

Article

Spatial-Aware Uncertainty Quantification for Low-Cost PM_{2.5} Sensors: GeoConformal Prediction vs. Monte Carlo Dropout Under Wildfire Distribution Shift

Anusha Srirenganathan Malarvizhi ¹, Kaylee Smith ², Seren Smith ¹, Joe T. Roberts ³, George Chang ³,
 Mohammad Pourhomayoun ⁴ and Chaowei Yang ^{1,*}

¹ Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA 22030, USA; asrireng@gmu.edu (A.S.M.); jsmit92@gmu.edu (S.S.)

² Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA; kayleesm@umich.edu

³ Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA; joe.t.roberts@jpl.nasa.gov (J.T.R.); george.w.chang@jpl.nasa.gov (G.C.)

⁴ Department of Computer Science, California State University, 5151 State University Drive, Los Angeles, CA 90032, USA; mpourho@calstatela.edu

* Correspondence: cyang3@gmu.edu

Abstract

Low-cost air quality sensors can expand PM_{2.5} monitoring networks but require calibration against regulatory-grade monitors to correct systematic bias towards trustworthy observations. Although interest in predictive uncertainty for air quality estimation has increased, the reliability of uncertainty under spatial sparsity and wildfire-induced distribution shift remains poorly understood. With the increasing use of low-cost sensors during wildfire-smoke episodes, uncertainty quantification during calibration is essential to determine when corrected PM_{2.5} estimates are reliable enough to support near-real-time alert, public-health, and exposure-related decisions. This study presents a spatial uncertainty evaluation framework for trustworthy AI-enabled environmental sensing. The framework is implemented with a transformer-based PM_{2.5} calibration model and two uncertainty quantification methods, GeoConformal Prediction (GCP) and Monte Carlo Dropout (MCD), using sensor pairs in California and the Northeast United States. The calibration model achieved strong performance at short spatial distances, with test R² values of 0.89 and 0.91 at the 1 km threshold in California and the Northeast, respectively, with accuracy declining as sensor separation increased. GCP generally produced calibration curves closer to the ideal diagonal, while MCD generated tighter prediction intervals under normal conditions. During wildfire events, the uncertainty performance depended on sensor separation. At short distances, MCD expanded its uncertainty intervals and captured PM_{2.5} spikes more effectively than GCP (71% vs. 61% coverage). At larger separations, MCD captured only 44% of elevated observations, whereas GCP widened its intervals and achieved 83% coverage. These results demonstrate that uncertainty reliability is strongly influenced by spatial separation and environmental conditions, highlighting the need for uncertainty-aware and spatially adaptive calibration of low-cost PM_{2.5} sensors for trustworthy actionable air quality observation.



Academic Editor: George Efthimiou

Received: 2 June 2026

Revised: 12 July 2026

Accepted: 12 July 2026

Published: 16 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: uncertainty quantification; model-agnostic uncertainty; model-based uncertainty; GeoConformal prediction; Monte Carlo dropout; out-of-distribution detection; dataset shift; wildfire smoke; low-cost sensor; PM_{2.5} calibration

1. Introduction

Air pollution, exacerbated by climate-driven events like wildfires, poses significant risks to both air quality and human health [1]. Fine particulate matter (PM_{2.5}) is of particular concern because it can penetrate deep into the respiratory system and is strongly associated with adverse health outcomes, underscoring the need for accurate monitoring [2]. Regulatory-grade sensors such as those from the United States Environmental Protection Agency's (EPA) AirNow network provide high-quality measurements but are limited in number and spatial coverage. As a result, they often fail to capture localized pollution variability, especially during extreme events [3]. This limitation becomes particularly important during wildfire events, where PM_{2.5} concentrations can vary substantially across space, even within the same region [4]. As a result, there is a growing need for monitoring approaches that can capture localized air quality dynamics beyond the coverage of existing regulatory networks for trustworthy AI for environmental sensing, which refers to AI systems that provide accurate predictions, reliable uncertainty estimates, robustness to spatial and environmental distribution shifts, and confidence measures that support operational decision making.

Low-cost sensors (LCSs) have emerged as a promising supplementary technology for regulatory monitoring networks owing to their affordability, accessibility, and small size, enabling denser spatial coverage [5]. Expanding low-cost sensor networks can help preserve air quality monitoring coverage when individual sensors fail or are destroyed during extreme events [6]. Among the most widely deployed LCS networks, PurpleAir operates tens of thousands of community-based sensors globally, offering near-real-time PM_{2.5} data at high spatiotemporal resolution [7]. While PurpleAir's real-time sensor readings are publicly viewable, bulk and historical data access through their API requires a paid subscription, limiting its utility for large-scale retrospective research [2]. Clarity, another LCS network, provides community-scale PM_{2.5} monitoring with publicly accessible data, making it a viable alternative for research applications [8]. Raw measurements from these LCS require calibration to correct systematic bias relative to regulatory-grade monitors [9]. This calibration is typically performed using statistical or machine learning models that learn relationships between sensor readings, meteorological variables, and reference observations [10]. Once calibrated, these models are used to generate spatial predictions or interpolations of PM_{2.5} concentrations, particularly in areas lacking dense monitoring coverage [11].

During wildfire smoke episodes, LCSs are increasingly used to provide real-time air quality information that helps inform near-real-time alerts, public health advisories, exposure assessments, and emergency response. These episodes produce rapidly changing and unusually high PM_{2.5} concentrations, including short-term concentrations far above typical ambient conditions, creating a more challenging operating environment for LCS [7,12]. Previous studies have shown that the raw LCS PM_{2.5} measurements can overestimate reference monitor concentrations, in some cases by approximately a factor of two, and may exhibit a nonlinear response at extreme smoke concentrations above 300 µg/m³ [7]. Although calibration can reduce this bias and produce corrected PM_{2.5} estimates, it does not quantify the uncertainty in those estimates [5,13]. Uncertainty quantification (UQ) is therefore essential to ensure the reliability of these corrected estimates when environmental conditions deviate from standard calibration ranges. UQ allows the interpretation of calibrated PM_{2.5} in terms of both magnitude and uncertainty, enabling the use of measurements when they are trustworthy for wildfire-related decision-making. Several model-based UQ methods, including Bayesian neural networks (BNNs), Monte Carlo Dropout (MCD), and deep ensembles (DEs), have been developed to quantify uncertainties by modeling variability in model outputs [14]. These UQ methods remain dependent on model assumptions and the

representativeness of the training distribution [15]. Under a wildfire-induced distribution shift, model-derived uncertainty may become miscalibrated, leading to overconfident or unreliable prediction intervals [16]. On the other hand, conformal prediction has emerged as a distribution-free alternative for UQ that constructs prediction intervals from observed residual errors, providing distribution-free coverage guarantees [17,18]. GeoConformal prediction (GCP) extends this framework to spatial settings by weighting calibration errors according to geographic proximity, producing location-dependent prediction intervals with formal coverage guarantees [15].

While interest in predictive uncertainty for air quality estimation has grown, a critical gap remains in the systematic evaluation of the reliability of predictive uncertainty under conditions of spatial sparsity and wildfire-induced distribution shifts. This study addresses this gap by systematically evaluating GCP and MCD reliability for low-cost PM_{2.5} sensor calibration across varying sensor separations and wildfire conditions, establishing a reliability-centered framework for UQ. The research objectives are listed below:

- Develop a transformer-based model for PM_{2.5} calibration and quantify uncertainty using GCP and MCD to produce reliable, well-calibrated estimates with uncertainty bounds across two study areas, California (CA) and the Northeast (NE) United States.
- Analyze how calibration performance and UQ reliability vary with increasing spatial distance between collocated sensor pairs (1 km, 5 km, and 10 km thresholds) across both study areas, to understand the spatial limitations of LCS calibration.
- Evaluate and compare the reliability of GCP and MCD under wildfire and non-wildfire conditions, focusing on empirical coverage and interval width to assess how wildfire-induced distribution shift affects each uncertainty method.

This study offers a practical basis for the trustworthy use of low-cost sensors in operational air-quality monitoring by quantifying calibrated PM_{2.5} concentrations and their associated uncertainty. Reliable uncertainty estimates allow decision-makers to distinguish measurements that can support wildfire-related public health advisories and exposure assessments from those that require caution. The reliability-centered framework developed here can further guide sensor network design and the deployment of low-cost sensors in regions with sparse regulatory coverage.

The remainder of this paper is organized as follows. Section 2 reviews prior research on low-cost PM_{2.5} sensor calibration and relevant uncertainty quantification methods. Section 3 introduces the methods and data used in this research, with a calibration modeling framework, experimental design, and evaluation metrics. Section 4 presents the UQ results and reliability analysis, followed by a discussion in Section 5 and conclusions in Section 6.

