Drift-Aware Gaussian Process Nowcasting of Urban $PM_{2.5}$ from Sparse Reference Monitors and Low-Cost Sensors

Recursive affine calibration without co-location, evaluated in a CPCB-parameterised simulation

[1]Chandrasekhar Panda, [2]Pravat Satpathy, [3]Sonia Mohapatra
[1]Researcher, [2]Researcher, [3]Researcher
[1,2,3]Research and Development Division,
Tekstein Solutions Private Limited, Bhubaneswar, Odisha, India
[1]contact@chandrasekharpanda.com, [2]analytics.pravat@gmail.com, [3]smctc98@gmail.com

Abstract*—Continuous reference-grade monitoring in Indian cities is sparse. A Tier-2 city such as Bhubaneswar relies on a handful of CPCB-grade analysers and on manual stations that sample twice a week, so an hourly $PM_{2.5}$ map is in practice an extrapolation from three or four points. Low-cost optical sensors can be deployed at ten times that density, but their readings carry multiplicative gain error, offset, humidity-driven over-reading and slow drift, and the correction fitted at deployment is rarely revisited. This paper treats drift correction and spatial nowcasting as a single estimation problem. Each sensor is modelled as an affine, humidity-dependent transform of the true concentration whose gain and offset follow a random walk; a windowed spatio-temporal Gaussian process (GP) anchored to the reference monitors estimates the field, and a block leave-one-sensor-out posterior at each sensor site supplies the regressor for a per-sensor Kalman update of the affine state. No co-location period is required, and one hourly update for 28 sites takes about 31 ms on a single core. In a 40-day winter simulation parameterised on CPCB and OSPCB statistics (city mean 61 µg m^{-3}, four references, 24 sensors reading 80 % high before correction), the method lowers grid-wide RMSE from 12.7 to 11.3 µg m^{-3} against a reference-only GP, lowers error at held-out reference sites by 30 %, and brings the error of the corrected sensor readings from 54.9 (raw) to 8.2 µg m^{-3} against a noise floor of 4.0, while a calibration frozen after one week has degraded to 12.7 by the end of the record. A 25 % fouling step is absorbed within two to four days. The shortfalls are reported with equal weight: uncalibrated sensors are worse than no sensors, dropping the humidity term removes the benefit of the sensors entirely, the 90 % predictive intervals cover only 58 % of truth, and a perfectly calibrated oracle shows that most of the remaining error is a representativeness floor set by sensor density rather than by calibration.

***Index Terms*—Air quality nowcasting, calibration transfer, Gaussian process regression, Kalman filter, low-cost sensors, PM2.5, sensor drift, CPCB.**

I. Introduction

India's regulatory network measures the right thing at too few places. Bhubaneswar, a notified non-attainment city under the National Clean Air Programme, is covered by six manual stations of the National Air Quality Monitoring Programme that sample PM_{10} and $PM_{2.5}$ on two days a week, giving 104 observations a year per site, and by a small number of continuous ambient air quality monitoring stations (CAAQMS) reporting to the Central Pollution Control Board (CPCB) [1], [2]. Modelled urban $PM_{2.5}$ for the city was 47.7 ± 9.4 µg m^{-3}, above the 40 µg m^{-3} annual standard, and the state board's own bulletins show $PM_{2.5}$ becoming the prominent pollutant in winter [2], [3]. At that density the gradient between a traffic corridor and a residential block two kilometres away, the scale at which exposure and enforcement operate, is simply unobserved.

Low-cost optical particle sensors change the arithmetic [13]. A unit that costs two orders of magnitude less than a beta-attenuation monitor can be placed on every third lamp post. The difficulty is well documented: a nephelometer-type sensor responds to light scattering, not mass, so its output is a gain and offset away from the true concentration, the gain rises steeply with relative humidity (RH) as hygroscopic particles take up water, and both gain and offset wander as the optics foul and the laser ages [4]–[7]. Field corrections such as the US-wide PurpleAir equation [6] or random-forest calibrations learned during co-location [4] are effective on the day they are fitted and degrade afterwards; Malings et al. report that corrections for the same sensor family had to be re-derived across seasons [5].

Spatial inference and sensor calibration are usually handled separately: kriging and Gaussian-process (GP) interpolation take their inputs as given [8], [9], and calibration studies treat each sensor in isolation against a co-located reference. The two problems are coupled. The best estimate of the concentration at a sensor site, which is what calibration needs, is the spatial posterior from everything else in the network, and that posterior needs sensors whose bias

has been removed. Datta et al. exploited this coupling for a Baltimore network with a static regression inside a spatial model [10]; drift, humidity and recursive operation were outside their scope.

