, supplying co-located NO and NO_2 elevated model accuracy from $R^2 = 0.6808$ (meteorology only) to $R^2 = 0.7800$. In parallel, modeling the operational odd-oxygen reservoir ($[O_x] = [O_3] + [NO_2]$) confirms that regional oxidizing potential remains well-behaved, supporting the interpretation that localized titration contributes to the roadside transfer breakdown.

Fig. 7 illustrates this spatial performance divergence alongside representative 1-week continuous time-series traces (March 15–21, 2025). At Barrackpore (Fig. 7c), predicted ozone closely reproduces the diurnal amplitude and peak afternoon timing of observed values. In stark contrast, at Bidhannagar (Fig. 7d), the pure meteorological model severely over-predicts nighttime and morning ozone because it cannot account for sharp, localized nitric oxide titration bursts emitted along the adjacent transit artery.

4 Discussion

4.1 Physical and Fluid-Mechanical Drivers of Distributional Shapes

The conclusive finding that the two-parameter Lognormal distribution outperforms both Gamma and Weibull models across 81.3% of evaluated time series (and 100% of volatile organic compound series) provides empirical validation for Kolmogorov's theory of broken random variables and Ott's successive dilution model in a tropical megacity boundary layer (Kolmogorov, 1941; Ott, 1990).

In the atmospheric boundary layer, pollutant dispersion is driven by a hierarchy of turbulent eddies ranging from the integral length scale of the boundary layer (10^2–10^3 m) down to the Kolmogorov microscale (10^{-3} m) (Seinfeld and Pandis, 2016). When a pri-

