Artificial Intelligence and Internet of Things for Air Pollution Monitoring and Prediction: A Critical Narrative Review

Mentor:
Prof. Kedar S. Redekar

Authors:
Vishakha B. Koli

Sanika K. Patil

Prajwal P. Mali

Siddhivinayak V. Patil

Parth K. Suryavanshi

Abstract

Air pollution monitoring is constrained by a persistent tension between the accuracy of reference-grade instruments and the spatial and temporal coverage required to describe neighbourhood-scale exposure. Internet of Things (IoT) networks can increase measurement density, while artificial intelligence (AI) methods can calibrate sensor signals, combine heterogeneous observations, map pollutant fields and forecast concentrations. This narrative review examines how these functions connect, rather than treating sensor deployment and prediction as separate technical problems. Recent field studies show that low-cost sensors can support local monitoring when they are calibrated and evaluated for the intended task; they do not establish that inexpensive devices are interchangeable with regulatory instruments. Machine-learning calibration results vary by pollutant, sensor, meteorology and location, and calibration transfer beyond co-location sites remains a central concern. Deep learning can represent temporal and spatial dependencies, but model complexity alone does not guarantee better forecasts or reliable transfer. The strongest practical direction is therefore a quality-managed pipeline that preserves provenance, validates models across time and sites, reports uncertainty, and communicates whether an output is a measurement, calibrated estimate, or forecast. Research priorities include external validation in undermonitored settings, drift-aware calibration, comparable evaluation protocols, interpretable models, and decision-focused communication. The reviewed evidence supports AIenabled IoT as a complement to established monitoring infrastructure, not a substitute for reference measurements or regulatory oversight.

Keywords: air pollution; Internet of Things; low-cost sensors; machine learning; deep learning; calibration; air quality forecasting; uncertainty

1. Introduction

Reliable air-quality information is needed to characterize pollution patterns, identify locations and periods of elevated concentration, and evaluate environmental interventions. Yet monitoring networks are rarely dense enough to resolve every neighbourhood, street or microenvironment. Reference-grade stations offer measurement quality suited to official monitoring, but their cost and operational demands constrain how many can be installed. In Delhi, for example, a study described 33 reference monitors across approximately 1,500 km² and used two years of data from 28 portable low-cost sensors to generate fine-grained pollution estimates [1]. The contrast illustrates the central engineering problem: sparse high-quality measurements and abundant but

noisier observations each leave an important information gap.

Internet of Things (IoT) technologies address the collection and transmission side of this problem. Distributed nodes can record particulate matter ($PM_{2.5}$ and PM_{10}), gases such as nitrogen dioxide (NO_2) and ozone (O_3), and meteorological variables; networked systems can transmit measurements to local gateways or cloud services. Artificial intelligence (AI), especially machine learning (ML), contributes at several distinct stages: identifying invalid readings, calibrating sensor response, filling spatial gaps, classifying air-quality states, and forecasting future concentrations. These functions are related but are not interchangeable. A calibrated concentration estimate is not the same as a forecast, and neither is automatically a regulatory measurement.

The field has moved from small sensor demonstrations toward networks and integrated analytical services. Early work emphasized how sensor networks could complement established instruments while noting that pollutant spatial homogeneity differs between cities and pollutants [2]. Subsequent research tested calibration transfer across network nodes and investigated probabilistic mapping, dense citywide forecasting, explainability and multi-network data fusion [3], [4]. This development shifts the key question from whether AI and IoT can be connected to whether the resulting outputs remain valid across time, geography, sensor types and operational uses.

This review synthesizes literature on four linked questions: (1) what IoT systems contribute to air-quality data collection; (2) how calibration and data-quality procedures shape the validity of those data; (3) how classical ML, deep learning and spatiotemporal methods support mapping and forecasting; and (4) what validation, interpretability and governance requirements determine whether the outputs can inform decisions. It focuses on ambient outdoor monitoring, while noting indoor and public-facing applications where they illuminate transfer or communication issues. It does not evaluate health effects of pollution or claim that prediction systems themselves reduce exposure.

2. Review approach and scope

A structured narrative review was prepared using the attached draft as the starting point and searching the academic indexes Semantic Scholar, OpenAlex and PubMed. The search was conducted on 3 November 2026. Topic searches covered IoT and low-cost air-quality sensor networks, machine-learning calibration and data quality, spatiotemporal air-quality prediction, explainability and uncertainty, and monitoring-system deployment. The synthesis prioritizes publications dated 2021–2026, including primary field studies and methodological papers, with review articles used to frame the scope of the field. Papers were selected for relevance to sensing architecture, calibration, forecasting or deployment rather than treated as a complete census of all publications.