2. Literature Review

2.1. Low-Cost Air Quality Sensor Calibration

Recent research on low-cost PM_{2.5} sensor calibration has increasingly emphasized field calibration under real-world conditions, as raw measurements often exhibit systematic bias relative to regulatory-grade sensors. One study demonstrated that accounting for PM_{2.5} levels and temperature can reduce bias and improve agreement with reference monitors [19]. Specifically, LCS are highly sensitive to temperature, humidity, and pollutant composition, and calibration approaches that do not account for this variability may not generalize across diverse settings, highlighting the risk of deployment without local calibration [20,21].

Table 1 summarizes recent LCS PM_{2.5} calibration studies with a wide range of performance depending on model family and the type of uncertainty information provided. Among classical AI/ML approaches, Park, Yoo, Park and Lee [13] achieved an R² of 0.93 with a Hybrid Long Short-Term Memory network (LSTM), and Guruprakash and Chouhan [22] reported an R² of 0.93 with a Graph Neural Network (GNN). Sousan,

et al. [23] obtained an R^2 between 0.91 and 1.00 under chamber-controlled exposure, although such high performance is not necessarily generalizable to field conditions. Hua, et al. [24] reported an R^2 between 0.83 and 0.86 with Generalized Additive Models (GAMs) under different humidity conditions, and Adong, et al. [25] applied Random Forest (RF) across a large-scale deployment. However, these studies primarily emphasized predictive accuracy, with limited or no formal UQ. For uncertainty-aware approaches, Bush, et al. [26] reported an R^2 of 0.91 with expanded uncertainty estimates within $\pm 28\%$, meeting European indicative monitoring thresholds. Zhivkov and Fidanova [27] reported an R^2 of 0.75 and produced distance-scaled 95% prediction intervals using an empirical uncertainty growth rate of 3.84–5.62% per km. Gómez, et al. [28] applied conformal prediction intervals, achieving 94–96% empirical coverage at a single urban site. Together, these studies demonstrate substantial progress in calibration accuracy for low-cost $PM_{2.5}$ sensors. However, most calibration studies report prediction accuracy without formal UQ, leaving prediction reliability uncharacterized. Among the few studies that incorporate UQ, evaluations are limited to single-site deployments and do not assess how interval coverage degrades with spatial transfer or under distribution shifts such as wildfire events.

Table 1. Summary of recent low-cost $PM_{2.5}$ sensor calibration studies, including model type, predictive performance, uncertainty quantification approach, reported uncertainty, and highlights.

Study	Model	Approach Type	R^2	UQ Method	Uncertainty	Highlights
[27]	RF	UQ-aware	0.53–0.75	Distance-based uncertainty	Uncertainty grows 3.84–5.62% per km from the reference	Prediction intervals based on distance-scaled uncertainty
[28]	RF/XGBoost	UQ-aware	>0.70	Conformal prediction	94–96% of observations fall within predicted intervals	Single sensor calibration; limited generalizability
[23]	Linear, RF, GBR	Classical	0.91–1.00	None	–	Chamber-controlled; high R^2 not generalizable to the field
[22]	GNN	Classical	0.93	None	–	Spatial dependencies modeled; no formal UQ
[26]	RF	UQ-aware	0.91	Expanded uncertainty	$\pm 28\%$ around each measurement (within European Union (EU) limits)	Meets EU indicative monitoring thresholds
[25]	RF	Classical	0.92	None	–	Large-scale deployment; no formal UQ
[13]	Hybrid LSTM	Classical	0.93	None	–	Single sensor calibration; limited generalizability; no formal UQ
[24]	GAMs	Classical	0.83–0.86	None	–	Performance varies with humidity; no formal UQ

2.2. Low-Cost Air Quality Sensing During Wildfires

Wildfires are an increasingly common source of extreme $\text{PM}_{2.5}$ exposure, yet regulatory monitoring networks often lack the spatial coverage needed to capture localized smoke impacts [29]. Studies demonstrated that networks of LCS can fill critical spatial gaps, enabling detailed monitoring of rapid $\text{PM}_{2.5}$ increases across indoor, outdoor, and personal exposure settings [3]. When appropriate device-specific adjustment factors are applied, LCSs achieve mean absolute errors below $10 \mu\text{g}/\text{m}^3$ and substantially improve the accuracy of wildfire smoke detection [12,30]. Integrating LCS with regulatory monitoring networks also improves area-wide concentration estimation during wildfire events and reduces interpolation error by over 20% [31]. These findings support the use of LCS for capturing local $\text{PM}_{2.5}$ concentrations during wildfire events, especially in regions lacking coverage from regulatory monitors.

However, wildfire events can introduce a significant dataset shift because $\text{PM}_{2.5}$ concentrations and environmental conditions during smoke episodes may differ substantially from those represented in the model training data [7]. Such shifts can reduce the generalizability of calibration models and lead to degraded performance, as distributional differences between training and deployment data are a recognized challenge in geospatial machine learning applications [32]. Identifying when an observation falls outside the training distribution, therefore, requires formal out-of-distribution detection, which quantifies how far smoke-affected conditions deviate from the training feature space. Furthermore, uncertainty estimates that appear well calibrated under in-distribution conditions may become unreliable under dataset shift [16]. While previous studies have primarily focused on improving calibration accuracy during wildfire events, limited attention has been given to systematically evaluating the reliability of calibrated $\text{PM}_{2.5}$ estimates and their associated uncertainty estimates under these conditions.

2.3. Uncertainty in LCS Calibration

In this section, two forms of spatial shift in $\text{PM}_{2.5}$ calibration are reviewed: variation in $\text{PM}_{2.5}$ characteristics and calibration performance across geographic regions and increasing distance between LCS and reference monitors. Firstly, calibration performance can vary across geographic regions due to differences in aerosol composition, meteorology, emissions, and environmental conditions [33]. These regional differences can reduce model transferability and lead to degraded calibration accuracy when models are applied outside the conditions represented in the training data [34]. To address this, a recent work has advocated for region-specific calibration and validation strategies that account for local aerosol and meteorological conditions [35]. Secondly, fine-scale spatial variability in $\text{PM}_{2.5}$ measurements can occur even between nearby sensors, reflecting local pollution sources and heterogeneity in observed concentrations [36]. This spatial heterogeneity results in reduced agreement between low-cost and regulatory $\text{PM}_{2.5}$ sensors with increasing distance from reference monitors [37,38]. Zhivkov and Fidanova [27] showed that the calibration uncertainty increases with distance from reference monitors, leading to reduced reliability of measurements at greater distances. However, their quantified uncertainty is a distance-dependent empirical growth rate but does not provide individual prediction-level uncertainty or adapt to changing environmental conditions.

Together, these findings highlight that while LCS calibration has advanced considerably through various approaches, most studies have focused on predictive accuracy and have lacked systematic uncertainty quantification. While our previous work systematically compared model-based UQ methods such as BNNs, MCD, and DEs for geospatial air quality monitoring [39], it did not evaluate spatially adaptive, distribution-free UQ methods for calibrated $\text{PM}_{2.5}$ estimates. This study addresses these gaps by systematically

evaluating both model-based and model-agnostic UQ methods for transformer-based PM_{2.5} calibration. A major contribution of this study is the evaluation of GCP as a spatially adaptive, distribution-free UQ framework that generates geographically informed prediction intervals with formal coverage guarantees. In addition, this study systematically examines how uncertainty reliability varies across increasing collocation distances between low-cost and reference sensors and under wildfire-induced distribution shifts.

3. Methodology

Figure 1 illustrates the overall research workflow as a spatial uncertainty evaluation framework for AI-enabled environmental sensing, beginning with data acquisition from OpenAQ LCS and EPA regulatory monitors, followed by data collocation, preprocessing, and train-test splitting. The processed data are used to train a Transformer-based calibration model, which produces calibrated PM_{2.5} estimates. These calibrated PM_{2.5} estimates are then used to quantify uncertainty and assess reliability under wildfire events.