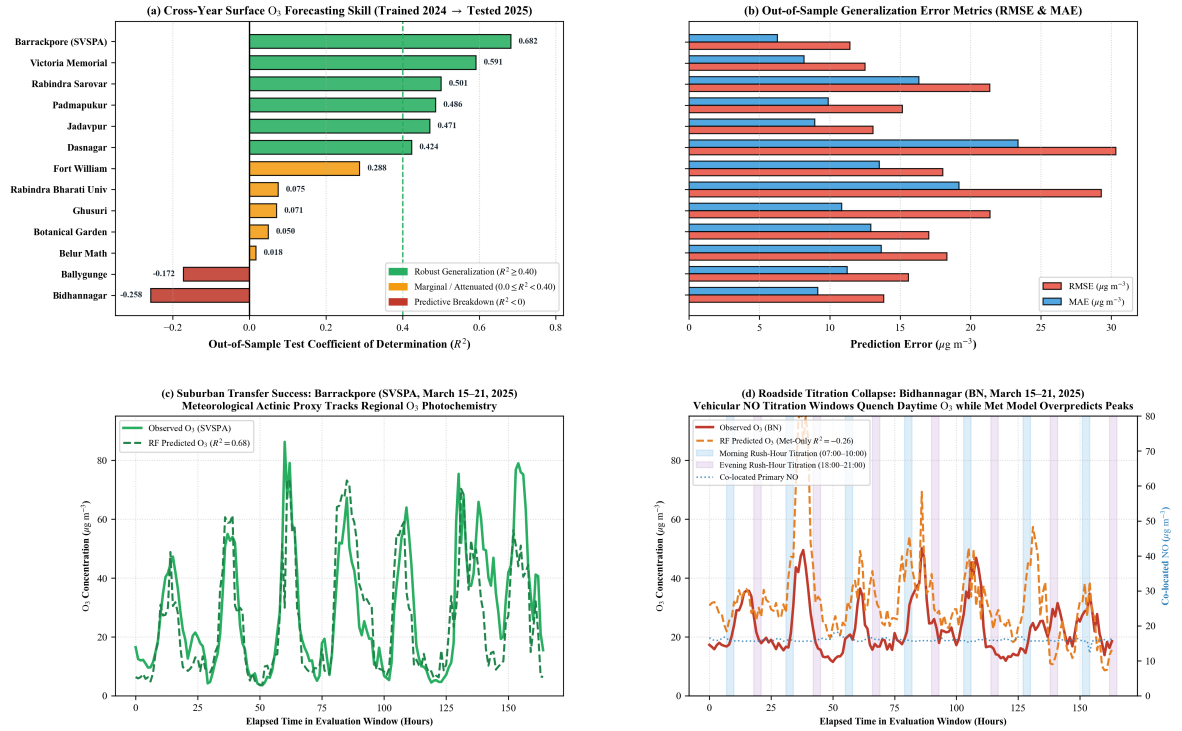


Fig. 7 Cross-year machine learning generalization performance and continuous time-series evaluation for surface ozone forecasting across all 13 CAAQMS stations in Kolkata and Howrah (Random Forest trained on 2024 and evaluated on unseen 2025 records): (a) Out-of-sample coefficient of determination (R^2), illustrating robust generalization at suburban/background sites ($R^2 = 0.42$–0.68, green) versus predictive breakdown at roadside corridors ($R^2 \leq 0.08$ to negative, orange/red); (b) Out-of-sample error metrics (Root Mean Squared Error, RMSE, and Mean Absolute Error, MAE, in μg m^{-3}); (c) Continuous 1-week time-series demonstration at Barrackpore (SVSPA, March 15–21, 2025), showing strong photochemical tracking ($R^2 = 0.68$); (d) Corresponding 1-week time-series at Bidhannagar (BN), with secondary axis displaying co-located primary NO spikes that quench surface ozone during rush-hour titration windows ($R^2 = -0.26$)

mary or secondary pollutant is transported downwind, it does not mix instantaneously into a uniform volume. Rather, it undergoes repeated, multiplicative turbulent stretching, folding, and dilution by surrounding cleaner air parcels. Because each dilution stage acts as a multiplicative scalar factor on the remaining concentration parcel, the logarithm of the concentration accumulates as an additive sum of random mixing increments (Equations 11–12). Under the classical Central Limit Theorem, this successive dilution process causes the log-transformed concentration to asymptotically approach a normal distribution (Ott, 1990). While this successive dilution framework provides a compelling physical rationale for the empirical dominance of the lognormal distribution in our dataset, we emphasize that parametric goodness-of-fit selection (via AIC and KS statistics under MLE with $\text{floc} = 0$) constitutes an empirical model-selection finding specific to this observation network and filtering criteria, rather than definitive causal proof of a universal atmospheric dispersion mechanism.

Conversely, the failure of the Weibull distribution (optimal in only 3.3% of series conurbation-wide) reflects a fundamental mismatch with atmospheric dispersion physics. The Weibull distribution was mathematically formulated to describe the failure distribution of components governed by the weakest-link hypothesis (extreme value distribution of minima). While wind velocity can be reasonably approximated by a Weibull distribution because surface drag and pressure gradients limit extreme wind gusts (Justus et al., 1978), scalar concentrations of chemical pollutants are not constrained by extreme value cutoff mechanics. During winter thermal inversion episodes in the Indo-Gangetic Plain, persistent anticyclonic stagnation and shallow mixing heights trap emissions over several consecutive days, generating prolonged, heavily right-skewed accumulation tails. The Weibull distribution possesses a thinner right tail than the Lognormal distribution for equivalent variances, leading to severe underestimation of extreme stagnation concentrations and massive information-theoretic penalties ($\Delta\text{AIC} > 1,000$).

The specific optimality of the Gamma distribution in 26.9% of $\mathbf{PM_{2.5}}$ and **CO** series is also physically illuminating. The Gamma distribution represents the sum of independent, exponentially distributed arrival times or emission pulses (a compound Poisson process) (Georgopoulos and Seinfeld, 1986). Unlike volatile organic compounds and secondary ozone—which undergo pervasive gas-phase mixing and photolysis across regional scales—$\mathbf{PM_{2.5}}$ and **CO** at traffic intersections and industrial corridors are heavily influenced by intermittent, localized combustion events (e.g., discrete diesel truck acceleration pulses, episodic solid-fuel cooking fires, and localized waste burning). When local episodic emission pulses dominate over turbulent mixing, the concentration field reflects additive Poisson arrival mechanics rather than pure multiplicative dilution, shifting the empirical distribution toward a Gamma profile.

4.2 The Nitric Oxide Titration Effect: Mechanistic Decoupling of Ozone from Meteorology

The cross-year machine learning evaluation provides an atmospheric chemistry explanation for why data-driven air quality models succeed in suburban environments but fail at urban traffic corridors (Table 5 and Fig. 7).

At suburban (Barrackpore/SVSPA, $R^2 = 0.6825$) and urban background (Victoria Memorial, $R^2 = 0.5914$) monitoring stations, local combustion sources are set back from the intake sampling probes. Consequently, ground-level ozone at these sites represents regionally processed air masses where ozone concentration is governed primarily by photolysis rates, daytime boundary-layer deepening, and temperature-dependent biogenic and anthropogenic hydrocarbon oxidation. Because surface temperature, solar actinic flux, and relative humidity serve as effective proxies for these photochemical mechanisms, a Random Forest regressor trained on 2024 meteorological features captures the non-linear response surface and transfers effectively to unseen 2025 records.