This paper closes that loop. The contributions are: (i) a formulation of drift-aware calibration as state estimation, in which each sensor's affine parameters are a Kalman state driven by a random-walk prior and observed through the GP's block leave-one-sensor-out posterior at the sensor's own location; (ii) an identifiability result showing why at least one reference monitor is necessary and sufficient to anchor the scale of the whole network; (iii) a windowed spatio-temporal GP nowcaster with a humidity-aware observation model that runs at interactive speed; and (iv) a simulation study, parameterised on CPCB and OSPCB statistics, that reports where the approach helps and where it does not, including an oracle bound that separates the calibration error from the representativeness error of the network geometry.

II. Related Work

Table I summarises the strands this work draws on. Every calibration study that followed sensors for more than a season found that a single correction does not hold [5], [7], [11]; the dominant error is multiplicative and RH-dependent rather than additive [6], [12]; and the spatial-inference literature has scalable GP machinery [9] and urban inference systems [14], [15] that treat sensor error as white noise. The method below keeps the GP as the spatial engine and places the calibration state inside its observation model.

TABLE I. Related work and the gap addressed

Work	Approach	Limitation for drifting networks
Zimmerman et al. [4]	Random-forest calibration from co-location	Requires co-location; no spatial model; no drift tracking
Malings et al. [5]	Long-term PM corrections for RAMP/PurpleAir	Seasonal re-fitting needed; per-sensor, not network-wide
Barkjohn et al. [6]	Nationwide affine + RH correction	Static coefficients; single sensor family
Crilley et al. [12]	κ-Köhler hygroscopic correction of OPC-N2	Humidity only; gain and offset left uncorrected
Datta et al. [10]	Statistical field calibration within spatial model	Static regression; no recursion or drift
Zheng et al. [14]; Cheng et al. [15]	Urban inference from heterogeneous data	Sensor bias treated as white noise
This work	GP nowcast + per-sensor Kalman affine state	Simulation only; variance under-dispersed (Section VI-E)

III. Problem Formulation

Let $c(x, t)$ denote the true $PM_{2.5}$ concentration (µg m^{-3}) at location x in a city domain and hour t. A reference monitor j at x_j reports $y_j^{r}(t) = c(x_j, t) + e_j$, with $e_j \sim N(0, \sigma_r^2)$. A low-cost sensor i at x_i reports

$$y_i(t) = a_i(t) + b_i(t)\, h(\mathrm{RH}_t)\, c(x_i, t) + \varepsilon_i(t), \qquad (1)$$

where a_i is the offset, b_i the gain, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ the measurement noise and $h(\cdot)$ a hygroscopic growth factor. Following the κ-Köhler form used for optical particle counters [12], the factor is taken as

$$h(\mathrm{RH}) = 1 + \kappa\, \mathrm{RH} / (1 - \mathrm{RH}), \qquad (2)$$

with κ an effective hygroscopicity for the sensor family. Equation (2) is a deliberate simplification of a mass-growth law: it multiplies the gain by 1.1 at 50 % RH and by 1.6 at 85 % RH for $\kappa = 0.10$, which is the range of over-reading reported for Plantower-class sensors in humid conditions [6], [12]. The gain and offset evolve slowly,

$$\theta_i(t+1) = \theta_i(t) + w_i(t), \quad \theta_i = [a_i, b_i]^{\mathrm{T}}, \quad w_i \sim N(0, Q), \qquad (3)$$

a random walk whose covariance Q sets how fast the estimator is allowed to forget. Nothing is co-located: the a_i, b_i and $c(\cdot, \cdot)$ are all unknown, and the task at each hour is to produce the posterior mean and variance of c on a grid from the reference and sensor streams up to that hour.

Proposition 1 (scale anchoring). *Under the model of (1) with no reference monitors, the likelihood of the sensor data is invariant to the transformation $c' = \beta c + \alpha$, $b'_i = b_i/\beta$, $a'_i = a_i - b_i\alpha/\beta$ for any $\beta > 0$ and α, so the field is identified only up to an affine map. One reference monitor with non-degenerate spatial covariance to the sensors removes the invariance.*

Proof. Substituting the primed quantities into (1) gives $a'_i + b'_i\, h\, c' = a_i - b_i\alpha/\beta + (b_i/\beta)\, h\, (\beta c + \alpha) = a_i + b_i\, h\, c + b_i\alpha(h - 1)/\beta$. The last term vanishes when $h \equiv 1$ and is otherwise a known deterministic function of RH that can be absorbed into $a'_i(t)$ within the random-walk prior, so the sensor likelihood is unchanged. A reference observation $y^{r} = c + e$ is not invariant: the transformed field predicts $\beta c + \alpha$ at x_j, which conflicts with y^{r} unless $\beta = 1$ and $\alpha = 0$. Because the GP prior couples $c(x_j, t)$ to $c(x_i, t)$ with non-zero covariance, the constraint propagates to every sensor. ∎