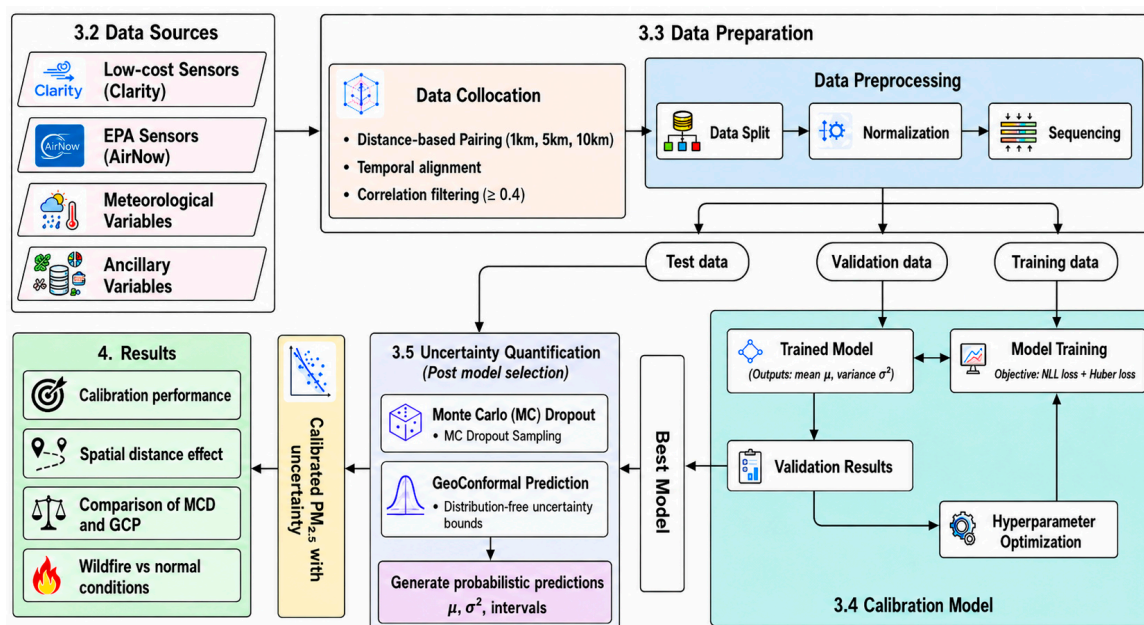


Figure 1. Workflow of PM_{2.5} sensor calibration and spatial uncertainty evaluation framework for AI-enabled environmental sensing, including data acquisition, collocation, preprocessing, Transformer-based calibration, uncertainty estimation, and reliability assessment.

3.1. Study Area

This study focuses on two geographically distinct study areas, California and the Northeast United States. These two regions were selected because they represent contrasting aerosol compositions, regulatory monitoring densities, and wildfire exposure regimes, enabling evaluation of model generalizability across climatically and geographically diverse conditions. Figure 2a,b depict the locations of AirNow and Clarity sensors across both study areas. In California, 147 AirNow monitors and 1022 Clarity sensors were used, concentrated in the Los Angeles and San Francisco metropolitan areas. In the Northeast United States, 157 AirNow monitors and 101 Clarity sensors were used. AirNow monitors provide broad spatial coverage across both study areas, while Clarity sensors offer dense coverage over urban regions, which is critical for capturing localized PM_{2.5} variability and supporting high-resolution calibration during extreme events.

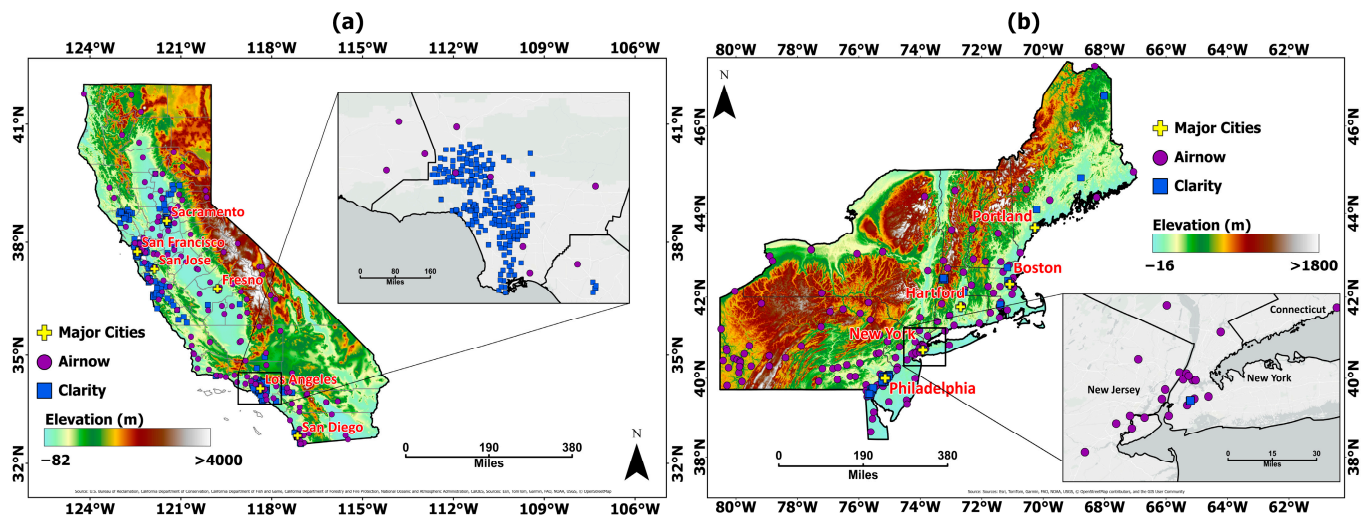


Figure 2. Spatial distribution of AirNow (red circles), the U.S. EPA reference-grade monitors and Clarity (blue squares) $PM_{2.5}$ monitoring sites in (a) California and (b) the Northeast, with elevation (m) shown as background shading.

Two wildfire events were used as case studies to evaluate model performance under extreme $PM_{2.5}$ conditions. In California, the Palisades Fire and Eaton Fire ignited on 7 January 2025, becoming among the most destructive wildfires in California history [40]. The Palisades Fire burned approximately 23,448 acres in the Pacific Palisades and Malibu areas, while the Eaton Fire burned approximately 14,021 acres near Altadena and Pasadena, collectively causing severe $PM_{2.5}$ degradation across the Los Angeles basin [41]. In the Northeast United States, the Jennings Creek Wildfire ignited in November 2024 along the border of New York and New Jersey, burning approximately 5000 acres, prompting health advisories for unhealthy air quality across the region [42]. Together, these wildfire events provide distinct and complementary case studies for evaluating the reliability of UQ methods under wildfire-induced distribution shift across both study areas.

3.2. Data Sources

3.2.1. AirNow

Hourly ground-level $PM_{2.5}$ concentrations from 2022 to 2025 were obtained from the U.S. EPA AirNow network via the AirNow API [43]. AirNow is a national air quality reporting partnership that aggregates monitoring data from federal, state, tribal, and local agencies across the United States. Monitoring agencies submit observations from regulatory-grade monitors designated as Federal Reference Method (FRM) or Federal Equivalent Method (FEM) instruments to AirNow's centralized system, which provides standardized national reporting and near-real-time quality control. EPA guidance defines acceptable $PM_{2.5}$ measurement uncertainty goals as 10% coefficient of variation for total precision and $\pm 10\%$ for total bias. These observations served as reference measurements for sensor calibration and uncertainty evaluation.

3.2.2. Clarity

The Clarity Node-S is a low-cost optical particle counter that measures $PM_{2.5}$ via laser light scattering, with raw measurements recorded at minute-level temporal resolution [21]. For this study, hourly aggregated $PM_{2.5}$ data were downloaded from the OpenAQ measurements API [44] for the period 2022–2025. These data were collocated with AirNow regulatory-grade monitors and used as the primary LCS input to develop and train calibration models.

3.2.3. Meteorological Variables

Hourly meteorological variables, including 2 m temperature, 2 m relative humidity, 10 m wind speed, and pressure, were obtained from the National Oceanic and Atmospheric Administration (NOAA) High-Resolution Rapid Refresh (HRRR) model, a real-time 3 km resolution hourly updated atmospheric model covering the contiguous United States. HRRR data were downloaded from the NOAA archive hosted on Amazon Web Services (AWS) Simple Storage Service (S3) [45] using the 2D surface-level fields at a 3 km spatial resolution. Although some Clarity sensor locations report temperature and relative humidity through the OpenAQ API, an assessment of the data used in this study showed that more than 97% of these local meteorological observations were missing. Therefore, HRRR was used to provide spatially and temporally complete and consistent meteorological inputs across all collocated sensor-reference pairs. Relative humidity and temperature are the primary meteorological drivers of the bias in the LCS, as high humidity causes hygroscopic particle growth, which inflates optical scattering signals, while temperature affects sensor electronics and particle composition, both of which systematically distort PM_{2.5} readings [46,47]. These variables were extracted at the collocated sensor locations and temporally matched to hourly PM_{2.5} observations.

3.2.4. Ancillary Variables

Land-use and land-cover (LULC) data were obtained from the National Land Cover Database (NLCD), produced by the United States Geological Survey (USGS) at 30 m spatial resolution [48]. Federal Information Processing Standards (FIPS) codes were used to identify county-level administrative boundaries for each sensor location [49]. Geographic coordinates, including latitude and longitude, were included to enable spatial matching between collocated sensor pairs. Temporal categorical variables such as hour of day, day, month, season, day of year (DoY), day of week (DoW), and week of year (WoY) were derived from each observation's timestamp to capture diurnal, weekly, and seasonal cycles in PM_{2.5} concentrations.