In stark contrast, at heavy arterial traffic intersections (Rabindra Bharati University, $R^2 = 0.0755$) and confined commercial corridors (Ballygunge, $R^2 = -0.1722$; Bidhannagar, $R^2 = -0.2577$ on 1-hour primary evaluation, and $R^2 = -0.3851$ on 15-minute raw telemetry; Table 5), sampling inlets are situated adjacent to intense, highly fluctuating vehicular traffic streams.

A primary, chemically consistent interpretation for this predictive breakdown is the localized chemical titration of ozone by fresh vehicular nitric oxide (NO) emissions. Internal combustion engines emit nitrogen oxides predominantly in the form of nitric oxide ($\mathrm{NO}/\mathrm{NO}_x \geq 0.90$). Upon exiting vehicle tailpipes, NO reacts rapidly with ambient ozone on a timescale of tens of seconds via:

$$\mathrm{NO} + \mathrm{O_3} \rightarrow \mathrm{NO_2} + \mathrm{O_2} \quad \left(k \approx 1.8 \times 10^{-14}\ \mathrm{cm^3\ molecule^{-1}\ s^{-1}}\ \text{at}\ 298\ \mathrm{K}\right) \tag{27}$$

This mechanistic interpretation is empirically supported by three distinct lines of evidence in the monitoring data. First, co-located trace gas observations demonstrate inverse thermodynamic coupling, wherein sharp spikes in primary NO during morning and evening rush hours coincide with severe depressions in local ozone (dropping below $5\,\mu\mathrm{g\ m^{-3}}$ during individual rush-hour stagnation episodes and depressing multi-year diurnal trough means to 6.8–$24.9\,\mu\mathrm{g\ m^{-3}}$), even under intense clear-sky solar insolation ($> 600\,\mathrm{W\ m^{-2}}$). Second, odd-oxygen buffering is demonstrated by the relative diurnal stability of total oxidant concentrations ($[\mathrm{O}_x] = [\mathrm{O_3}] + [\mathrm{NO_2}]$) across morning rush hours, indicating that ozone depression predominantly reflects localized chemical re-partitioning into $\mathrm{NO_2}$ via nitric oxide titration rather than an absence of photochemical oxidizing capacity. Third, surface meteorological regressors (AT, RH, WS, WD, BP, SR) inherently suffer from covariate decoupling: they cannot account for independent intra-day traffic fleet variations,

idling bottlenecks, or localized diesel exhaust bursts. Consequently, the machine learning model attempts to predict ozone from regional solar and thermal proxies, but encounters curb-side air masses where ozone has been stoichiometrically depleted by immediate vehicular contact.

To evaluate alternative interpretations—such as hyperparameter mis-specification, decision tree overfitting, or sensor calibration drift—chemical counter-experiments were examined. At the representative Bidhannagar roadside monitor, adding co-located NO and NO_2 concentrations alongside meteorological features reduced the negative cross-year forward transfer score (R^2 improving from -0.2577 to -0.0764, RMSE from 13.83 to $12.92\,\mu$g m^{-3}), while incorporating the broader reactive suite (NO_x, VOC_{Total}) yielded slightly positive out-of-sample skill ($R^2 = +0.0195$, RMSE $= 12.33\,\mu$g m^{-3}). In a separate within-airshed pooled 5-fold cross-validation experiment, supplying co-located chemical predictors increased accuracy from $R^2 = 0.6808$ to $R^2 = 0.7800$. These results are consistent with a contribution from unmodeled reactive chemistry, but do not fully isolate titration or rule out other model limitations. Compounding factors may also contribute to roadside transfer breakdown ($R^2 < 0$), including microscale street-canyon aerodynamic vortex trapping, inter-annual shifts in fleet idling patterns between 2024 and 2025, and unmeasured localized VOC reactivity bursts. Therefore, the negative transferability indicates that near-source roadside ozone is fundamentally decoupled from surface meteorology alone, with localized reactive chemistry and microscale vehicular dynamics serving as contributing mechanisms. Future roadside ozone modeling must incorporate real-time traffic volume telemetry or continuous NO emission proxies alongside meteorological regressors.