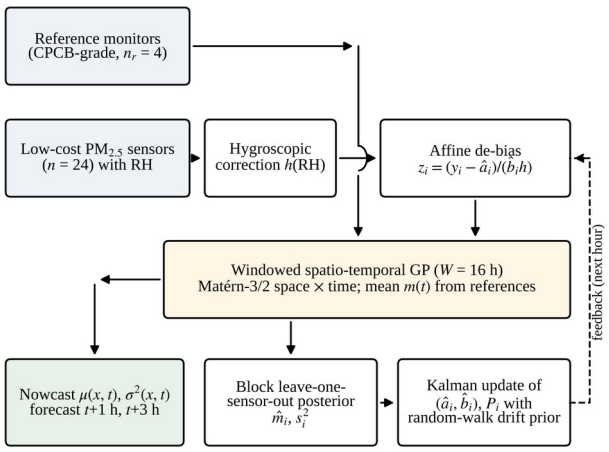


Fig. 1. DAC-GP architecture. Reference monitors and humidity-corrected, de-biased low-cost readings feed a windowed spatio-temporal GP. The GP's block leave-one-sensor-out posterior at each sensor site is the regressor for a Kalman update of that sensor's gain and offset, which closes the loop at the next hour.

A sensor network with no reference anchor can be made internally consistent but not correct; one CAAQMS inside the GP correlation length fixes the scale of all of them. The simulation uses four, closer to what a Tier-2 city has.

IV. Proposed Method: DAC-GP

Fig. 1 shows the drift-aware calibrated GP (DAC-GP). Each hourly cycle performs three operations: de-bias the sensor stream with the current affine state, solve a windowed GP for the field, and update each sensor's affine state using the GP's leave-one-sensor-out view of that sensor's location.

A. Windowed spatio-temporal GP

The field is decomposed as $c(x, t) = m(t) + f(x, t)$, where $m(t)$ is the mean of the reference readings at hour t and f carries the spatial anomaly. The anomaly is given a zero-mean GP prior with a separable product kernel,

$$k((x, t), (x', t')) = \sigma_f^2 \, k_{3/2}(\|x - x'\|/\ell_s) \, k_{3/2}(|t - t'|/\ell_t), \quad (4)$$

$$k_{3/2}(r) = (1 + \sqrt{3}\, r) \exp(-\sqrt{3}\, r), \quad (5)$$

with spatial and temporal length scales ℓ_s (km) and ℓ_t (h). The Matérn-3/2 form was preferred to the squared exponential because plumes from corridors and kilns have sharper edges than a Gaussian kernel admits. Only the most recent $W = 16$ h of data enter each solve, which caps the training set at $N = W(n_r + n)$ points (448 for four references and 24 sensors) and makes the exact GP affordable. Each reference observation enters with noise variance σ_r^2. Each sensor observation is first inverted through (1) with the current state estimate,

$$z_i = (y_i - \hat{a}_i) / (\hat{b}_i \, h), \quad (6)$$

$$v_i = (\sigma_\varepsilon^2 + P_{aa}) / (\hat{b}_i \, h)^2 + z_i^2 \, P_{bb} / \hat{b}_i^2, \quad (7)$$

where P is the state covariance of sensor i and (7) is the first-order (delta-method) variance of z_i. A sensor whose gain is still uncertain is therefore down-weighted by the GP automatically. With K the $N \times N$ kernel matrix plus the diagonal of noise variances, and r the vector of residuals after subtracting m, the nowcast at grid cell x and horizon $\tau \geq 0$ is

$$\mu(x, t{+}\tau) = m(t) + k^{\mathrm{T}} K^{-1} r, \quad \sigma^2(x, t{+}\tau) = \sigma_f^2 - k^{\mathrm{T}} K^{-1} k, \quad (8)$$

with k the kernel vector between the target and the training points. For $\tau > 0$ the temporal kernel pulls the anomaly back toward $m(t)$ while the background mean is held, so the method has short-range skill but is not a forecast model.