This is a narrative synthesis, not a registered systematic review or meta-analysis. It does not claim exhaustive retrieval, and no quantitative pooling was attempted because studies differ in pollutant, sensor technology, validation design, forecast horizon and reported metric. The synthesis uses the abstracts available for most studies; accordingly, detailed information absent from those abstracts—such as full preprocessing protocols, complete baselines and subgroup results—is not inferred. Search access was restricted to the selected academic indexes and the supplied draft; web sources, standards databases and regulatory guidance were not searched. Findings are therefore interpreted at the level supported by the available article abstracts and are not presented as a complete assessment of every system or jurisdiction.

3. From sensor node to information service

An IoT air-quality system is best understood as an end-to-end measurement pipeline. At the edge, a sensor converts environmental conditions into electrical or optical signals. A microcontroller timestamps and packages the signals, and a communication link transfers them to a server, database or gateway. The analytical layer checks data completeness and plausibility, applies calibration or quality flags, and may aggregate observations across

sensors. A user-facing layer then presents measured values, estimates, maps, alerts or forecasts. Each transition can introduce error: a poorly placed sensor may not represent the target location; a clock mismatch can distort temporal alignment; communication loss can create gaps; and an unlabelled calibration or forecast can lead users to interpret a model output as a direct observation.

Low-cost sensors are attractive because their price and deployment flexibility can support denser monitoring than a network composed exclusively of expensive reference instruments. The benefit is not simply more devices. A denser network can reveal variation missed by sparse stations, but only if the observations are sufficiently characterized and their uncertainty is visible. A review of sensor-network evidence emphasizes that spatial representativeness is pollutant- and city-dependent: ozone was more spatially homogeneous in the evaluated US contexts, whereas $PM_{2.5}$, CO and NO_2 varied more by city [2]. A uniform network design therefore cannot be assumed to work equally well for all pollutants or urban forms.

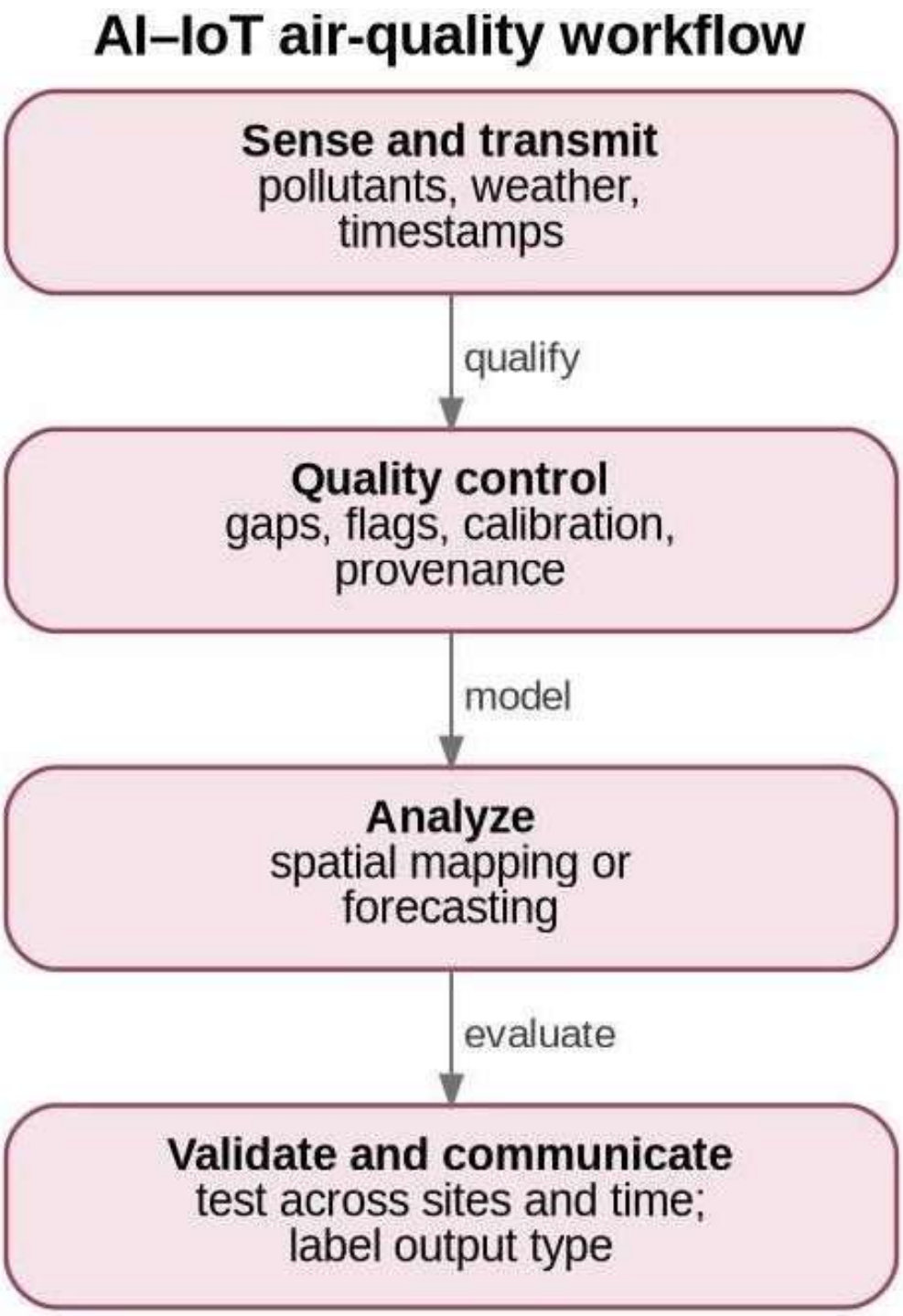
Network design should follow the intended use. A study directed toward hotspot discovery may require broad spatial coverage and repeated measurements near likely sources. A trend-monitoring application needs stable operation over time and attention to sensor drift. An exposure-assessment system requires spatial and temporal coverage appropriate to human activity and a calibration approach that preserves uncertainty. A public warning service additionally needs reliable latency, understandable thresholds and clear rules for missing or suspect data. These different objectives should be stated before selecting hardware, communication technology or algorithms.