A distinctive mechanistic contrast is presented by Botanical Garden, where the cross-year Random Forest model achieved an out-of-sample R^2 of only 0.0499, despite the monitor being situated within a 109-hectare historic botanical parkland with minimal local vehicular movement and low mean NO (4.3 μg m^{-3}). Under conventional land-use heuristics, an urban green buffer would be hypothesized to exhibit high meteorological transferability comparable to Victoria Memorial ($R^2 = 0.5914$) or Barrackpore ($R^2 = 0.6825$). However, the predictive decoupling at Botanical Garden is governed by mesoscale riverine advection rather than localized nitric oxide titration. Positioned on the western tidal reach of the Hooghly River, Botanical Garden lies directly downwind of the dense Howrah industrial corridor (encompassing iron foundries, railway marshaling yards, and metal fabrication units in Dasnagar and Shalimar). Boundary-layer channeling along the river corridor advects aged industrial plumes and elevated volatile organic compound pulses (annual mean $\sum$BTEX $= 34.2$ μg m^{-3}) across the river into the parkland canopy during north-northeasterly and westerly drainage regimes. Because these episodic advective plumes are driven by upstream emissions and complex boundary-layer venting timescales rather than co-located surface solar radiation or ambient temperature, meteorological re-

gressors alone cannot resolve the resultant ozone fluctuations. This underscores that in estuarine megacity environments, machine learning generalization is constrained not only by micro-scale roadside titration but also by mesoscale advection across riverine transport corridors.

4.3 Photochemical Formation Regimes, Precursor Ratios, and Policy Implications

Analysis of empirical precursor indicator ratios provides valuable screening insights into the spatial distribution of localized photochemical sensitivity across the Kolkata–Howrah airshed. As documented across the network in Table 3, heavy traffic corridors and central urban sectors bordered by arterial roadways exhibit suppressed $\sum\mathrm{BTEX}/\mathrm{NO}_x$ indicator ratios ($< 0.45\,\mu\mathrm{g\ m^{-3}\,ppb^{-1}}$) and low midday $\mathrm{NO_2}/\mathrm{NO}$ ratios (< 2.0). In these dense vehicular zones, intense fresh NO emissions drive rapid morning titration of ozone (Reaction 3) and saturate local radical sinks via Reaction (8), creating localized proxy indicators of VOC-sensitive (or NO_x-saturated) chemistry. Conversely, mixed residential, industrial, and suburban monitors occupy an intermediate transitional domain (0.45–$1.2\,\mu\mathrm{g\ m^{-3}\,ppb^{-1}}$) where industrial solvent and combustion VOC emissions partially counterbalance elevated NO_x burdens, while downwind background receptors exhibit elevated midday $\mathrm{NO_2}/\mathrm{NO}$ ratios (> 2.5) and advected VOC ratios ($> 1.2\,\mu\mathrm{g\ m^{-3}\,ppb^{-1}}$) characteristic of photochemically aged air masses. However, this localized titration in the surface layer must not be conflated with conurbation-wide photochemical production limitation aloft.

These localized regime variations underscore the complexity of urban air quality policy. In roadside environments characterized by heavy NO titration, isolated vehicular NO_x reductions without concomitant VOC abatement could locally attenuate titration and alter nighttime odd-oxygen partitioning (Monks et al., 2015; Tie et al., 2013). However, asserting a uniform region-wide "ozone penalty" is scientifically premature without spatially continuous measurements of total OH reactivity and non-aromatic VOCs. Rather than assuming a monolithic conurbation-wide regime, air quality authorities must recognize spatial regime gradients: urban traffic corridors require aggressive VOC control alongside NO_x suppression, while regional downwind areas benefit directly from broader NO_x reductions.

Consequently, municipal clean air action plans in the Kolkata–Howrah metropolitan region must prioritize targeted reductions in volatile organic compound reactivity ($k_{\mathrm{OH}}[\mathrm{VOC}]$) coordinated with regional NO_x abatement strategies. Because aromatic and olefinic species exhibit elevated reaction rate coefficients with hydroxyl radicals ($k_{\mathrm{OH}} \approx 10^{-11}$ to 10^{-10} $\mathrm{cm^3\ molecule^{-1}\ s^{-1}}$), mitigation policies must focus specifically on high-reactivity emission sources. In retail petroleum distribution, mandatory installation and

continuous vapor-tight operation of Stage I and Stage II vapor recovery systems are critical to curb fugitive benzene, toluene, and volatile alkane venting during bulk tanker offloading and vehicle refueling. Concurrently, environmental regulations should mandate industrial solvent reformulation across the Howrah foundry, coating, and metal fabrication clusters, incentivizing rapid adoption of low-VOC or water-based coatings, degreasers, and industrial sealants. Stringent surveillance and enforcement against the open combustion of municipal solid waste, plastics, and scrap tire burning remain equally vital, as unconstrained low-temperature incineration generates substantial, localized pulses of highly reactive aromatic and aliphatic intermediates (Venkataraman et al., 2018; Majumdar et al., 2008). Finally, accelerating the fleet transition of commercial transport, freight trucks, and para-transit three-wheelers to battery-electric or compressed natural gas propulsion provides dual photochemical benefits, simultaneously curbing primary nitric oxide titration bursts and reducing unburned reactive hydrocarbon exhaust.