B. Recursive affine calibration

The regressor for calibrating sensor i must not contain sensor i's own data, otherwise the update would confirm whatever bias the sensor already has. Let I index the points in the current window that belong to sensor i and let $A = K^{-1}$. The block leave-one-out identity gives the posterior of the field at those points from all other data as

$$\mu_{-*|*} = r_I - (A_{II})^{-1} (A\, r)_I, \quad \Sigma_{-*|*} = (A_{II})^{-1} - \mathrm{diag}(v_I), \quad (9)$$

at the cost of one $|I| \times |I|$ inversion per sensor, since A is already available from the GP solve. The current-hour element gives $\hat{m}_i = m(t) + [\mu_{-*|*}]_{\text{last}}$ and its variance s_i^2. Equation (1) is then linear in θ_i given $\hat{m}_i$:

$$y_i(t) = H_i \, \theta_i + \eta, \quad H_i = [1, \, h\, \hat{m}_i], \quad \eta \sim \mathrm{N}(0, R_i), \quad (10)$$

with $R_i = \sigma_\varepsilon^2 + \hat{b}_i^2 \, h^2 \, s_i^2$, and the standard Kalman predict–update pair [16] with process covariance Q from (3) follows. The regressor uncertainty s_i^2 enters the measurement noise, so a sensor sitting far from any other site, where the GP knows little, receives a small Kalman gain and its calibration moves slowly; a sensor near a reference is calibrated quickly.

C. Algorithm and complexity

Algorithm 1 lists one hourly cycle. The dominant cost is the Cholesky factorisation and inversion of K, $O(N^3)$ with $N \approx 450$, followed by the grid prediction, $O(N^2 G)$ for G cells, and n block inversions of size at most W. Memory is $O(N^2 + NG)$. With $G = 576$ and $n = 24$ the measured cost is 31 ms per hour on one core of a general-purpose CPU, and the sweep in Section VI-D shows 54 ms at $n = 40$. The three hyper-parameters (σ_f^2, ℓ_s, ℓ_t) are estimated once by maximising the marginal likelihood on the window at the end of a seven-day burn-in, after the sensors have been calibrated, and are then frozen.

ALGORITHM 1. One DAC-GP cycle at hour t

Step	Operation
1	Predict: $P_i \leftarrow P_i + Q$ for every sensor i reporting at t
2	Form window: references with variance σ_r^2; sensors as (z_i, v_i) from (6)–(7); residuals r = value − $m(\cdot)$
3	Factorise K; compute $A = K^{-1}$, $\alpha = A\, r$
4	Nowcast and forecast on the grid for $\tau \in \{0, 1, 3\}$ h using (8)
5	For each reporting sensor i: block LOO (9) → $\hat{m}_i$, s_i^2; Kalman update of θ_i, P_i with (10); clip $\hat{b}_i \geq 0.2$
6	Emit μ, σ^2, θ; advance to $t + 1$

V. Simulation Design

There is no public hourly ground truth for an Indian city at 0.5 km resolution against which a nowcast could be scored, so the study is a controlled simulation whose statistics are tied to what the CPCB and the Odisha State Pollution Control Board report. All parameters are listed in Table II; the code and seed are in the reproducibility bundle.

Truth field. A 12 × 12 km domain on a 0.5 km grid is driven for 40 winter days by (i) a city background with a synoptic day-to-day AR(1) component, a winter diurnal cycle peaking at night under the inversion, and a Sunday reduction; (ii) local sources: two traffic corridors with morning and evening peaks, a brick-kiln cluster in the north-west, a construction site active in daytime for 18 days, and a night-time open-burning zone, each an anisotropic Gaussian plume displaced and elongated by an hourly wind; and (iii) a stochastic residual built from 40 random Fourier modes with AR(1) coefficients, equivalent to a stationary GP with $\ell_s = 2.5$ km and $\ell_t = 3$ h. The resulting city mean is 61 µg m^{-3}, the hourly city mean ranges 20–96 µg m^{-3}, 60 % of days exceed the 60 µg m^{-3} 24-h standard, and the mean within-hour spatial standard deviation is 12 µg m^{-3}. These sit where the 2015–2023 OSPCB tabulations for Bhubaneswar place a winter month [2], [3].

Network. Four reference monitors with $\sigma_r = 2$ µg m^{-3} are placed at generic urban positions, none on a hotspot. Twenty-four sensors are scattered uniformly. Each has offset $a_i \sim \mathrm{N}(3, 2^2)$, initial gain $b_i \sim \mathrm{N}(1.40, 0.25^2)$, a per-sensor κ drawn from U(0.08, 0.12) while the estimator assumes 0.10, a linear gain trend of ±18 % over the record plus a random walk, noise $\sigma = 3 + 0.05c$, 5 % random dropouts and a 48-h blackout for three sensors. Sensors 1 and 2 suffer a 25 % gain step at day 30 to represent fouling. Before correction the sensors read 80 % high on average, the mean absolute gain change over the record is 17 %, and RH averages 71 % (range 42–95 %).