3.3. Data Preparation

Raw PM_{2.5} measurements from Clarity sensors were first subjected to outlier removal using a rolling standard deviation filter. Records with a rolling standard deviation of zero were excluded to remove inactive or non-reporting sensor readings. Following outlier removal, Clarity sensors were spatiotemporally collocated with the nearest AirNow regulatory-grade monitors at distance thresholds of 1 km, 5 km, and 10 km, retaining only sensor pairs with a Pearson correlation coefficient of 0.4 or higher to ensure a minimum level of measurement agreement before calibration.

Table 2 summarizes the collocated sensor pair statistics across both study areas. In the Northeast, the number of collocated pairs increased from 4 at 1 km to 19 at 10 km, with mean correlations ranging from 0.87 at 1 km to 0.81 at 10 km. In California, pairs increased from 11 at 1 km to 27 at 10 km, with mean correlations ranging from 0.80 at 1 km to 0.77 at 10 km. Across both study areas, the correlation declined as sensor separation distance increased. In contrast, the number of available pairs increased, reflecting a trade-off between spatial proximity and the availability of collocated sensor pairs for model training. Although the number of collocated pairs at the 1 km threshold is limited, each pair contributes a continuous multi-year hourly time series, yielding ~75,000 training samples in California and ~12,000 in the Northeast.

Table 2. Summary of spatiotemporal collocation statistics for Clarity and AirNow sensor pairs across three distance thresholds (1 km, 5 km, and 10 km) in California and the Northeast United States.

Metric	NE 1 km	NE 5 km	NE 10 km	CA 1 km	CA 5 km	CA 10 km
Total pairs	4	11	19	11	17	27
Mean distance (m)	48.8	1476.1	3843.7	117.9	1132.4	3604.5
Max distance (m)	83.7	4104.2	9717.7	464.4	4164.0	9892.9
Mean correlation	0.87	0.85	0.81	0.80	0.78	0.77
Min correlation	0.75	0.67	0.65	0.52	0.52	0.52
Max correlation	0.96	0.96	0.96	0.95	0.95	0.95
Pairs with $r > 0.7$	4	10	16	9	13	20
Pairs with $r > 0.9$	2	4	4	2	2	2

This study applied several preprocessing steps to the collocated dataset before training, including gap-filling for missing values, data splitting, feature scaling, and sequence construction. Missing PM_{2.5} values were imputed using an XGBoost-based model trained separately for each region, achieving an R² of 0.88 and an RMSE of 2.81 µg/m³ for California, and an R² of 0.89 and an RMSE of 1.43 µg/m³ for the Northeast [50]. The imputation model was trained on spatial, temporal, and auxiliary predictors, capturing nonlinear PM_{2.5} relationships during rapidly changing pollution conditions. Because spatial and temporal dependencies are represented only through the predictors, the imputation reflects the structure present in the observed data [51]. FIPS codes and LULC were incorporated as categorical variables to capture county-level administrative context and surface environment characteristics at each sensor location. The collocated dataset was split using two complementary validation schemes. First, a time-based split was applied, with 70% of the data used for training, 10% for validation, and 20% for testing, ensuring that all sets are temporally non-overlapping, preventing data leakage, and reflecting real-world deployment conditions. Second, a spatial hold-out validation was performed to assess geographic transferability, in which the data were partitioned by the U.S. climate divisions designated by the National Oceanic and Atmospheric Administration (NOAA) [52], holding out entire divisions for testing while training on the remaining divisions. Then the input features were subsequently normalized using a standard scaler to ensure consistent scaling across continuous variables. Finally, the data were organized into sequential input windows of 23 preceding hourly observations, with each sequence used to calibrate the current-hour PM_{2.5} measurement. The sequence length of 23 was selected to capture near-diurnal temporal dependencies while maintaining computational efficiency [53].

3.4. Calibration Model

The proposed calibration model is based on a transformer architecture, which was introduced by [54]. A schematic overview of the model architecture is presented in Figure 3. Transformers capture long-range temporal dependencies through self-attention mechanisms while processing entire input sequences in parallel. This makes them well-suited for PM_{2.5} calibration, where capturing temporal patterns across the input window is essential [55].

The proposed model takes continuous and categorical spatiotemporal variables as input and produces a multi-step sequence of calibrated PM_{2.5} estimates. Categorical temporal variables, including hour, day, season, month, DoY, DoW, and WoY, as well as location indicators, are represented using learnable embedding layers that map each category to a continuous vector representation. These embeddings allow the model to capture relation-

ships among discrete temporal and spatial features. The resulting categorical embeddings are concatenated with continuous input features, including temperature, pressure, and humidity, to form the complete input sequence. This combined representation allows the model to jointly learn from both discrete and continuous spatiotemporal information.

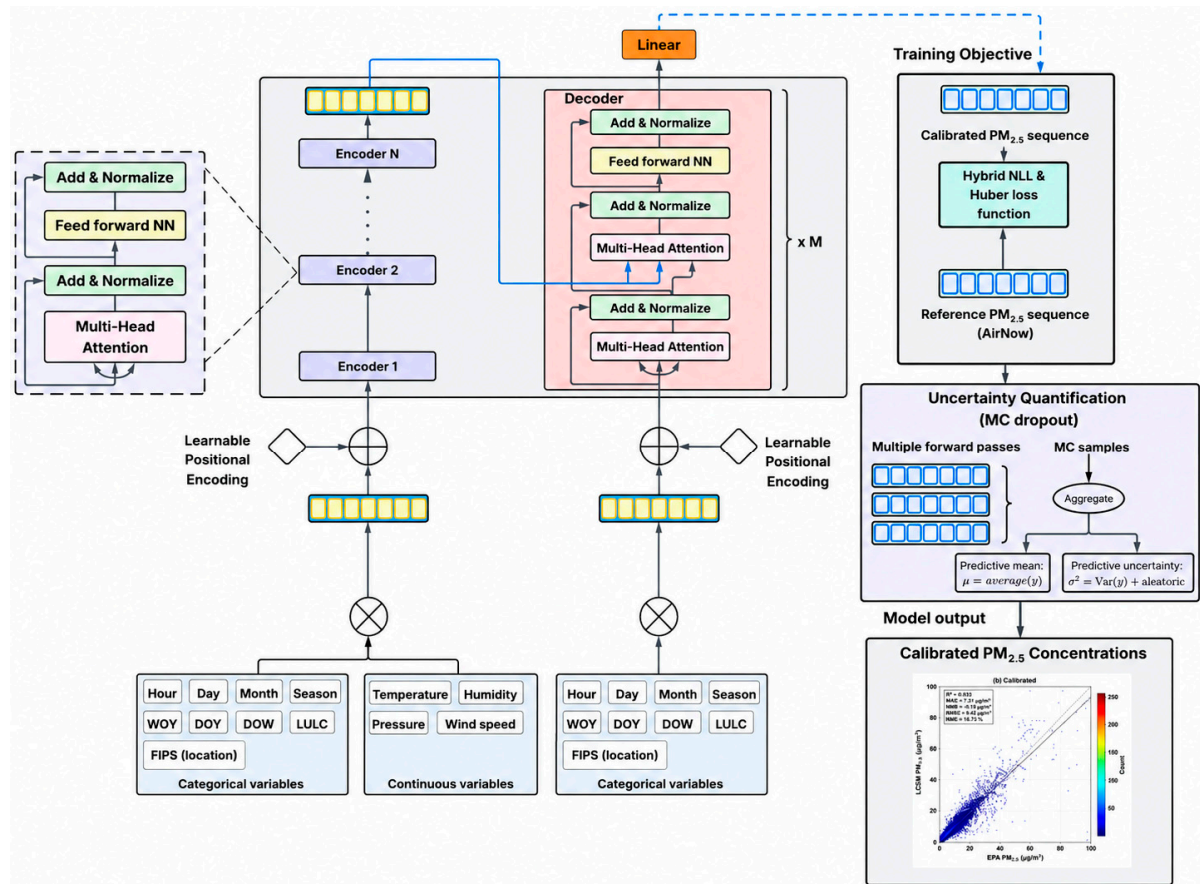


Figure 3. Transformer-based calibration architecture for PM_{2.5} estimation with encoder–decoder attention mechanisms, and Monte Carlo dropout for UQ. The inset (left) details a single encoder block, vertical dots denote the stacked intermediate encoder layers between Encoder 2 and Encoder N, where N and M are the number of stacked encoder and decoder layers, respectively.

Learnable positional embeddings are added to the projected sequence representation to preserve temporal relationships within the input data. Since the Transformer architecture does not inherently encode sequence order, positional information must be explicitly incorporated into the input representations [54]. These embeddings provide information about the position of each element in the sequence, enabling the model to distinguish between different temporal arrangements. This allows the self-attention mechanism to account for sequence order, enabling the model to capture temporal structure during both training and inference.

The encoder consists of a stack of identical encoder layers. Each encoder layer includes multi-head self-attention followed by a feed-forward network. Residual connections and layer normalization are applied around both sublayers [54].