4.4 Important Scientific Limitations of Precursor Screening Ratios

It is critical to articulate the scientific boundaries of empirical precursor indicator ratios. As emphasized in classical atmospheric chemistry literature (Sillman and He, 2002; Monks et al., 2015; Jin and Holloway, 2015), a simple ratio of measured BTEX to $\mathbf{NO}_x$ cannot definitively establish the precise photochemical transition line for an urban airshed. Incomplete VOC speciation presents a notable analytical constraint, as routine CAAQMS gas chromatography quantifies only five monoaromatic hydrocarbons (benzene, toluene, ethylbenzene, *m*, *p*-xylene, and *o*-xylene). Highly reactive light alkenes (e.g., ethene, propene, 1,3-butadiene), oxygenated carbonyls (formaldehyde, acetaldehyde), and biogenic terpenes and isoprene emitted from the nearby Sundarbans mangrove biosphere and urban tree canopies are not captured in continuous telemetry. Because formaldehyde and terminal alkenes exhibit hydroxyl radical reaction rate coefficients (k_{OH}) orders of magnitude higher than benzene, the true radical propagation rate of the ambient VOC pool can diverge substantially from BTEX mass fractions (Xue et al., 2014). Furthermore, non-local mesoscale transport and vertical boundary-layer mixing decouple surface measurements from column radical budgets: ground-level indicator ratios capture surface concentrations at fixed shelter heights (3–15 m), whereas secondary ozone formation occurs throughout the evolving convective boundary layer as air parcels traverse diverse emission landscapes. Consequently, definitive determination of ozone chemical regimes in this tropical conurbation ultimately necessitates observation-based box modeling (OBM) constrained by comprehensive VOC and radical speciation (such as the Master Chemical Mechanism) or 3D chemical transport modeling (e.g., WRF-Chem, CMAQ) driven by high-resolution local emission inventories (Jin and Holloway, 2015; Vieno et al., 2010). Within this context, the precursor ratios presented here serve as an empirical screening

framework that delineates localized proxy gradients while underscoring the necessity of intensive photochemical field campaigns.

4.5 Synthesis with Global Megacity Findings

The findings of this investigation align with and expand upon observations across major megacities worldwide. In the Beijing–Tianjin–Hebei and Yangtze River Delta regions of China, Xue et al. (2014) and Jin and Holloway (2015) observed widespread VOC-limited regimes across industrial and dense urban cores, transitioning to $\mathbf{NO}_x$-limited regimes in downwind rural sectors. During the COVID-19 lockdown in China, drastic vehicular $\mathbf{NO}_x$ reductions without corresponding VOC controls triggered historic ozone spikes in urban centers, demonstrating the real-world hazard of the ozone titration penalty (Le et al., 2020). Similarly, in the Mexico City Metropolitan Area, Lei et al. (2007) demonstrated that high aromatic hydrocarbon emissions from industrial solvents and domestic LPG leakage sustain a persistent VOC-limited regime across the urban core. Furthermore, in European and North American airsheds, long-term statistical evaluations have consistently confirmed that urban air pollutant concentrations adhere to lognormal distributions due to turbulent atmospheric mixing, while Weibull models systematically fail to capture stagnation accumulation tails (Georgopoulos and Seinfeld, 1986; Vieno et al., 2010).

4.6 Methodological Limitations and Future Research Priorities

While this investigation establishes a comprehensive multi-station empirical baseline, several limitations provide avenues for future scientific research. First, monitoring was restricted to surface-level sampling inlets (3–15 m above ground level). Continuous vertical profiling of planetary boundary layer height, temperature inversions, and vertical ozone profiles via ceilometers, micropulse lidars, and tethered balloon soundings is needed to quantify nocturnal residual layer ozone storage and morning fumigation dynamics. Second, direct measurements of short-lived radical intermediates ($\mathbf{OH}^{\bullet}$, $\mathbf{HO}_2^{\bullet}$, $\mathbf{RO}_2^{\bullet}$) and photolysis frequencies ($J(\mathbf{NO}_2)$, $J(\mathbf{O}^1\mathbf{D})$) were unavailable in routine monitoring telemetry, necessitating reliance on empirical chemical proxies. Third, future research should investigate secondary organic aerosol (SOA) formation pathways coupled with ozone photochemistry, exploring how high particulate surface area during winter haze episodes modulates heterogeneous radical uptake and actinic flux attenuation.