Field examples demonstrate why a service-oriented architecture must include data quality rather than end at transmission. A low-cost network study describes calibration as a prerequisite for using measurements in decision-support services and links system development to local municipal needs [5]. A hybrid network study used one public station and ten low-cost devices, and reported that calibration could be propagated from a calibrated low-cost node to an uncalibrated node, reducing reported RMSE for NO_2 and PM_{10} in its deployment [6]. These results suggest that networks can share calibration information, but they do not establish that propagation will work regardless of sensor model, distance, weather or pollution regime.

Communication and storage choices also shape practical performance. Continuous transmission enables near-real-time visualization but can be limited by connectivity, energy consumption and operating cost. Local processing can screen implausible values or reduce data transfer, while centralized processing can support model updates and comparisons across sites. A system should preserve raw measurements alongside corrected outputs, calibration version, timestamps, sensor identity and quality flags. Without this provenance, later recalibration, audit or comparison to a reference instrument becomes difficult. These recommendations follow from the literature's recurring distinction between raw sensor outputs and calibrated data, and from its concern about the transferability of models beyond the co-location setting [3].

The overall lesson is that IoT creates the possibility of high-resolution monitoring, not an assurance of high-quality monitoring [5], [7]. Deployment studies emphasize that decision-support services depend on accurate, timely input data and describe calibration as necessary before low-cost measurements are used [5]. Low-cost sensors can add spatial sampling, but their readings require processing and calibration to support reliable interpretation [3], [7].

Figure 1 summarizes the end-to-end logic supported by sensor-network deployments and calibration research: data collection must retain quality and provenance before analytical outputs are interpreted [3], [5], [7].



Conceptual pipeline from sensing and transmission through quality control, modelling, validation and communication. The sequence is a proposed synthesis of the reviewed system components.

4. Calibration, sensor reliability and data quality

Calibration is the bridge between a low-cost device's raw signal and an estimate intended to approximate a reference measurement. Optical particulate sensors, electrochemical gas sensors and other devices respond differently to environmental conditions and cross-sensitivities. Relative humidity and temperature can alter sensor response, while ageing, contamination and device-to-device variation can change it over time. A model fitted under one combination of sensor, pollutant and environment may therefore perform poorly after relocation or extended operation.

Field calibration studies show both the potential and the conditional nature of ML. In a large-scale AirQo deployment in Sub-Saharan urban settings, Random Forest performed best for $PM_{2.5}$ calibration and Lasso

regression for PM_{10}; the reported $PM_{2.5}$ RMSE fell from 18.6 to 7.2 µg/m³, while PM_{10} RMSE fell from 13.4 to 7.9 µg/m³. The study used cross-unit and cross-site validation and deployed the resulting approach across a

network of more than 120 devices [8]. This is an important practical example because the evaluation considered deployment across a network rather than only model fit at a single calibration location. It also cautions against assuming that one algorithm is optimal for multiple pollutants.

In a separate field evaluation, two low-cost sensor systems were compared with an urban standard station over October–December 2023. The sensors showed systematic overestimation, particularly for PM_{10}, and Random Forest and XGBoost outperformed other tested calibration models in the reported test comparison [9]. The difference between this result and the AirQo study is not necessarily a contradiction: the sensor systems, settings, period and targets differed. Rather, both studies reinforce the need to assess the complete sensor–model combination, including the pollutant and field conditions, rather than rank algorithms in isolation.

Calibration transfer remains a particularly consequential limitation because co-location performance does not establish network-wide validity [3]. Calibration models are often trained at locations where low-cost and reference monitors are co-located, but performance at those sites may not predict performance elsewhere in the network [3]. A Denver network study explicitly examined transferability and hotspot-detection implications, identifying sparse evidence on how co-location duration and site representativeness affect downstream use [3]. Another study used a hybrid network with one public monitoring station and ten low-cost nodes to test calibration propagation across devices; its improvements were pollutant-specific and reported in the deployed network rather than establishing universal transferability [6]. Together, these studies make geographic and temporal validation central design requirements, not optional refinements.

