© Volume 2, Issue 9, ***September*** 2026 | JATIR
ISSN: 3139-8464 (p)

Simulation Modelling of Urban Air Pollution: A Gaussian Dispersion–Machine Learning Hybrid Framework for Roadside Pm2.5 Prediction

[1]Prakash L Pathak, [2]Dr. Pradip D Jadhao
[1]Research Scholar, Department of Civil Engineering, K.K. Wagh Institute of Engineering Education and Research, Nashik, India
[2]Professor and Head, Department of Civil Engineering, K.K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract—**Urban roadside PM2.5 pollution has become one of the more pressing problems facing fast-growing cities, driven by rising traffic and boundary layers that often do not ventilate near-road emissions effectively. In this paper we set out to build and test a hybrid modelling framework that couples a physically based Gaussian line-source dispersion model with a data-driven Random Forest model, in a model-output-statistics (bias-correction) configuration. We want to state clearly, up front, what kind of study this is: because our aim here was methodological — to work through the full modelling, validation and comparison pipeline end to end, and to leave a workflow that is directly reusable once real data are available — we built a synthetic, thirty-day hourly case study for a representative urban arterial corridor. Traffic, meteorological and atmospheric-stability inputs were drawn from patterns and emission factors reported in the published dispersion literature, not from a specific city's monitoring network. The "observed" concentrations used to validate the models are pseudo-observations, generated by adding representativeness-type noise to the physical model's own output, rather than independent field measurements.**

We compared three configurations - the Gaussian dispersion model on its own, a standalone Random Forest model, and the hybrid combination — on an identical held-out test partition, using six standard air-quality validation statistics: RMSE, MAE, R^2, mean bias error (MBE), fractional bias (FB), and Willmott's index of agreement (IOA). The Gaussian model came out ahead on every overall-accuracy statistic (RMSE = 5.10 µg/m³, R^2 = 0.80, IOA = 0.94); this is expected, since the pseudo-observations were generated from that same model plus noise, so it really sets an upper bound for this exercise rather than an independent benchmark. The standalone Random Forest model trailed on every measure (RMSE = 8.01 µg/m³, R^2 = 0.50, IOA = 0.79). The hybrid model recovered almost all of the physical model's accuracy while roughly halving the fractional bias (FB = 0.008 versus

0.015), which suggests that residual correction earns its keep mainly by removing systematic bias rather than by smoothing random noise. A feature-importance analysis pointed to traffic volume, wind direction and mixing height as the strongest predictors of roadside PM2.5 in this setup. Taken together, the results support a practical way of pairing the two modelling philosophies: keep the physical dispersion model as the primary predictive engine, and use machine learning mainly for bias correction, gap-filling and short-term forecasting rather than as a full replacement. We close with some thoughts on what it would take to extend this framework — real monitoring data, multi-pollutant chemistry, street-canyon-resolving CFD — into something usable for actual urban air-quality decisions.

Index Terms—**Air pollution modelling; Gaussian dispersion; Urban PM2.5; Machine learning; Random Forest; model validation; Hybrid simulation; Traffic emissions**

I. INTRODUCTION

Air pollution in urban areas results from a complex interaction between emission sources, atmospheric transport and dispersion processes, meteorology, and topography. Among the criteria pollutants, fine particulate matter (PM2.5) is of particular concern due to its ability to penetrate deeply into the respiratory tract and its established associations with cardiovascular and respiratory morbidity and mortality. In most rapidly urbanizing regions, road traffic is the dominant near-source contributor to ambient PM2.5 in city centers and along arterial corridors, compounded by industrial emissions, construction dust, biomass burning and secondary aerosol formation at the regional scale.

Air pollution simulation modelling provides civil and environmental engineers with a quantitative tool to translate emission inventories and meteorological conditions into predicted ambient concentrations, without the cost and lag associated with dense monitoring networks. Such models support source apportionment, health-impact assessment, urban and transport planning, early-warning systems, and evaluation of control strategies (e.g., low-emission zones, fleet electrification, traffic management) before they are implemented. Broadly, two modelling philosophies have matured over the past four decades: (i) physically based deterministic models, rooted in the advection–diffusion equation and typified by Gaussian plume/puff models, Eulerian grid models and computational fluid dynamics (CFD); and (ii) data-driven statistical and machine-learning (ML) models that learn empirical relationships between meteorological/traffic covariates and observed pollutant concentrations without explicitly solving the transport equations.