In the self-attention sub-layer, the input representation is linearly projected into query (Q), key (K), and value (V) matrices. The relationships between different positions in the sequence are computed using scaled dot-product attention, defined as Equation (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V denote the query, key, and value matrices, and d_k represents the dimensionality of the key vectors. The resulting attention outputs from multiple heads are concatenated and passed to the feed-forward network within the encoder layer. The encoder produces a transformed sequence representation that serves as memory for the decoder.

The decoder also consists of stacked decoder layers. Each decoder layer includes self-attention, encoder–decoder attention, and a feed-forward network, with residual connections and layer normalization applied after each component. The decoder processes the encoded sequence representation to generate the final calibrated $\text{PM}_{2.5}$ estimates.

The decoder output is passed through a hidden linear projection layer to produce the predicted mean (μ) and variance (σ^2) of $\text{PM}_{2.5}$ concentrations. Model parameters are optimized using a hybrid loss function that combines Gaussian negative log-likelihood (NLL) and a weighted Huber loss [56]. The NLL term enables the model to capture aleatoric uncertainty by penalizing errors relative to the predicted variance. At the same time, the weighted Huber loss improves robustness to extreme $\text{PM}_{2.5}$ values by reducing the disproportionate influence of large errors. The combined loss function, defined in Equation (2), balances uncertainty calibration with robustness to high $\text{PM}_{2.5}$ outliers.

$$L = \alpha \cdot L_{\text{NLL}} + (1 - \alpha) \cdot L_{\text{Huber}} \quad (2)$$

The Gaussian Negative Log-Likelihood loss is defined as Equation (3):

$$L_{\text{NLL}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{2} \log \sigma_i^2 + \frac{(y_i - \mu_i)^2}{2\sigma_i^2} \right] \quad (3)$$

where μ_i and σ_i^2 are the predicted mean and variance for the sample i , and $\alpha = 0.5$ balances the two loss components.

Hyperparameter optimization was performed using random search over a predefined configuration space [57]. The search space included embedding dimensions, the number of attention heads, the number of layers, the feed-forward dimension, the dropout rate, the learning rate, and the batch size. The dropout rate was searched over the range of 0.1 to 0.5 and applied during both training and inference for MCD-based uncertainty estimation, with values between 0.1 and 0.2 yielding stable validation performance [39]. Each configuration was trained with early stopping based on validation R^2 , and the best-performing configuration was selected for final evaluation. The number of attention heads (8), number of layers (5), feed-forward dimension (1024), and optimizer (AdamW) remained consistent across all configurations, while the embedding dimension, batch size, dropout rate, and number of epochs varied across regions and distance thresholds, with the final configurations reported in Table 3.

Table 3. Optimized hyperparameters selected via random search for each region and spatial distance threshold.

Parameter	CA 1 km	NE 1 km	CA 5 km	NE 5 km	CA 10 km	NE 10 km
Epochs	20	35	20	35	20	32
Batch Size	64	64	128	128	128	128
Embedding Dimension	256	256	512	512	512	512
Dropout	0.1	0.2	0.1	0.2	0.1	0.2
Learning Rate	1×10^{-4}	1×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}	2×10^{-4}

3.5. Uncertainty Quantification

In this study, two complementary UQ methods are considered: MCD, a model-based method grounded in approximate Bayesian inference, and GCP, a distribution-free framework that provides statistically valid prediction intervals.

MCD, introduced by Gal and Ghahramani [58] provides an efficient framework for estimating predictive uncertainty in deep neural networks by interpreting dropout as a Bayesian approximation to a probabilistic deep Gaussian process. Under this framework, multiple stochastic forward passes are performed during inference with dropout active, producing a distribution of predictions from which the predictive mean and variance are derived. In this study, MCD is extended to jointly capture both epistemic and aleatoric uncertainties by predicting the mean and variance of the target distribution, with the training objective modified to a hybrid NLL loss to enable learning of input-dependent uncertainty. Equation (4) defines the total predictive uncertainty as the sum of epistemic uncertainty and aleatoric uncertainty:

$$\sigma_{\text{total}}^2 = \underbrace{\text{Var}_t(\mu_t)}_{\sigma_{\text{epistemic}}^2} + \underbrace{\frac{1}{T} \sum_{t=1}^T \sigma_t^2}_{\sigma_{\text{aleatoric}}^2} \quad (4)$$

where μ_t and σ_t^2 denote the predicted mean and variance obtained from the t -th stochastic forward pass with dropout enabled, $\text{Var}_t(\mu_t)$ quantifies epistemic uncertainty across all stochastic predictions, and T is the total number of stochastic forward passes, set to 30 in this study. This value is consistent with the original MCD formulation, in which Gal and Ghahramani [58] showed that the reasonable estimates of the predictive mean and uncertainty can be obtained with as few as 10 stochastic forward passes.

GCP is a model-agnostic UQ framework that extends conformal prediction to spatial prediction tasks [15]. Conformal prediction constructs prediction intervals from a calibration dataset by computing nonconformity scores, typically defined as residuals, and generates intervals with valid marginal coverage under minimal distributional assumptions [17]. The GCP incorporates geographic weighting to account for spatial heterogeneity and covariate shift through a Gaussian distance-decay kernel. It assigns greater weight to nearby calibration points and lower weight to distant ones, where the weight assigned to each calibration point is defined in Equation (5):

$$w(X_i) = K \left(\frac{|l_i - l_{n+1}|}{b} \right) \quad (5)$$

where l_i is the location of the calibration point i , l_{n+1} is the test location, and b is the bandwidth parameter controlling the rate of distance decay. Smaller b values assign a stronger weight to nearby calibration residuals, whereas larger b values produce smoother and more uniform geographic weighting; in this study, $b = 0.15$ was selected through validation-set tuning. The neighborhood included all calibration points without a fixed spatial threshold, and the bandwidth parameter was held constant across all test locations, yielding locally adaptive prediction intervals that reflect spatial heterogeneity.

Out-of-Distribution (OOD) Detection. Mahalanobis distance was used to identify OOD observations in the test set. For a feature vector x , the squared Mahalanobis distance is defined in Equation (6):

$$d^2(x) = (x - \mu)^\top \Sigma^{-1} (x - \mu) \quad (6)$$

where x is the feature vector and μ and Σ are the mean vector and covariance matrix of the feature representations estimated from the training set. An observation was flagged

as OOD when its squared Mahalanobis distance exceeded the 97.5th percentile of the training-set distance distribution.

Spatial Autocorrelation Analysis. Spatial autocorrelation in the calibration residuals was evaluated using Moran's I to determine whether residual errors were geographically clustered after model calibration. Positive Moran's I values indicate clustering of similar residuals, negative values indicate spatial dispersion, and values near zero indicate a random spatial pattern.

3.6. Evaluation Metrics

To comprehensively evaluate both calibration performance and uncertainty reliability, this study employs a combination of standard regression metrics and uncertainty-specific metrics summarized in Table 4. While traditional metrics assess how well the calibration model matches observed $\text{PM}_{2.5}$ concentrations, UQ metrics evaluate the reliability of the uncertainty intervals. Standard regression metrics, including R^2 , MAE, MSE, and RMSE, were used to evaluate calibration accuracy, as defined in [39].

Table 4. Summary of evaluation metrics used to assess predictive accuracy and uncertainty reliability.

Metric	Type	Description
R^2	Accuracy	Proportion of variance in observed $\text{PM}_{2.5}$ explained by the calibrated estimates; values closer to 1 indicate better model fit.
MAE	Accuracy	Mean absolute difference between calibrated and observed $\text{PM}_{2.5}$ values; lower values indicate smaller average errors.
MSE	Accuracy	Mean squared deviation between calibrated and observed $\text{PM}_{2.5}$ values; lower values indicate lower squared error.
RMSE	Accuracy	Calibration error expressed in the original $\text{PM}_{2.5}$ units; lower values indicate better predictive accuracy.
ECE	Uncertainty	Difference between predicted confidence and observed coverage; lower values indicate better uncertainty calibration.
PICP	Uncertainty	Proportion of observations falling within the prediction interval; higher values indicate better interval coverage.
PIW	Uncertainty	Mean width of the prediction intervals; lower values indicate narrower and more efficient intervals.

Expected Calibration Error (ECE) was used to quantify the difference between predicted confidence and observed coverage, as defined in [39]. Additionally, Prediction Interval Coverage Probability (PICP) and Prediction Interval Width (PIW) were used to evaluate the uncertainty intervals, as defined in Equation (7) and Equation (8), respectively:

$$\text{PIW} = \frac{1}{n} \sum_{i=1}^n (u_i - l_i) \quad (7)$$

where l_i and u_i denote the lower and upper bounds of the prediction interval for observation i .

$$\text{PICP} = \frac{1}{n} \sum_{i=1}^n 1(l_i < y_i < u_i) \quad (8)$$

where the indicator function equals 1 if the observed value lies within the interval and 0 otherwise.