Uncertainty-aware calibration is valuable where decisions depend on whether a concentration crosses a threshold. A Baltimore study trained probabilistic gradientboosted models directly on raw $PM_{2.5}$ measurements and used them to create neighbourhood-scale probabilistic exposure surfaces. The approach reportedly improved point and distributional accuracy over a linear model, particularly above 25 µg/m³ and at monitors excluded from training; its output represented the probability of exceeding an exposure threshold rather than only a single concentration estimate [10]. This illustrates a shift from asking only "How close is the point prediction?" to asking "How uncertain is the estimate, and what is the probability that a decision-relevant threshold is exceeded?"

Long-term reliability requires repeated quality checks. In Zagreb, a city-wide evaluation involved 35 sensor locations, one local reference station and three national reference stations for PM_{10} and NO_2. XGBoost calibration using spatiotemporal, meteorological and traffic proxy features reduced RMSE by up to 82% in the reported assessment; a rollingerror analysis suggested recalibration was needed within one to six months [11]. The interval should be interpreted as a result of that particular network and study, not as a universal recalibration schedule. Nonetheless, it demonstrates why deployment protocols should define drift monitoring and recalibration triggers in advance.

Recent multi-city work also indicates that the best calibration method can vary by pollutant, sensor and reference instrument. A 2026 study compared PurpleAir and OPCN3 devices in Antwerp, Oslo and Zagreb and reported different performance patterns for $PM_{2.5}$ and PM_{10}, with a hybrid calibration approach competitive with Random Forest in some comparisons but not uniformly superior [12]. This is a more informative conclusion than a broad claim that hybrid models dominate: calibration selection is conditional, and robust assessment should report pollutant- and site-specific results.

Thus, ML can improve low-cost sensor data, but the evidence supports calibrated, application-specific use—not the substitution of model outputs for instrument validation. Stronger practice combines reference co-location, cross-site and time-separated validation, bias and error reporting, drift surveillance, and uncertainty estimates. A single random train–test split may not adequately represent the spatial or temporal transfer a

network will face.

5. Machine learning for mapping and forecasting

5.1 Baselines and model selection

Air-quality tasks include regression on pollutant concentrations, classification into categories, anomaly detection, spatial interpolation and forecasting. Linear regression is valuable as a transparent baseline and can be adequate when relationships are approximately linear and the data regime is limited. Tree-based methods can capture nonlinear interactions and often require less feature engineering than deep neural networks. Support Vector Regression may be useful for nonlinear prediction with appropriate kernels, while neural networks can learn complex representations from larger multivariate datasets. None of these method families is inherently best: the appropriate comparison depends on the target, data volume, deployment hardware, forecast horizon and validation design.

A systematic comparison of eleven ML models and five software packages for PurpleAir $PM_{2.5}$ calibration found that the software implementation affected reported accuracy. LSTM obtained R^2 values around 0.856–0.857 and RMSE values around 4.25–4.26 µg/m³ in that study, but its training time was reported as 1.5 hours; Random Forest and XGBoost had slightly lower performance metrics but trained in under a minute [13]. This comparison illustrates a practical trade-off: an offline calibration workflow may tolerate longer training, while an operational system needing rapid updates may prioritize computational efficiency. The same source also warns that model comparisons depend on implementation choices, not only algorithm labels.

5.2 Temporal models and short-horizon prediction

Air-pollution observations are time series influenced by persistence, daily cycles, meteorology and changing emissions. Recurrent neural networks, including Long ShortTerm Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, are designed to represent sequential dependencies. Their use is common, but high performance in a temporal test set does not establish transfer to another city or season. Forecast horizons should be explicit because errors typically become more consequential as prediction lead time increases, and a model optimized for one-hour prediction need not be useful for a day-ahead warning.

A low-cost IoT system for indoor ambient-assisted living combined calibration with a multi-headed convolutional/GRU model and transfer learning. Its authors reported a 55.42% improvement in RMSE for prediction at a new target system with insufficient local data [14]. While the setting is indoor and thus not directly interchangeable with outdoor city monitoring, the study addresses a transferable problem: new locations often lack a long local training history. Transfer learning may reduce this data barrier, but its safe use depends on whether the source and target environments share relevant pollution dynamics and sensor characteristics.

At a broader spatial scale, one study used 417 micro-monitoring stations in Lanzhou, China, to produce $PM_{2.5}$ forecasts at 500 m × 500 m spatial and one-hour temporal resolution. Its spatiotemporal model reported RMSE of 2.12 µg/m³ for a one-hour horizon and 4.09 µg/m³ for a six-hour horizon [15]. The increasing error with horizon is consistent with the added uncertainty of longer forecasts. However, these values should not be compared directly with error metrics from other cities without accounting for pollutant distributions, measurement sources, aggregation interval and evaluation design.