5 Conclusion

This investigation provides an empirical examination of ground-level ozone photochemistry, chemical regime indicators, parametric probability tail distributions, and out-of-

sample machine learning generalization across all 13 Continuous Ambient Air Quality Monitoring System stations in the Kolkata–Howrah conurbation during 2024 and 2025 (16.1 million continuous raw parameter readings). Ambient particulate concentrations persistently breached national benchmarks, with network-wide annual mean $PM_{2.5}$ reaching $49.5 \pm 41.1\,\mu g\ m^{-3}$, driven by severe winter stagnation inversions across industrial foundry clusters in Howrah (61.4–$66.1\,\mu g\ m^{-3}$) and dense traffic corridors in Kolkata. Diurnal trace gas cycles revealed pronounced spatial divergences in photochemical processing: suburban and background parkland stations exhibited unimodal daytime photochemical accumulation peaking at 41.2–$52.5\,\mu g\ m^{-3}$ between 13:00 and 14:00 IST alongside afternoon NO troughs (4.6–$12.1\,\mu g\ m^{-3}$). In contrast, heavy arterial intersections and industrial centers displayed pronounced nocturnal and rush-hour titration troughs (6.8–$24.9\,\mu g\ m^{-3}$), driven by intensive primary vehicular nitric oxide emissions that convert ozone into NO_2 while preserving total odd oxygen ($[O_x] = [O_3] + [NO_2]$).

Rigorous information-theoretic model selection across 182 spatiotemporal monitored pollutant series provides strong empirical support for Kolmogorov's theory of broken random variables and Ott's successive random dilution model in a tropical megacity boundary layer. The two-parameter Lognormal distribution achieved decisive empirical superiority, attaining minimum AIC in 81.3% of series (148/182) and minimum Kolmogorov–Smirnov statistics in 74.7% of series (136/182), reaching 100% optimality for volatile organic compounds and 73.1% for surface ozone. Conversely, the two-parameter Weibull distribution was optimal in only 3.3% of series (6/182) and incurred a conurbation-wide arithmetic mean ΔAIC penalty of $+5,763.6$, demonstrating that standard two-parameter Weibull formulations fail to capture the extended multiplicative dispersion tails and heavy stagnation accumulations characteristic of urban air quality. The Gamma distribution proved optimal in a secondary fraction of series (15.4%), concentrated within primary particulate matter ($PM_{2.5}$, 26.9%) and carbon monoxide (26.9%), where additive, discrete combustion pulses compete with atmospheric turbulent dilution.

Cross-year forward machine learning forecasting demonstrated a pronounced contrast in predictive transferability governed by micro-scale chemical environments. Random Forest models trained on 2024 meteorological features achieved robust out-of-sample generalization across unseen 2025 records at suburban ($R^2 = 0.6825$) and urban background ($R^2 = 0.5914$) sites, where regional actinic flux and thermal boundary-layer dynamics govern ozone production. In sharp contrast, model transferability collapsed to near-zero or negative skill at roadside corridors ($R^2 = 0.0755$ at RBU, -0.1722 at Ballygunge, and -0.2577 at Bidhannagar), while riverine transport from upwind industrial plumes similarly uncoupled ozone from local meteorology at Botanical Garden ($R^2 = 0.0499$). Diagnostic chemical counter-experiments at the representative Bidhannagar roadside monitor showed that incorporating co-located nitric oxide and nitrogen dioxide predictors reduced the negative cross-year forward transfer score (R^2 improving from -0.2577 to -0.0764),

while incorporating the broader reactive suite yielded slightly positive skill ($R^2 = +0.0195$). In separate within-airshed pooled 5-fold cross-validation, incorporating co-located chemical predictors increased accuracy from $R^2 = 0.6808$ to $R^2 = 0.7800$. These findings are consistent with unmodeled reactive chemistry contributing to roadside predictive breakdown, though localized titration operates alongside complex microscale street-canyon aerodynamics and fleet dynamics. Consequently, air quality management in tropical estuarine megacities must account for fine-scale chemical regime gradients, coordinating solvent recovery and fugitive vapor controls within VOC-sensitive traffic and industrial cores while pursuing regional NO_x abatement to avert localized ozone titration penalties.

Supplementary Information

The online version contains supplementary material available as Online Resource 1 (ESM_1.pdf), which includes CAAQMS station metadata, analytical instrument specifications, goodness-of-fit catalogs across criteria pollutants, and machine learning counter-experiment diagnostic matrices.