4. Results

4.1. Statistical Analysis

Exploratory statistical analysis was conducted on PM_{2.5} concentrations from both AirNow and Clarity sensors across two study areas. Boxplots of PM_{2.5} concentrations were used to summarize the central tendency and variability of the data from both sensor types across three distance thresholds (1 km, 5 km, and 10 km). In California (Figure S1), the two datasets show similar central tendencies, with median values of 5.1 µg/m³ for AirNow and 4.77 µg/m³ for Clarity. The interquartile ranges are comparable, spanning 3.1–8.2 µg/m³ for AirNow and 2.89–8.07 µg/m³ for Clarity. Despite these similarities, AirNow shows greater variability, with a standard deviation of 6.05 µg/m³ compared to 4.78 µg/m³ for Clarity. AirNow reaches a maximum of 401.1 µg/m³, while Clarity values reach a maximum of 320 µg/m³. The timing of these extreme observations aligns with the 2025 Palisades and Eaton Fires, indicating that these outliers correspond to real pollution events rather than measurement noise.

In the Northeast United States (Figure S2), both sensor types again show comparable central tendencies, though at lower concentrations than in California. AirNow reaches a maximum of 150 µg/m³, while Clarity reaches a maximum of 102 µg/m³, with extreme observations corresponding to smoke transport from Canadian wildfires [59].

To further examine the distributional characteristics, histograms of PM_{2.5} concentrations were analyzed for both AirNow and Clarity sensors across the two study areas, as shown in Figures S3 and S4. In California (Figure S3), both AirNow and Clarity exhibit a strongly right-skewed distribution, with the majority of observations concentrated between approximately 3 and 8 µg/m³. Both sensors report similar typical PM_{2.5} values, with long upper tails corresponding to infrequent but extreme pollution events associated with the 2025 Palisades and Eaton Fires. In the Northeast United States (Figure S4), a similar right-skewed distribution is observed. However, extreme values are notably lower than in California, with upper-tail observations corresponding to smoke transport from Canadian wildfires.

4.2. Model Performance

4.2.1. California

Table 5 presents the model performance metrics across three spatial distance thresholds for California. The model demonstrates strong calibration performance at shorter distances, with test R² values of 0.89, 0.88, and 0.84 at the 1 km, 5 km, and 10 km thresholds, respectively, indicating a gradual decline in accuracy with increasing spatial separation. This decline suggests that increasing spatial separation introduces greater variability and reduces the model's ability to capture localized PM_{2.5} patterns. The gap between training and test performance remains relatively small at shorter distances, indicating limited overfitting when sensors are closely collocated. At larger distances, the increase in error metrics and bias (to 0.2606 µg/m³ at 10 km) suggests reduced generalization and greater difficulty in modeling spatial heterogeneity. These results confirm that calibration accuracy is sensitive to spatial distance, with closer sensor pairs yielding more reliable and consistent predictions.

Table 6 presents the California spatial hold-out results across the three distance thresholds, where the reported metrics are averaged across the held-out climate divisions. The mean test R² values of 0.82, 0.81, and 0.83 at the 1 km, 5 km, and 10 km thresholds are slightly lower than those obtained under the temporal split (0.89, 0.88, and 0.84). The wider gap between training and test performance under spatial hold-out reflects the greater difficulty of extrapolating to spatially disjoint locations than to unseen time periods within observed locations. Despite this gap, the mean test R² remains near 0.82 across all thresh-

olds, indicating that the model captures transferable spatial relationships rather than relying on dependencies among nearby stations.

Table 5. Model performance metrics across three spatial distance thresholds (1 km, 5 km, and 10 km) for train, validation, and test splits under temporal splitting in California.

Distance Threshold	Split	R ²	RMSE (µg/m ³)	MSE (µg/m ³)	MAE (µg/m ³)	Bias (µg/m ³)
1 km	Train	0.91	2.33	5.45	1.20	−0.0013
	Validation	0.89	2.68	7.18	1.28	0.0927
	Test	0.89	2.81	8.06	1.44	0.1075
5 km	Train	0.90	2.64	6.97	1.26	−0.0140
	Validation	0.88	2.88	8.29	1.29	0.0809
	Test	0.88	3.27	10.72	1.55	−0.2474
10 km	Train	0.85	3.08	9.49	1.25	−0.0196
	Validation	0.83	3.44	11.83	1.38	0.1243
	Test	0.84	4.17	17.42	1.74	0.2606

Table 6. Model performance metrics across three spatial distance thresholds (1 km, 5 km, and 10 km) for train, validation, and test splits under spatial hold-out validation in California.

Distance Threshold	Split	R ²	RMSE (µg/m ³)	MSE (µg/m ³)	MAE (µg/m ³)	Bias (µg/m ³)
1 km	Train	0.87	2.68	7.18	1.30	−0.014
	Validation	0.87	2.62	6.86	1.45	−0.140
	Test	0.82	3.64	13.25	1.50	−0.271
5 km	Train	0.88	2.62	6.86	1.29	−0.013
	Validation	0.88	2.89	8.35	1.58	−0.097
	Test	0.81	3.93	15.44	1.69	−0.160
10 km	Train	0.91	2.22	4.93	1.20	−0.005
	Validation	0.90	3.12	9.73	1.56	0.098
	Test	0.83	3.28	10.76	1.50	−0.163

Figure 4 compares uncalibrated and calibrated PM_{2.5} predictions against reference AirNow observations. The uncalibrated model (Figure 4a) shows substantial dispersion relative to the 1:1 line, with an R² of 0.601, an RMSE of 5.42 µg/m³, and a positive bias of 0.55 µg/m³, indicating systematic overestimation. In contrast, the calibrated model (Figure 4b) exhibits a much tighter alignment with the 1:1 line, with R² improving to 0.893 and RMSE reducing to 2.81 µg/m³. Notably, calibration also substantially reduces the bias from 0.55 to 0.12 µg/m³, demonstrating the model's effectiveness in correcting systematic overestimation of PM_{2.5} concentrations.

4.2.2. Northeast

Table 7 presents the model performance metrics for the Northeast United States across three spatial distance thresholds. Model performance remains strong across all distance thresholds, with only modest increases in error as spatial separation increases. At the 1 km threshold, the model achieves a test R² of 0.91 and RMSE of 1.38 µg/m³, indicating accurate calibration under closely collocated sensor conditions. Compared to California, error metrics in the Northeast are substantially lower across all thresholds, suggesting more consistent and homogeneous PM_{2.5} conditions in this region [60]. As spatial distance increases, performance declines modestly, with test R² decreasing from 0.91 at 1 km to 0.88 at 10 km and RMSE increasing from 1.38 to 1.63 µg/m³. The gap between training and test performance remains small across all distance thresholds, suggesting minimal overfitting and strong generalization. Bias values are consistently negative across

all thresholds (ranging from -0.1245 to $-0.3200 \mu\text{g}/\text{m}^3$), indicating a systematic tendency for the model to slightly underestimate $\text{PM}_{2.5}$ concentrations, though the magnitude remains small.

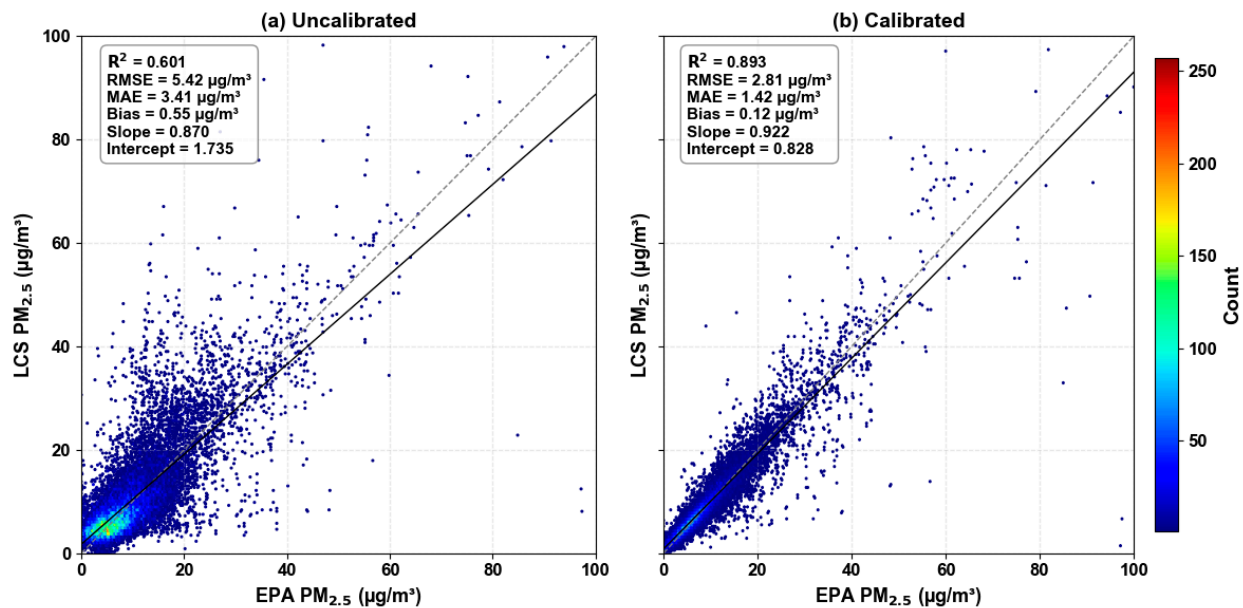


Figure 4. Scatter plots of LCS $\text{PM}_{2.5}$ vs. EPA $\text{PM}_{2.5}$ for (a) uncalibrated and (b) calibrated Clarity sensors at the 1 km distance threshold in California.