A 2022 study for NEOM City combined residual feature extraction with ConvLSTM for spatiotemporal prediction and reported next-hour MASE values of 9.13 for PM_{10} and 13.57 for $PM_{2.5}$ [16]. Because the abstract does not provide enough common evaluation detail to align these values with the Lanzhou RMSE, the studies are best read side by side as demonstrations of different architectures and locations, not as a league table. The key synthesis is that temporal deep learning can exploit sequential structure, while its performance remains tied to the data setting and metric.

5.3 Spatial structure, sensor fusion and unmonitored locations

A city is not a single time series. Nearby sites can share meteorological and transport influences, but local emissions, street geometry and sensor placement can produce sharp gradients. Spatial prediction methods therefore seek to combine observations across monitoring locations, often using convolutional networks, graph structures, recurrent networks or geostatistical methods. The benefit is especially clear where the network is sparse or where forecasts are needed between fixed monitors.

In Delhi, a model combining message-passing recurrent neural networks with spatiotemporal geostatistics used two years of data from 28 portable low-cost sensors. The study reported mean absolute percentage errors of 9.4% over low-cost monitor locations, 10.5% over reference monitors and 9.6% over the combined network for onehour estimates [1]. This provides evidence that complementary statistical and neural components can address high variability and intermittent sensor data. It does not demonstrate that an equivalent error would hold for other pollutants, seasons or cities.

Forecasting for locations without sensors is a related but distinct problem. A study in Expert Systems with Applications compared ML and deep-learning approaches for unmonitored urban locations and found that the best-performing model family differed by pollutant: LSTM for O_3, a multipath CNN–BiLSTM for NO_2, GRU for $PM_{2.5}$ and BiLSTM for PM_{10} [17]. This result argues against a single universal architecture and emphasizes the need to state the pollutant-specific target. The study's comparison also highlights that spatial inference is not simply temporal forecasting at a new point; it must account for what is known about nearby sites and how representative they are.

For a citywide application, an integrated deep-learning study reported more than 31% improvement in prediction accuracy while maintaining approximately 80% prediction coverage for 24-hour forecasts in two real-world construction and infrastructure settings [18]. The combination of coverage and accuracy is important: a model that produces precise estimates only for a subset of observations may not be operationally useful, while a model that produces complete but poorly calibrated outputs may mislead. Future evaluations should report coverage, error and uncertainty together.

5.4 Emerging architectures and interpretability

Newer model families include attention-based networks, graph neural networks and transformer variants [19], [20], [21]. These architectures introduce different ways to represent spatial and temporal structure, while predictive performance and feature interpretation remain distinct questions to evaluate [19], [21]. A 2024 study proposed an interpretable spatiotemporal model using feature selection and channel attention for forecasting across the Beijing–Tianjin–Hebei region, combining Shapley-based analysis with attention mechanisms [19]. A 2025 study proposed a spatiotemporal inverted Transformer using multistation pollutant and meteorological inputs, and reported improvements over its benchmarks in experiments in Beijing [21]. Such results show active methodological development; they do not by themselves establish that attention weights or feature-attribution scores are causal explanations.

A 2025 hybrid framework combined LSTM, CNN and graph neural networks and reported distinct performance on Beijing data and a cross-dataset test trained on Beijing and tested on Los Angeles [20]. Cross-dataset testing is more informative about transfer than a random split within one city, although one source–target pair does not establish general global transportability. Separately, a study using convergent cross-mapping for feature selection reported lower test-set mean squared error than SHAP-based feature selection for PM_1 and $PM_{2.5}$ calibration in its OPC-N3 evaluation [22]. The comparison demonstrates that interpretability methods and predictive performance can depend on the feature-selection strategy and test setting.

Model explanations should be evaluated as part of the scientific claim. Feature importance can describe associations learned by a predictive model, but it does not prove that changing a feature would change pollution. Attention weights are internal model operations, not automatically interpretable causal effects. More defensible interpretability combines domain-

relevant variables, sensitivity analyses, stability checks across sites and time, and clear disclosure of what the explanation method can and cannot infer.

6. Comparative evidence and evaluation standards

The following studies illustrate different objectives, data designs and reported metrics. The numbers are reproduced in the framing reported by each source and are not pooled because their pollutants, sensors, geographic settings, horizons and metrics differ.