Acknowledgements

The author gratefully acknowledges the Central Pollution Control Board (CPCB) and the West Bengal Pollution Control Board (WBPCB) for providing open-access ambient air quality monitoring telemetry across the Kolkata and Howrah conurbation.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723, https://doi.org/10.1109/TAC.1974.1100705

Balakrishnan K, Dey S, Gupta T, Dhaliwal RS, Brauer M, Cohen AJ, Stanaway JD, Beig G, Joshi TK, Krishna B, Aggarwal A, Paul VK, Kumar R, Sankar K, Salvi S, Agarwal A, Shukla DK, Arora NK, Ghosh S, Dandona R, Dandona L (2019) The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. The Lancet Planetary Health 3(1):e26–e39, https://doi.org/10.1016/S2542-5196(18)30261-4

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer-Verlag, New York, https://doi.org/10.1007/b97636

Chakroborty B, Rudra K, Chakrabarty D, Naskar S, Das S (2025) Decoding spatiotemporal dynamics of air pollution and its underlying drivers in Kolkata metropolitan area through integrated field investigation and geospatial analysis. Discover Cities 2:77, https://doi.org/10.1007/s44327-025-00118-7

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, Feigin V, Freedman G, Hubbell B, Jobling A, Kan H, Knibbs L, Liu Y, Martin R, Morawska L, Pope CA, Shin H, Straif K, Degenhardt L, Vos T, Murray CJL, Forouzanfar MH (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 389(10082):1907–1918, https://doi.org/10.1016/S0140-6736(17)30505-6

De Marco A, Proietti C, Anav A, Ciancarella L, D'Elia I, Fares S, Fornasier MF, Fusaro L, Gualtieri M, Manes F, Marchetto A, Mircea M, Paoletti E, Piersanti A, Rogora M, Salvati L, Salvatori E, Screpanti A, Vialetto G, Vitale M, Leonardi C (2019) Impacts of air pollution on human and ecosystem health, and implications for the National Emission Ceilings Directive: Insights from Italy. Environment International 125:320–333, https://doi.org/10.1016/j.envint.2019.01.064

Georgopoulos PG, Seinfeld JH (1986) Mathematical modeling of turbulent reacting plumes—I. General theory and model formulation. Atmospheric Environment 20(9):1791–1807, https://doi.org/10.1016/0004-6981(86)90128-9

Ghosh B, Barman HC, Padhy PK (2023) Analysis of spatiotemporal distribution of air quality index (AQI) in the state of West Bengal, India from 2016 to 2021. Discover Atmosphere 1(1):1, https://doi.org/10.1007/s44292-023-00001-3

Ghude SD, Jena C, Chate DM, Beig G, Pfister GG, Kumar R, Ramanathan V (2014) Reductions in India's crop yield due to ozone. Geophysical Research Letters 41(15):5685–5691, https://doi.org/10.1002/2014GL060930

Jin X, Holloway T (2015) Spatial and temporal variability of ozone sensitivity over China observed from the Ozone Monitoring Instrument. Journal of Geophysical Research: Atmospheres 120(14):7229–7246, https://doi.org/10.1002/2015JD023250

Justus CG, Hargraves WR, Mikhail A, Graber D (1978) Methods for estimating wind speed frequency distributions. Journal of Applied Meteorology 17(3):350–353, https://doi.org/10.1175/1520-0450(1978)017<0350:MFEWSF>2.0.CO;2

Karmakar J (2020) Air pollution in Kolkata: Emerging challenges and dynamics. Journal of Atmospheric Science Research 3(4):1–9, https://doi.org/10.30564/jasr.v3i4.2231

Kolmogorov AN (1941) Über das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstückelung. Doklady Akademii Nauk SSSR 31(2):99–101

Le T, Wang Y, Liu J, Yang J, Linn RR, Seinfeld JH, Yung YL (2020) Unexpected air pollution with marked emission reductions during the COVID-19 lockdown in China. Science 369(6504):702–706, https://doi.org/10.1126/science.abb7431

Lei W, de Foy B, Zavala M, Volkamer R, Molina LT (2007) Characterizing ozone production in the Mexico City Metropolitan Area: a case study using a chemical transport model. Atmospheric Chemistry and Physics 7(5):1347–1366, https://doi.org/10.5194/acp-7-1347-2007

Leighton PA (1961) Photochemistry of Air Pollution. Academic Press, New York, https://doi.org/10.1016/B978-0-12-442250-6.X5001-6

Lu HC (2002) The statistical characters of PM_{10} concentration in Taiwan area. Atmospheric Environment 36(3):491–502, https://doi.org/10.1016/S1352-2310(01)00245-X

Majumdar D, Dutta C, Mukherjee AK, Sen S (2008) Source apportionment of VOCs at the petrol pumps in Kolkata, India; exposure of workers and assessment of associated health risk. Transportation Research Part D: Transport and Environment 13(8):524–530, https://doi.org/10.1016/j.trd.2008.09.011