Comparators. REF-MEAN assigns the reference mean everywhere; IDW-raw is inverse-distance weighting of all raw readings; REF-GP is the same GP using references only; RAW-GP feeds uncalibrated sensors with σ_ε noise; STATIC-GP calibrates with DAC-GP during the burn-in and then freezes θ, which is the co-location-once practice; DAC-GP (no RH) sets $h \equiv 1$; ORACLE-GP replaces each sensor with its true concentration plus 3 µg m^{-3} noise, an upper bound on what any calibration could achieve at these sites. Metrics on every grid cell and hour after burn-in are RMSE, MAE, R^2, coverage of the nominal 90 % interval, RMSE on the top decile of cells per hour (hotspots), and the F_1 score for hourly exceedance of 60 µg m^{-3}. A leave-one-reference-out (LORO) test, the only one of these that could be run on real data, withholds each reference in turn and scores the nowcast at its site against its readings.

TABLE II. Simulation parameters

Parameter	Value
Domain, grid, record	12 × 12 km, 0.5 km (576 cells), 40 d hourly
Truth residual GP	ℓ_s = 2.5 km, ℓ_t = 3 h, s.d. 7 µg m^{-3}
References / sensors	4 (σ_r = 2) / 24
Sensor gain, offset, κ	N(1.40, 0.25^2), N(3, 2^2), U(0.08, 0.12)
Gain drift	±18 % trend + RW (s.d. 0.0035 h$^{-1/2}$); 25 % step, 2 sensors
Estimator priors	a ~ N(0, 10^2), b ~ N(1, 0.5^2), Q = diag(0.02^2, 0.0035^2)
Estimator noise, κ	σ_ε = 4 µg m^{-3}, κ = 0.10
Window, burn-in	W = 16 h, 7 d
Fitted hyper-parameters	σ_f^2 = 58, ℓ_s = 2.57 km, ℓ_t = 1.77 h
Software	Python 3.13, NumPy 2.5, SciPy 1.18; seed 20261025

VI. Results and Discussion

A. Nowcast accuracy

Table III gives the grid-wide scores. Two baselines bracket the problem. REF-MEAN, which ignores space entirely, reaches 12.7 µg m^{-3}, and REF-GP is indistinguishable from it at 12.7: four references cannot resolve a 12 km city, so the GP collapses to its mean. Adding uncalibrated sensors is harmful, not neutral. RAW-GP and IDW-raw inherit the 80 % sensor over-reading and produce RMSEs of 44.2 and 48.9 µg m^{-3} with negative R^2; a network operator who maps raw low-cost data is worse off than one who maps nothing. DAC-GP brings the RMSE to 11.3 µg m^{-3} (R^2 = 0.67), lowers hotspot RMSE from 30.4 to 27.6 µg m^{-3}, and raises the exceedance F_1 from 0.79 to 0.83.

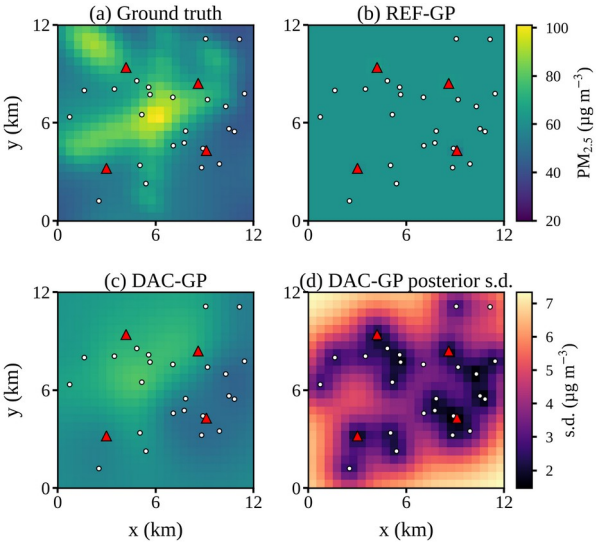


Fig. 2. Snapshot at 21:00 on day 33 (night-time peak, two sensors fouled since day 30). (a) Truth; (b) REF-GP; (c) DAC-GP; (d) DAC-GP posterior standard deviation. Triangles: reference monitors; circles: low-cost sensors. DAC-GP recovers the shape of the field at about half its amplitude; the posterior s.d. is lowest around references and sensors and rises in the unsampled corners.

TABLE III. Grid-wide nowcast performance after burn-in (33 days × 576 cells)

Method	RMSE	MAE	R^2	Cov_{90}	Hotspot RMSE	F_1
REF-MEAN	12.7	9.2	0.58	–	–	–
IDW-raw	48.9	38.5	-5.19	–	32.3	–
REF-GP	12.7	9.2	0.58	0.69	30.4	0.79
RAW-GP	44.2	33.0	-4.06	0.17	35.8	0.73
STATIC-GP	11.9	8.3	0.63	0.56	27.7	0.80
DAC-GP (no RH)	13.6	9.5	0.52	0.49	28.8	0.78
DAC-GP	**11.3**	**7.7**	**0.67**	0.58	**27.6**	**0.83**
ORACLE-GP	9.8	6.3	0.75	0.70	24.3	0.87

RMSE, MAE and hotspot RMSE in µg m^{-3}. Cov_{90}: empirical coverage of the nominal 90 % interval. F_1: hourly exceedance of 60 µg m^{-3}. ORACLE-GP uses perfectly calibrated sensors and bounds what calibration can deliver at these 24 sites.