Study and setting	Data and method	Reported result	What it supports—and what it does not
Delhi, fine-scale mapping [1]	28 low-cost sensors; two years; messagepassing recurrent model plus geostatistics	One-hour MAPE: 9.4% at low-cost sites, 10.5% at reference sites	Supports dense local mapping in the studied Delhi network; not a universal accuracy benchmark.
AirQo field calibration [8]	More than 120 devices; reference co-location; crossunit and cross-site validation	$PM_{2.5}$ RMSE: 18.6 to 7.2 µg/m³; PM_{10}: 13.4 to 7.9 µg/m³	Supports field ML calibration and network deployment in the studied Sub-Saharan context; model choice differed by pollutant.
Lanzhou citywide forecast [15]	417 micro-stations; spatiotemporal deep learning	$PM_{2.5}$ RMSE: 2.12 µg/m³ at 1 h; 4.09 µg/m³ at 6 h	Supports citywide forecasting at the stated resolution; horizon and metric prevent direct comparison to MAPE studies.
Baltimore probabilistic exposure [10]	Field-calibrated probabilistic gradient boosting; reference and network data	Improved point and distributional accuracy, notably above 25 µg/m³	Supports thresholdoriented uncertainty estimates; performance is tied to the Baltimore data and calibration design.

Study and setting	Data and method	Reported result	What it supports—and what it does not
Zagreb sensor network [11]	35 low-cost locations and reference stations; spatiotemporal XGBoost calibration	RMSE reduction up to 82%; recalibration indicated within 1–6 months	Supports drift-aware recalibration in that deployment; not a general maintenance interval.
Three-city sensor calibration [12]	PurpleAir and OPCN3; $PM_{2.5}$ and PM_{10}; three European cities	Hybrid and RF models varied in relative performance across pollutant, city and instrument	Supports conditional, multi-city evaluation rather than a single "best" model.

A credible evaluation should define the prediction target, pollutant, spatial unit, temporal aggregation, horizon and intended use, because the reviewed calibration, mapping and forecasting studies address distinct tasks [1], [3], [15]. For sensor calibration, the target is a reference instrument measurement; for forecasting, it is a future concentration or category; for mapping, it may be a spatial estimate at an unmonitored location. These tasks need different baselines and data splits. Reported metrics should include interpretable errors such as MAE or RMSE, relevant scale information, and complementary measures of bias, calibration or uncertainty. R^2 alone is insufficient because it does not communicate absolute error and can be influenced by the range of observations.

Validation should reflect intended deployment, because calibration performance at colocation sites may not transfer to the rest of a sensor network [3]. A Denver network study specifically identifies transferability as an evaluation concern and proposes metrics for assessing calibration models beyond their co-location sites [3]. The AirQo study also reports cross-unit and cross-site validation as part of its field calibration [8]. For forecasting systems, evaluations should separately report performance on later time periods; for expanding networks, they should report results at locations withheld from model fitting. These are recommended evaluation designs for this review, not a claim that a common protocol has already been adopted.

Results should also be stratified where plausible error mechanisms differ: pollutant level, season, humidity range, sensor age, site type and pollution episode [9], [11]. Aggregate accuracy can obscure concentration-specific behavior: the Baltimore probabilistic calibration study found stronger comparative benefit above 25 $\mu g/m^3$, while the Zagreb study reported larger gains during pollution peaks [10], [11]. These findings support testing performance across concentration regimes, although each threshold and setting requires local justification [10], [11].

Comparability across publications remains limited. Studies may report RMSE, MAPE, MASE, R^2 or classification accuracy, use different pollutants and reference methods, and evaluate different time horizons. Data can represent hourly averages, short-interval readings or aggregated maps. Therefore, a lower numerical error in one paper does not establish a superior model without aligned targets, measurement units, settings and test protocols. The useful comparative conclusion is methodological: transparent, leakageresistant and externally relevant evaluation is more informative than isolated benchmark rankings.

7. Deployment, public communication and responsible use

A technically accurate forecast becomes useful only when it is delivered at the right time and in a form that supports an appropriate action. Public dashboards can show pollutant values, trends, update time, location and alert categories. They should distinguish observed values from calibrated estimates and forecasts, indicate sensor status, and avoid false precision.
The system should communicate uncertainty and explain that a low-cost sensor can indicate local patterns without automatically satisfying regulatoryequivalence requirements. A sensor-network study explicitly notes that low-cost measurements require additional processing and do not inherently meet regulatory reference-equivalent performance [7].

Alerts should be tied to a defined decision. A warning to the public, a research hotspot map and a compliance assessment have different error costs. Missing an episode can delay protective action; a false alarm can weaken trust or prompt unnecessary responses.
Threshold selection should therefore be explicit, and predicted categories should be evaluated for false positives and false negatives at the intended lead time. Probabilistic estimates are especially useful when the decision concerns threshold exceedance, as demonstrated by neighbourhood-scale exposure mapping [10].

Data governance matters because sensor networks can reveal spatial patterns associated with particular communities, roads or facilities. Deployment plans should document locations, ownership, data access, maintenance responsibility and procedures for correcting known errors. Public-facing systems should not overstate what the underlying measurements can establish about individual exposure or health outcomes. A city-level ambient estimate cannot be treated as a personal exposure measurement unless the model and data explicitly represent a person's movement and microenvironment.