Table 7. Model performance metrics across three spatial distance thresholds (1 km, 5 km, and 10 km) for train, validation, and test splits under temporal splitting in the Northeast.

Distance Threshold	Split	R^2	RMSE ($\mu\text{g}/\text{m}^3$)	MSE ($\mu\text{g}/\text{m}^3$)	MAE ($\mu\text{g}/\text{m}^3$)	Bias ($\mu\text{g}/\text{m}^3$)
1 km	Train	0.95	1.03	1.06	0.60	-0.0047
	Validation	0.92	1.02	1.04	0.54	-0.0158
	Test	0.91	1.38	1.90	0.71	-0.1245
5 km	Train	0.94	1.19	1.41	0.64	-0.0018
	Validation	0.92	1.09	1.19	0.56	-0.0465
	Test	0.88	1.63	2.66	0.76	-0.3200
10 km	Train	0.92	1.31	1.72	0.58	-0.0104
	Validation	0.90	1.19	1.42	0.63	-0.0636
	Test	0.82	2.1	4.39	0.72	-0.3073

Table 8 presents the Northeast spatial hold-out results, where metrics are averaged across the held-out climate divisions. Test R^2 remains high and stable across thresholds, with values of 0.93, 0.90, and 0.90 at the 1 km, 5 km, and 10 km thresholds, respectively, and only a small gap between training and held-out performance. This consistency suggests that the calibration relationship is more spatially transferable across the Northeast divisions, allowing the model to generalize to unseen locations rather than relying only on dependencies among nearby stations. This behavior is consistent with the documented east–west gradient in $\text{PM}_{2.5}$ predictability, where models tend to perform more uniformly across the relatively flat eastern United States than in the more terrain- and composition-heterogeneous western regions [60].

Figure 5 compares uncalibrated and calibrated predictions against reference observations. The uncalibrated model (Figure 5a) shows substantial dispersion from the 1:1 line, with relatively low explanatory power ($R^2 = 0.402$) and higher error (RMSE = $3.60 \mu\text{g}/\text{m}^3$). In contrast, the calibrated model (Figure 5b) demonstrates a significant improvement, with

R^2 increasing to 0.913 and RMSE decreasing to $1.37 \mu\text{g}/\text{m}^3$. Additionally, the reduction in bias from $-0.98 \mu\text{g}/\text{m}^3$ to $-0.12 \mu\text{g}/\text{m}^3$ highlights the calibration model's effectiveness in correcting systematic underestimation. Overall, these results demonstrate that calibration substantially improves both accuracy and reliability in the Northeast region, with stronger performance compared to California, particularly at larger spatial distances. This is likely because the Northeast has lower $\text{PM}_{2.5}$ concentrations with less spatial variability, which reduces prediction difficulty [60], whereas California's complex terrain and episodic wildfire smoke introduce greater concentration heterogeneity, which challenges model generalization [61]. This is likely because the Northeast has lower $\text{PM}_{2.5}$ concentrations with less terrain variability, which reduces prediction difficulty [60], whereas California's complex terrain and episodic wildfire smoke introduce greater concentration heterogeneity that challenges model generalization [61].

Table 8. Model performance metrics across three spatial distance thresholds (1 km, 5 km, and 10 km) for train, validation, and test splits under spatial hold-out validation in the Northeast.

Distance Threshold	Split	R^2	RMSE ($\mu\text{g}/\text{m}^3$)	MSE ($\mu\text{g}/\text{m}^3$)	MAE ($\mu\text{g}/\text{m}^3$)	Bias ($\mu\text{g}/\text{m}^3$)
1 km	Train	0.95	0.96	0.92	0.53	−0.0016
	Validation	0.94	1.21	1.47	0.63	−0.0184
	Test	0.93	1.28	1.64	0.70	−0.1655
5 km	Train	0.93	1.28	1.65	0.61	−0.0092
	Validation	0.91	1.34	1.79	0.65	−0.0318
	Test	0.90	1.40	1.96	0.69	−0.0680
10 km	Train	0.92	1.33	1.77	0.70	−0.0049
	Validation	0.90	1.46	2.14	0.72	−0.0124
	Test	0.90	1.49	2.22	0.76	−0.0527

4.3. Comparison of MCD and GCP

To further quantify the reliability and practical usefulness of the uncertainty estimates, calibration metrics including PICP, PIW, and ECE were evaluated across multiple confidence levels.

4.3.1. California

The calibration curves for MCD and GCP at the 1 km distance threshold in California are presented in Figure 6. The MCD calibration curve lies predominantly above the ideal diagonal, indicating slight overcoverage at lower confidence levels where the observed PICP values exceed nominal targets. In contrast, the GCP calibration curve consistently falls below the diagonal across the full confidence range, reflecting systematic undercoverage. Notably, while MCD achieves a lower ECE (0.02 vs. 0.03), its miscalibration area (0.03) is higher than that of GCP (0.02), suggesting that GCP's curve adheres more closely to the ideal diagonal across the full confidence range despite its slightly higher point-level deviation.

Table S1 presents calibration and uncertainty-interval metrics, such as PICP, PIW, and ECE, for MCD and GCP approaches across confidence levels at the 1 km distance threshold in California. The MCD approach demonstrates strong calibration performance, with observed PICP values closely matching the nominal confidence levels. At lower confidence levels, the model exhibits slight underconfidence, with observed PICP values exceeding the nominal levels (e.g., at the 50% confidence level, the observed PICP is 0.55 compared to the ideal of 0.50). At higher confidence levels, however, the trend reverses, with mild undercoverage observed (e.g., at the 99% confidence level, the observed PICP is 0.97 compared to the ideal of 0.99). The relatively low ECE value of 0.02 further

confirms that the model is well-calibrated. The prediction intervals widen as expected with increasing confidence levels, with PIW values ranging from $2.37 \mu\text{g}/\text{m}^3$ at the 50% level to $9.12 \mu\text{g}/\text{m}^3$ at the 99% level, reflecting the model's ability to scale uncertainty with confidence appropriately.

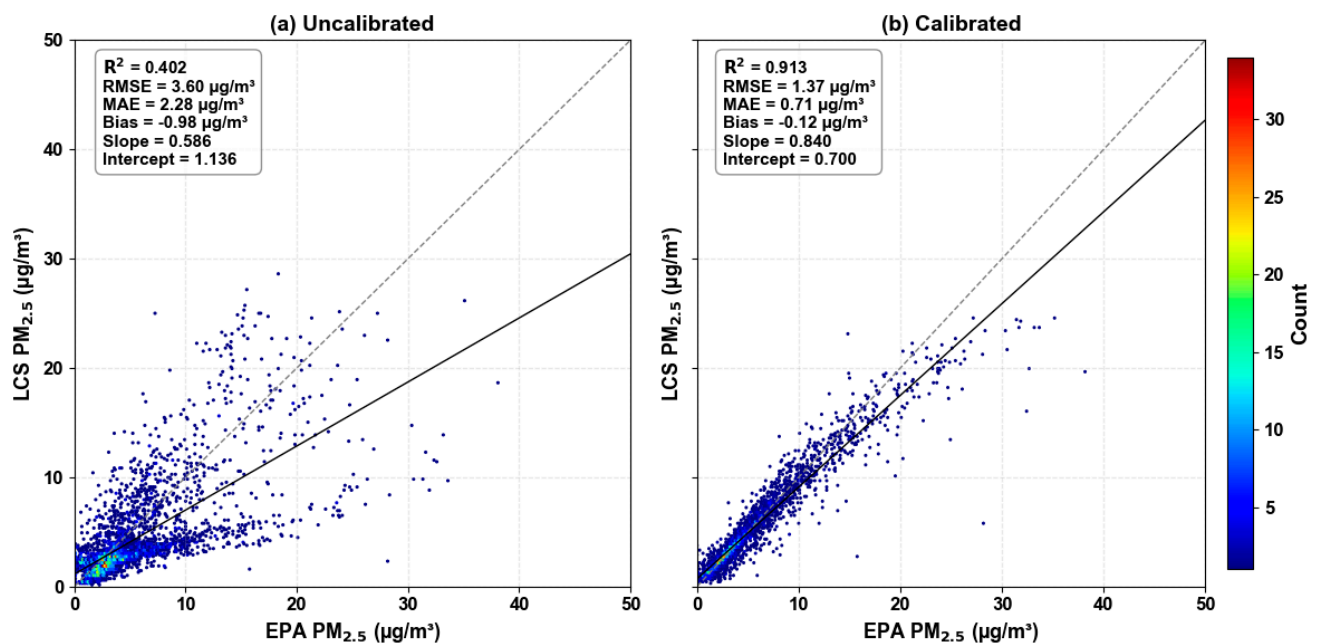


Figure 5. Scatter plots of LCS PM_{2.5} vs. EPA PM_{2.5} for (a) uncalibrated and (b) calibrated Clarity sensors at the 1 km distance threshold in the Northeast United States.