The oracle row is the most informative one. With every sensor perfectly corrected the same GP reaches 9.8 µg m^{-3}; DAC-GP recovers 48 % of the gap between the reference-only and oracle solutions. The remaining 9.8 µg m^{-3} is not a calibration problem. It is the error of a GP with a 2.6 km length scale trying to represent 0.5–0.9 km plumes from 28 points, and no correction algorithm will remove it; only more sites, or a source-aware mean function, can. Fig. 2 makes the point visually. At 21:00 on day 33 the reference-only map spans only 54–67 µg m^{-3} across the city. DAC-GP lifts the corridor junction to 71 against a true 95 and the kiln cluster to 66 against 85, lowers the clean south-west and north-east corners to 56 against 45–51, and misses the open-burning plume (60 against 65) because no sensor sits within its 0.7 km footprint. The shape is right; the amplitude is roughly halved.

B. Calibration tracking

Fig. 3 follows the gain estimates. Starting from a prior of b = 1, which is 34 % below the true population mean, the estimates settle within the first two days and the slow trends are followed thereafter. The gains are not recovered exactly. Over the final ten days the mean absolute relative gain error

across the 24 sensors is 11 %, with individual ratios $\hat{b}_i/b_i$ between 0.75 and 1.25 and no population bias (mean ratio 1.04). Two mechanisms account for the scatter. One is the offset–gain trade-off: where the concentration stays in a narrow band for days, $y = a + b\,h\,c$ is nearly collinear in (1, hc) and part of the gain migrates into the offset; the three sensors with the largest offset estimates (15–28 µg m^{-3} against true values of 4–6) are exactly those with gain ratios of 0.75–0.85. The other is representativeness: a sensor on a plume is compared with a posterior built from sites off the plume. Neither damages the quantity that matters for the map. The corrected reading $(y - \hat{a})/(\hat{b}h)$ at the sensor cells has an RMSE of 8.2 µg m^{-3} over the last ten days (bias -2.5), against 54.9 for raw readings, 12.7 for the frozen calibration and 4.0 for a sensor whose true parameters are known. The fouling step on day 30 is absorbed: the gain estimate of sensor 2 falls by 20 % of its pre-step value after 55 h and that of sensor 1 after 100 h, with the slower of the two sitting in a sparsely sampled part of the domain where s_i^2 is large and the Kalman gain correspondingly small. A larger Q would shorten the response at the cost of noisier estimates in quiet periods; Q was not tuned to this event.

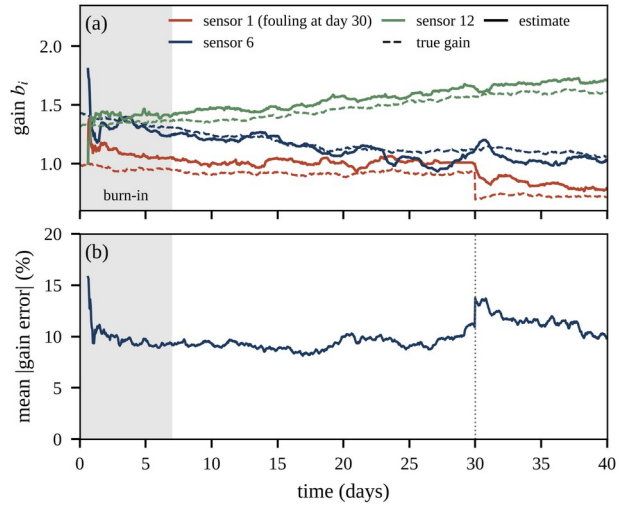


Fig. 3. Gain tracking. (a) True (dashed) and estimated (solid) gain for three sensors, including sensor 1 with a 25 % fouling step at day 30; grey band is the burn-in. (b) Mean absolute relative gain error across all 24 sensors.