The broader AI-IoT literature identifies a recurring tension between model performance, interpretability, data requirements and information sharing [23]. For operational systems, a dependable simpler model may be preferable to a complex model whose performance cannot be reproduced or whose failure modes are unknown. The decision criterion should be fitness for purpose: accuracy and uncertainty at the relevant spatial and temporal scale, stable operation, reasonable computational cost, and a comprehensible path for auditing output.

8. Research gaps and priorities

First, external validation remains uneven in the retrieved studies [3], [12], [17]. Studies demonstrate strong performance in particular networks, but calibration transfer and cross-city prediction remain study-specific questions rather than evidence of universal transportability [3], [12], [17]. A useful next step is prospective multi-city research that predefines a source-to-target transfer protocol and reports outcomes without retraining, after limited local recalibration, and after full local calibration [3], [12], [17]. This would clarify when sensor and model transfer is feasible and what adaptation costs are required. Existing work on calibration transfer, multi-city models and unmonitored locations provides a foundation, but it does not yet justify universal transfer claims [3], [12], [17].

Second, long-term drift and maintenance need shared operational benchmarks. The Zagreb study's one-to-six-month recalibration estimate and the multi-city sensor comparison literature show that model stability is part of the measurement system, not a one-time preprocessing step [11], [12]. Studies should report time since calibration, sensor replacement, data downtime and error trajectories, then evaluate trigger rules for recalibration. This will help distinguish whether accuracy declines because of sensor ageing, changing environmental conditions, model mismatch or reference-instrument differences.

Third, uncertainty should be integrated into both calibration and forecasting. Probabilistic calibration can support threshold-risk estimates, but many systems still report a single point estimate or headline metric. Future work should test interval coverage, calibration of predictive probabilities and decision performance across high-pollution episodes. Uncertainty should be decomposed where possible into measurement noise, model

uncertainty, spatial interpolation uncertainty and data-gap effects. An uncertainty interval that widens when the model is extrapolating can be more honest and actionable than a precise-looking point forecast.

Fourth, comparability requires more consistent evaluation. Researchers should publish data-split logic, reference method, sensor model and firmware, calibration duration, pollutant unit, time resolution, forecast horizon, missing-data treatment and baseline results. Metrics should be reported by pollutant and relevant concentration range, alongside deployment and computational costs. A shared benchmark spanning distinct climates, urban forms and sensor types would make it easier to assess whether claimed improvements are reproducible beyond one city.

Fifth, interpretability requires stronger evaluation than presenting feature-importance plots [19], [22]. A useful explanation should be stable under reasonable data perturbations, understandable to intended users and clearly separated from causal inference. The studies using SHAP-based feature analysis and causal feature selection indicate promising methodological paths, while also showing that different selection strategies can produce different model results [19], [22]. Research should test explanation fidelity and decision usefulness, not only whether an explanatory graphic can be generated.

Sixth, research should continue to evaluate under-monitored and resource-constrained settings. AirQo's large-scale deployment demonstrates that calibration and network operation can be studied in Sub-Saharan urban contexts [8]. Future programmes should budget for maintenance, reference co-location, training, reliable connectivity and local data stewardship rather than treating sensor purchase as the whole intervention. Lowcost systems can extend observational coverage, but their long-term value depends on sustainable operation and local institutional capacity.

Finally, research should evaluate the information service, not only the prediction model. User studies and operational pilots can determine whether maps are understood, whether uncertainty labels are interpreted correctly, whether alerts arrive in time and whether the information changes a decision. Public dashboards should be evaluated for accessibility, language and trust. These are empirical questions; model accuracy alone does not demonstrate public-health benefit or effective environmental governance.

9. Conclusion

AI and IoT can strengthen air-pollution monitoring by joining distributed sensing with calibration, spatial inference, forecasting and accessible information services. The reviewed studies establish that machine learning can improve low-cost sensor estimates and that spatiotemporal models can produce useful local forecasts in particular deployments. They also show that performance varies with pollutant, sensor, location, validation design and forecast horizon. The strongest conclusion is therefore conditional: these systems can complement sparse reference networks when measurements are calibrated, model transfer is tested, uncertainty and data quality are visible, and outputs are matched to a defined decision. Future progress depends less on adopting the newest architecture than on reproducible field evaluation, long-term maintenance, transparent reporting and responsible communication. AI-enabled sensor networks should extend the evidence available to communities and planners while preserving the distinct role of reference measurements and regulatory monitoring.

References

[1] S. R. Iyer et al., “Modeling fine-grained spatio-temporal pollution maps with lowcost sensors,” npj Climate and Atmospheric Science, 2022, doi: 10.1038/s41612-02200293-z.

[2] Y. Li et al., “From air quality sensors to sensor networks: Things we need to learn,” Sensors and Actuators B Chemical, 2021, doi: 10.1016/j.snb.2021.130958.

[3] P. deSouza et al., “Calibrating networks of low-cost air quality sensors,” Atmospheric measurement techniques, 2022, doi: 10.5194/amt-15-6309-2022.