Monks PS, Archibald AT, Colette A, Cooper O, Coyle M, Derwent R, Fowler D, Granier C, Law KS, Mills GE, Stevenson DS, Tarasova O, Thouret V, von Schneidemesser E, Sommariva R, Wild O, Young PJ (2015) Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. Atmospheric Chemistry and Physics 15(15):8889–8973, https://doi.org/10.5194/acp-15-8889-2015

Ott WR (1990) A physical explanation of the lognormality of pollutant concentrations. Journal of the Air & Waste Management Association 40(10):1378–1383, https://doi.org/10.1080/10473289.1990.10466789

Ran L, Zhao C, Geng F, Tie X, Tang X, Peng L, Zhou G, Yu Q, Xu J, Guenther A (2009) Ozone photochemical production in urban Shanghai, China: Analysis based on ground level observations. Journal of Geophysical Research: Atmospheres 114:D15301, https://doi.org/10.1029/2008JD010752

Sarkar M, Das A, Mukhopadhyay S (2021) Assessing the immediate impact of COVID-19 lockdown on the air quality of Kolkata and Howrah, West Bengal, India. Environment, Development and Sustainability 23(6):8613–8642, https://doi.org/10.1007/s10668-020-00985-7

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464, https://doi.org/10.1214/aos/1176344136

Seinfeld JH, Pandis SN (2016) Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 3rd edn. John Wiley & Sons, Hoboken, NJ

Sillman S, He D (2002) Some theoretical results concerning O_3-NO_x-VOC chemistry and NO_x-VOC indicators. Journal of Geophysical Research: Atmospheres 107(D22):4659, https://doi.org/10.1029/2001JD001123

Tie X, Geng F, Guenther A, Cao J, Greenberg J, Zhang R, Apel E, Li G, Weinheimer A, Chen J, Cai C (2013) Megacity impacts on regional ozone formation: observations and WRF-Chem modeling for the MIRAGE-Shanghai field campaign. Atmospheric Chemistry and Physics 13(11):5655–5669, https://doi.org/10.5194/acp-13-5655-2013

Venkataraman C, Brauer M, Tibrewal K, Sadavarte P, Ma Q, Cohen A, Chaliyakunnel S, Frostad J, Klimont Z, Martin RV, Millet DB, Philip S, Walker K, Wang S (2018) Source influence on emission pathways and ambient $PM_{2.5}$ pollution over India (2015–2050). Atmospheric Chemistry and Physics 18(11):8017–8039, https://doi.org/10.5194/acp-18-8017-2018

Vieno M, Dore AJ, Stevenson DS, Doherty R, Heal MR, Reis S, Hallsworth S, Tarrason L, Wind P, Fowler D, Simpson D, Sutton MA (2010) Modelling surface ozone during the 2003 heat-wave in the UK. Atmospheric Chemistry and Physics 10(16):7963–7978, https://doi.org/10.5194/acp-10-7963-2010

World Health Organization (2021) WHO Global Air Quality Guidelines: Particulate Matter ($\mathbf{PM_{2.5}}$ and $\mathbf{PM_{10}}$), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva, URL https://iris.who.int/handle/10665/345329

Xue L, Wang T, Gao J, Ding A, Zhou X, Blake DR, Wang X, Saunders SM, Fan S, Zuo H, Zhang Q, Wang W (2014) Ground-level ozone in four Chinese cities: precursors, regional transport and heterogeneous processes. Atmospheric Chemistry and Physics 14(23):13175–13188, https://doi.org/10.5194/acp-14-13175-2014

Statements and Declarations

Funding: The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The author has no relevant financial or non-financial interests to disclose.

Author Contributions: Saptarshi Sadhu conceived the study, curated the telemetry datasets, developed the mathematical and machine learning pipelines, performed the statistical validations, and wrote the manuscript.

Data Availability: Continuous ambient air quality telemetry is publicly accessible through the Central Pollution Control Board (CPCB) CCR portal (https://airquality.cpcb.gov.in). All reproducible Python analysis scripts, statistical validation pipelines, and processed matrices are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable. This study utilizes non-human ambient atmospheric telemetry and does not involve human participants or animal subjects.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable. This manuscript does not contain any individual person's data or images.

Computational Framework and AI Disclosure: All data processing, statistical modeling, and validation pipelines were implemented in Python 3.13. Generative AI tools (Anthropic Claude and Google Gemini) assisted in computational script development, mathematical formalization, structuration of theoretical linkages, and manuscript drafting and language editing. The author conceived the research objectives, curated and verified all primary datasets, independently audited all mathematical formulations, numerical outputs, and statistical interpretations, and takes full responsibility for the integrity and accuracy of the manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PreprintPaper2ESM1Supplementary.pdf