C. Static calibration and humidity

STATIC-GP, the co-location-once practice, scores 11.9 µg m^{-3} on the main run, only 5 % behind DAC-GP. With ±18 % drift over 40 days, a calibration fitted in week one is still serviceable in week six, and the drift narrative should be tempered accordingly. The drift sweep in Fig. 4(c) shows where the two part company. At twice the nominal drift STATIC-GP reaches 13.9 µg m^{-3} while DAC-GP stays at 11.6, and with zero drift the two are equal (11.1 versus 11.1), which confirms that the recursion costs nothing when there is nothing to track. The humidity term, by contrast, is not optional. Removing it raises the RMSE to 13.6 µg m^{-3}, worse than REF-GP. The reason is structural: a time-varying multiplicative error with a 24-h cycle cannot be absorbed by a random-walk gain whose process noise is tuned for slow drift, so the sensors oscillate around the truth and the GP, which believes their stated variance, follows them.

D. Short-range forecast, sensor count and LORO

Table IV reports the τ = 1 h and 3 h forecasts. DAC-GP is indistinguishable from persistence of its own nowcast at 1 h (12.7 against 12.7 µg m^{-3}) and marginally ahead at 3 h (16.3 against 16.5); the small 3 h margin comes from the temporal kernel decaying the anomaly toward the background, not from dynamical skill, and a forecast product would need a model of $m(t)$. Fig. 4(a) shows the error by hour of day. Both methods are most accurate in the early afternoon, when the mixed layer flattens the field, and least accurate through the morning inversion. The sensors buy a fairly uniform 1.5–1.8 µg m^{-3} from 05:00 to 20:00 but only 0.8–1.2 µg m^{-3} between 21:00 and 03:00: the night-time open-burning plume is narrower than the sensor spacing, and neither method sees it. Fig. 4(b) sweeps the sensor count. With eight sensors DAC-GP already improves on REF-GP (11.8 µg m^{-3}); the gain from 24 to 40 sensors is small (11.3 to 11.0), consistent with the representativeness floor seen in the oracle. RAW-GP worsens with every added sensor. The LORO test, which is what a practitioner could compute, is where the method looks strongest: at held-out reference sites the DAC-GP nowcast has RMSE 6.8 µg m^{-3} against 9.7 for REF-GP, a 30 % reduction. The two numbers differ from the grid-wide ones because reference sites are, by design, not on hotspots.

TABLE IV. Short-range forecast RMSE (µg m^{-3}) and leave-one-reference-out (LORO) error

Method	τ = 1 h	τ = 3 h	LORO RMSE	LORO bias
Persistence of nowcast	12.7	16.5	–	–
REF-GP	13.5	16.5	9.7	0.00
DAC-GP	**12.7**	**16.3**	**6.8**	-1.22
ORACLE-GP	11.6	15.9	–	–

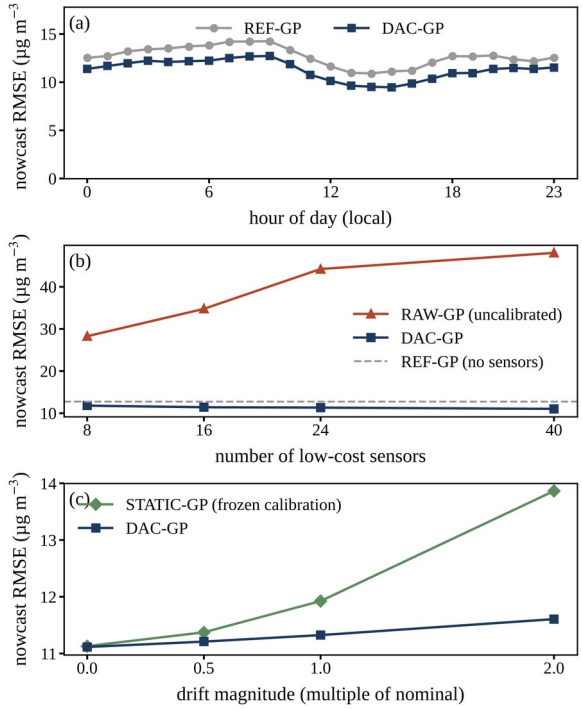


Fig. 4. (a) Nowcast RMSE by hour of day for REF-GP and DAC-GP. (b) Sensor-count sweep; REF-GP (dashed) uses no sensors. (c) Drift-magnitude sweep: STATIC-GP freezes the calibration after burn-in, DAC-GP keeps tracking.

E. Negative findings

Three results argue against over-selling the method. First, the 90 % predictive intervals are under-dispersed everywhere: DAC-GP covers 58 % of the truth and even the oracle covers only 70 %, so the shortfall is a property of the stationary GP prior facing non-stationary plumes rather than of the calibration. REF-GP's higher coverage (69 %) is not a virtue; its intervals are wider because it knows less. A variance inflation factor fitted on LORO residuals, or a non-stationary kernel, is the obvious remedy and was not implemented here. Second, the grid-wide RMSE improvement over REF-GP is 11 %, modest next to the 30 % LORO improvement; readers comparing with studies that report only station-level cross-validation should keep the distinction in mind. Third, the single-window marginal-likelihood fit returned $\ell_s = 2.6$ km, which reproduces the 2.5 km of the stochastic residual but not the 0.5–0.9 km plumes; one stationary length scale cannot serve both, and re-estimating it every hour would not change that, since the limitation is the kernel family rather than its value.