[4] C. Heffernan, K. Koehler, D. R. Gentner, R. D. Peng, and A. Datta, “Unified calibration and spatial

mapping of fine particulate matter data from multiple low-cost air pollution sensor networks in baltimore, maryland," Environmetrics, 2026, doi: 10.1002/env.70100.

[5] T. Veiga et al., "From a low-cost air quality sensor network to decision support services: Steps towards data calibration and service development," Sensors, 2021, doi: 10.3390/s21093190.

[6] I. Vajs, D. Drajic, and Z. Cica, "Data-driven machine learning calibration propagation in a hybrid sensor network for air quality monitoring," Sensors, 2023, doi: 10.3390/s23052815.

[7] S. Felici-Castell, J. Segura-Garcia, J. Solano, R. Fayos-Jordan, A. Soriano-Asensi, and J. Alcaraz-Calero, "AI-IoT low-cost pollution-monitoring sensor network to assist citizens with respiratory problems," Italian National Conference on Sensors, 2023, doi: 10.3390/s23239585.

[8] P. Adong, E. Bainomugisha, D. Okure, and R. Sserunjogi, "Applying machine learning for large scale field calibration of low-cost PM 2 .5 and PM 10 air pollution sensors," Applied AI Letters, 2022, doi: 10.1002/ail2.76.

[9] Z. Feng, L. Zheng, B. Ren, D. Liu, J. Huang, and N. Xue, "Feasibility of low-cost particulate matter sensors for long-term environmental monitoring: Field evaluation and calibration." Science of the Total Environment, 2024, doi: 10.1016/j.scitotenv.2024.174089.

[10] A. Patton et al., "Non-linear probabilistic calibration of low-cost environmental air pollution sensor networks for neighborhood level spatiotemporal exposure assessment," Journal of Exposure Science and Environmental Epidemiology, 2022, doi: 10.1038/s41370-022-00493-y.

[11] V. Petrić et al., "Assessment of sensor data from an air quality monitoring network — the need for machine learning-based recalibration and its relevance in health impact analysis of local pollution events," Atmosphere, 2025, doi: 10.3390/atmos16121358.

[12] I. Yaqoob, S. Nepal, V. Kumar, and S. A. Chaudhry, "Multi-city bi-pollutant calibration of low-cost PM sensors using machine learning," IEEE Access, 2026, doi: 10.1109/access.2026.3656120.

[13] S. Smith et al., "A systematic study of popular software packages and AI/ML models for calibrating in situ air quality data: An example with purple air sensors." p. 1028, 2025. doi: 10.3390/s25041028.

[14] S. Sonawani and K. Patil, "Air quality measurement, prediction and warning using transfer learning based IOT system for ambient assisted living," International Journal of Pervasive Computing and Communications, 2023, doi: 10.1108/ijpcc-07-2022-0271.

[15] R. Guo, Q. Zhang, X. Yu, Y. Qi, and B. Zhao, "A deep spatio-temporal learning network for continuous citywide air quality forecast based on dense monitoring data," Journal of Cleaner Production, 2023, doi: 10.1016/j.jclepro.2023.137568.

[16] K. Elbaz, "Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM." Chemosphere, vol. 313, p. 137636, 2022, doi: 10.1016/j.chemosphere.2022.137636.

[17] F. Illescas-Martinez, L. García, A.-J. García-Sánchez, R. Asorey-Cacheda, and J. Garc a-Haro, "Air quality forecasting in non-monitored urban areas through machine and í deep-learning model," Expert Systems with Applications, 2025, doi: 10.1016/j.eswa.2025.127749.

[18] H. Nguyen, T. H. Le, M. Azzi, and Q. P. Ha, "Monitoring and estimation of urban emissions with low-cost sensor networks and deep learning," Ecological Informatics, 2024, doi: 10.1016/j.ecoinf.2024.102750.

[19] W. Yang, H. Li, J. Wang, and H. Ma, "Spatio-temporal feature interpretable model for air quality forecasting," Ecological Indicators, 2024, doi: 10.1016/j.ecolind.2024.112609.

[20] S. Raja, A. Damodaran, and G. Manohar, "Enhanced forecasting of air quality index through an

integrated deep learning framework," Environmental science and pollution research international, 2025, doi: 10.1007/s11356-025-37269-4.

[21] R. Zhang and N. Awang, "A novel ST-iTransformer model for spatio-temporal ambient air pollution forecasting," Journal of Big Data, 2025, doi: 10.1186/s40537-02501150-5.

[22] V. Sooriyaarachchi, D. J. Lary, L. O. H. Wijeratne, and J. Waczak, "Causalitydriven feature selection for calibrating low-cost airborne particulate sensors using machine learning," Sensors, 2024, doi: 10.3390/s24227304.

[23] S. M. Popescu et al., "Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management," Frontiers in Environmental Science, vol. 12, p. 1336088, 2024, doi: 10.3389/fenvs.2024.1336088.