Each philosophy has complementary strengths. Physical models are interpretable, extrapolate more safely to new emission scenarios, and are consistent with the underlying physics of turbulent dispersion; however, they can be sensitive to parameterization errors, and their accuracy in complex urban geometries (street canyons, building wakes) is limited unless computationally expensive CFD is used. ML models can achieve strong short-term predictive accuracy where dense historical data exist, and they naturally absorb unmodelled local effects, but they are data-hungry, can behave unpredictably outside their training envelope, and offer limited mechanistic insight. This has motivated growing interest in hybrid frameworks that use ML to correct or downscale the residual bias of a physical model rather than replace it outright — an approach sometimes termed model-output-statistics (MOS) in the atmospheric-science literature.

This paper develops and demonstrates such a hybrid simulation framework for urban roadside PM2.5. Because the purpose of the paper is methodological - to demonstrate the modelling, validation and comparison workflow in full — a representative synthetic case study is used, built from diurnal traffic and meteorological patterns and emission factors consistent with published urban corridor studies, rather than a specific city's proprietary monitoring dataset. The same workflow, code structure and validation protocol are directly transferable to any urban corridor once real traffic-count, meteorological and monitoring data are substituted for the illustrative inputs used here.

1.1 Research Problem Statement

i. Monitoring networks are sparse. Most cities operate only a handful of continuous ambient air-quality stations — nowhere near enough to resolve the sharp spatial gradients of traffic-related PM2.5 near busy corridors, intersections and street canyons.

ii. There is a persistent trade-off in model choice. Purely physical dispersion models are transparent and physically consistent, but carry systematic bias when local emission factors, stability classification or urban morphology are imperfectly known. Purely statistical/ML models can fit historical data well, yet generalise poorly the moment traffic, meteorological or emission-control conditions move outside what the model has seen before.

iii. There is also a validation gap in the literature. Many urban air-quality studies report a single model's fit statistics in isolation, without a like-for-like comparison against an alternative modelling philosophy — which makes it hard for a practitioner to pick the right tool for a given planning question. Regulatory screening and short-term public-health forecasting are quite different jobs, and they don't necessarily call for the same model.

1.2. Research Objectives

i. To develop a hybrid air pollution prediction framework.

ii To simulate and predict hourly roadside $PM_{2.5}$ concentrations.

iii To evaluate and compare the performance.

iv To identify the key factors influencing roadside $PM_{2.5}$ concentrations and assess the applicability of the hybrid modelling framework.

in the methodology and results that follow.

ISSN: 3139-8464 (p)

II. METHODOLOGY

2.1 Overall Framework

The modelling framework comprises four stages, illustrated conceptually in Figure 1's underlying workflow: (i) construction of an hourly emission and meteorological input dataset for a representative urban arterial corridor; (ii) simulation of the corresponding roadside PM2.5 concentration field using a Gaussian line-source dispersion model, treated as the physically based 'first-guess' prediction; (iii) generation of a pseudo-observed concentration series by adding representativeness/instrument-type noise to the physical simulation, used here in place of a monitored dataset to demonstrate the full validation workflow; and (iv) training and validation of a Random Forest regression model, and of a hybrid Gaussian-plus-RF-residual-correction model, against a held-out test partition of the pseudo-observed series.

2.2 Gaussian Line-Source Dispersion Model

Vehicular emissions along an urban arterial are represented as a continuous, finite line source of length L, oriented at bearing θ_road, with time-varying emission strength q(t) (g m^{-1} s^{-1}) obtained from hourly traffic volume N(t) (veh/h) and a composite fleet-average PM2.5 emission factor EF (g veh^{-1} km^{-1}):

$$q(t) = N(t) \times EF / 3600 / 1000 \quad (1)$$

The steady-state ground-level concentration increment at a receptor located at perpendicular distance x from an (effectively infinite relative to σy at short range) line source, following the standard Gaussian line-source formulation used in CALINE-type models, is:

$$C(x, t) = [2\, q(t) / (\sqrt{2\pi}\, \sigma z(t)\, u(t) \sin \varphi(t))] \times \exp[-H^2 / (2\, \sigma z(t)^2)] \qquad (2)$$

where u(t) is wind speed (m/s), φ(t) is the angle between wind direction and the road axis, σz(t) is the vertical dispersion coefficient (m) determined from the Pasquill–Gifford atmospheric stability class prevailing at hour t, and H is the effective release height of traffic-generated PM2.5 (taken as 0.5 m). Stability class was assigned using a simplified day/night rule: convective/near-neutral daytime conditions (07:00–17:00) were assigned a larger σz (more effective vertical mixing), evening/early-morning transition hours an intermediate σz, and night-time hours (23:00–04:00) the smallest σz, representing shallow, stable boundary layers that trap traffic emissions near the surface.

A nocturnal dilution correction, $f\_dil = \min(3.0, \max(0.7, 600/H\_mix(t)))$, was applied to the line-source term to represent enhanced near-surface accumulation when the mixing height H_mix(t) collapses at night, and an urban background term B(t) — representing regional and secondary aerosol contributions with its own diurnal cycle — was added to obtain the total simulated concentration:

$$C\_total(t) = C(x, t) \times f\_dil(t) + B(t) \qquad (3)$$

ISSN: 3139-8464 (p)

Table 2. Model inputs and their representation in the thirty-day case-study simulation.

Model input / parameter	Representation in the simulation
Road geometry	Single line source, length 1000 m, receptor set back 30 m from kerb, release height 0.5 m
Traffic volume	Diurnal bimodal pattern (AM peak ≈ 09:00, PM peak ≈ 18:30) with ±6% stochastic variability, 720 hourly values (30 days)
Emission factor	Composite fleet-average PM2.5 factor of 0.045 g/veh-km (representative of a mixed petrol/diesel/CNG urban fleet)
Wind speed / direction	Diurnal sinusoidal pattern with random perturbation; wind–road incidence angle computed each hour
Atmospheric stability	Simplified Pasquill–Gifford day/night classification controlling vertical dispersion coefficient (σz)
Mixing height	Diurnal boundary-layer growth/collapse cycle (120–1150 m) used to scale nocturnal trapping of traffic plume
Urban background	Regional/secondary aerosol contribution with diurnal and slow (30-day) trend components
Temperature / relative humidity	Diurnal cycle with stochastic noise; used only as ML predictors, not in the physical dispersion term

2.3 Data-Driven (Random Forest) Model

A Random Forest regression ensemble (300 trees, maximum depth 10, minimum leaf size 3) was trained to predict hourly PM2.5 directly from seven covariates — traffic volume, wind speed, wind direction, temperature, relative humidity, mixing height and hour-of-day — without reference to the Gaussian dispersion equations. The dataset (n = 720 hourly records) was split 75:25 into training and test partitions using a fixed random seed to ensure reproducibility, and performance was evaluated only on the held-out 25% test partition, consistent with standard machine-learning validation practice.

2.4 Hybrid (Bias-Correction) Model

A second Random Forest (300 trees, maximum depth 6, minimum leaf size 5 — a shallower ensemble appropriate for modelling a lower-variance residual signal) was trained to predict the residual, $r(t) = C\_obs(t) − Gaussian(t)$, from the same seven covariates. The hybrid prediction was formed by adding the predicted residual back onto the physical model's output, $Hybrid(t) = Gaussian(t) + \hat{r}(t)$, evaluated on the identical held-out test partition used for the standalone models, ensuring a fair, like-for-like three-way comparison.

ISSN: 3139-8464 (p)

2.5 Validation Statistics

Model performance was quantified using six standard air-quality model evaluation statistics computed on the test partition: root-mean-square error (RMSE) and mean absolute error (MAE), which measure overall prediction error in the units of the pollutant; the coefficient of determination (R^2); mean bias error (MBE) and fractional bias (FB = 2·mean(pred-obs)/mean(pred+obs)), which measure systematic over- or under-prediction; and Willmott's index of agreement (IOA), a bounded (0–1) measure of overall model skill that is less sensitive to outliers than R^2 alone. These are the same statistics recommended in USEPA and WMO model-evaluation guidance for regulatory and research-grade air-quality models.