VII. Limitations

The study is a simulation. The CPCB and OSPCB statistics anchor the magnitude and seasonality of the field, not its spatial structure, which was designed from the source types known to exist in the city rather than from measurements. The sensor model is affine with a single κ; real optical sensors show particle-size and composition sensitivity that an affine map cannot capture [7], [12]. Stuck or saturated sensors are not simulated. The reference-mean background $m(t)$ assumes that the references are spatially representative of the city mean, which is untrue where a single CAAQMS sits beside a road. The exact GP limits the window to around 500 points; a sparse variational GP [9] would be needed for a network of several hundred sensors. Finally, the results are one realisation of one random seed with fixed source geometry; the numbers in Tables III and IV should be read as such.

VIII. Conclusion

Drift correction and spatial nowcasting of low-cost $PM_{2.5}$ networks are one estimation problem, and solving them together removes the need for co-location: each sensor is calibrated against the GP's view of its own location, with the reference monitors anchoring the scale of the whole network (Proposition 1). In a CPCB-parameterised winter simulation the scheme lowers grid RMSE by 11 % and held-out reference error by 30 % relative to a reference-only GP, cuts the corrected-reading error of a calibration frozen after one week by 36 %, absorbs a 25 % fouling step within two to four days, and runs at 31 ms per hourly cycle. The same experiments show that uncalibrated sensors degrade a map rather than improve it, that the humidity term is indispensable, that a static calibration survives mild drift, and that beyond roughly 24 sensors the limiting error is network geometry, not sensor quality. Deployment on a real CAAQMS-plus-sensor network in Bhubaneswar, with the LORO protocol as the acceptance test and a variance-inflation step for the intervals, is the natural next stage.

References

[1] Central Pollution Control Board, *National Ambient Air Quality Standards*, Notification B-29016/20/90/PCI-L, New Delhi, India, Nov. 2009.

[2] Odisha State Pollution Control Board, “Ambient air quality status of different cities and towns of Odisha during 2023,” Bhubaneswar, India, 2024.

[3] S. K. Guttikunda et al., “Air pollution knowledge assessment (APnA) city program: Bhubaneswar,” UrbanEmissions.info, 2017.

[4] N. Zimmerman et al., “A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring,” *Atmos. Meas. Tech.*, vol. 11, no. 1, pp. 291–313, 2018.

[5] C. Malings et al., “Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation,” *Aerosol Sci. Technol.*, vol. 54, no. 2, pp. 160–174, 2020.

[6] K. K. Barkjohn, B. Gantt, and A. L. Clements, “Development and application of a United States-wide correction for $PM_{2.5}$ data collected with the PurpleAir sensor,” *Atmos. Meas. Tech.*, vol. 14, no. 6, pp. 4617–4637, 2021.

[7] J. Tryner et al., “Laboratory evaluation of low-cost PurpleAir PM monitors and in-field correction using co-located portable filter samplers,” *Atmos. Environ.*, vol. 220, Art. no. 117067, 2020.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[9] J. Hensman, N. Fusi, and N. D. Lawrence, “Gaussian processes for big data,” in *Proc. 29th Conf. Uncertainty in Artificial Intelligence (UAI)*, Bellevue, WA, USA, 2013, pp. 282–290.

[10] A. Datta et al., “Statistical field calibration of a low-cost $PM_{2.5}$ monitoring network in Baltimore,” *Atmos. Environ.*, vol. 242, Art. no. 117761, 2020.

[11] N. Castell et al., “Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?” *Environ. Int.*, vol. 99, pp. 293–302, 2017.

[12] L. R. Crilley et al., “Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring,” *Atmos. Meas. Tech.*, vol. 11, no. 2, pp. 709–720, 2018.

[13] P. Kumar et al., “The rise of low-cost sensing for managing air pollution in cities,” *Environ. Int.*, vol. 75, pp. 199–205, 2015.

[14] Y. Zheng, F. Liu, and H.-P. Hsieh, “U-Air: When urban air quality inference meets big data,” in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 1436–1444.

[15] Y. Cheng et al., “AirCloud: A cloud-based air-quality monitoring system for everyone,” in *Proc. 12th ACM Conf. Embedded Network Sensor Systems (SenSys)*, Memphis, TN, USA, 2014, pp. 251–265.

[16] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.