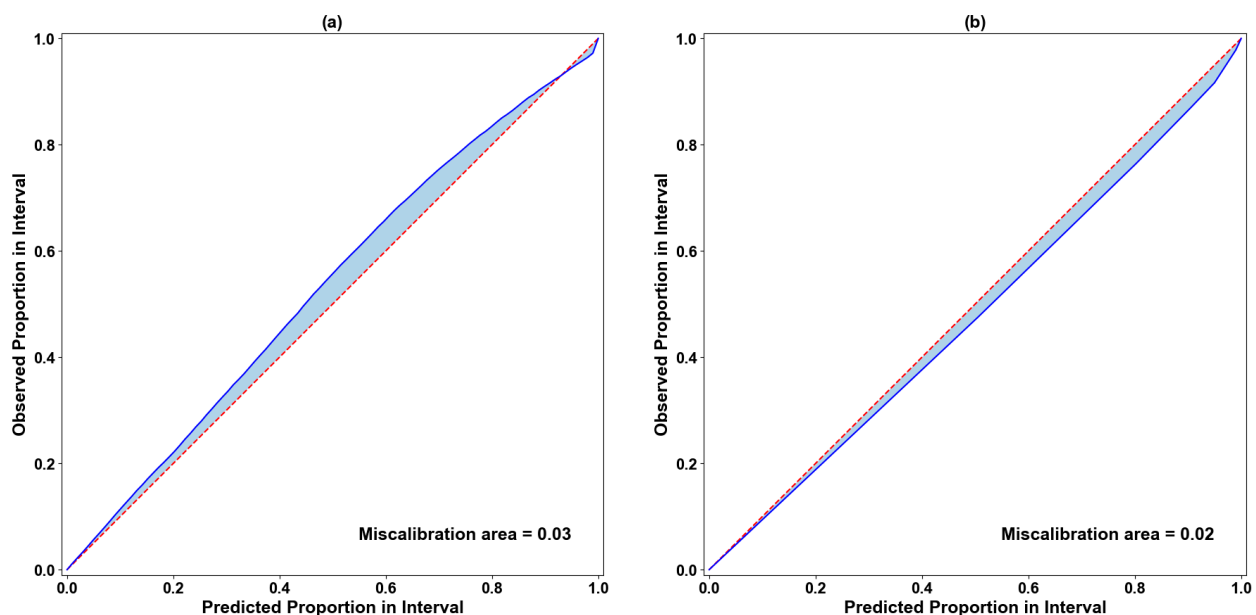


Figure 6. Calibration curves for MCD (a) and GCP (b) uncertainty estimates in California. The red dashed line denotes perfect calibration; the blue curve shows observed coverage, with the shaded area representing the miscalibration area.

The GCP uncertainty approach was evaluated using the same uncertainty metrics to assess its calibration performance. GCP produces values that are generally close to the nominal confidence levels but exhibits consistent undercoverage across all levels. For example, at the 80% and 90% confidence levels, observed PICP values of 0.76 and 0.86 indicate mild undercoverage. The ECE of 0.03 suggests good overall calibration, though

slightly worse than MCD. In terms of prediction interval width, GCP produces narrower intervals at lower confidence levels (e.g., $1.67 \mu\text{g}/\text{m}^3$ at 50% vs. $2.37 \mu\text{g}/\text{m}^3$ for MCD) but wider intervals at the 99% confidence level ($12.90 \mu\text{g}/\text{m}^3$ vs. $9.12 \mu\text{g}/\text{m}^3$), suggesting that GCP expands its intervals more aggressively at higher confidence levels to compensate for undercoverage.

4.3.2. Northeast

The calibration curves for MCD and GCP at the 1 km distance threshold in the Northeast are presented in Figure 7. The MCD calibration curve (Figure 7a) lies above the ideal diagonal across much of the confidence range, indicating overcoverage and a higher miscalibration area of 0.08. In contrast, the GCP calibration curve (Figure 7b) falls below the diagonal, indicating systematic undercoverage, but exhibits a lower miscalibration area of 0.03 and remains closer to the ideal diagonal overall. The largest deviations for GCP occur at mid-to-high confidence levels, particularly around the 80% and 90% confidence bounds.

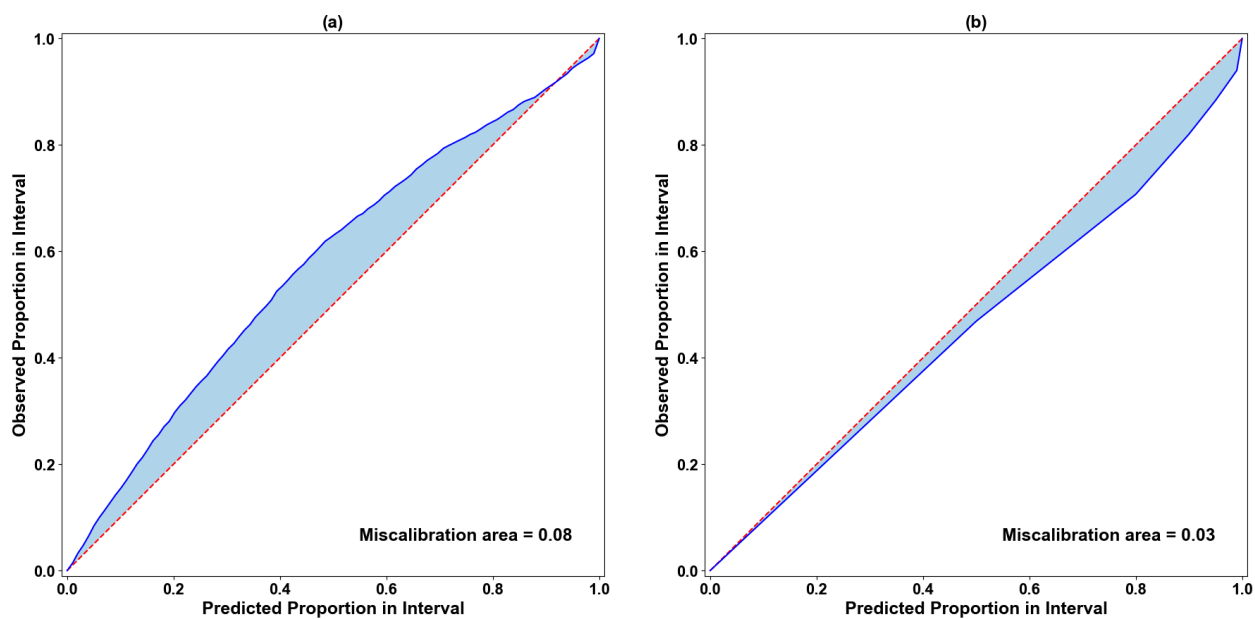


Figure 7. Calibration curves for MCD (a) and GCP (b) uncertainty estimates in the Northeastern United States. The red dashed line denotes perfect calibration; the blue curve shows observed coverage, with the shaded area representing the miscalibration area.

Table S2 presents the PICP, PIW, and ECE metrics for MCD and GCP approaches at the 1 km distance threshold in the Northeast. The MCD approach demonstrates good calibration performance, with observed PICP values generally aligning with nominal confidence levels. However, at the 50% confidence level, notable overcoverage is observed (observed PICP of 0.67 vs. ideal of 0.50), while higher confidence levels remain well calibrated (e.g., 90%: 0.91, 99%: 0.97). The ECE of 0.06 is higher than the value observed in California (0.02), indicating a slightly weaker overall calibration in the Northeast. PIW values are consistently lower than in California, ranging from $1.23 \mu\text{g}/\text{m}^3$ at 50% to $4.74 \mu\text{g}/\text{m}^3$ at 99%, suggesting tighter prediction intervals in this region.

The GCP approach in the Northeast shows consistent undercoverage across all confidence levels, with observed PICP values falling short of the nominal levels. Undercoverage is most pronounced at lower confidence levels, with observed PICP values of 0.47, 0.71, and 0.82 at the 50%, 80%, and 90% confidence levels, respectively. The ECE of 0.06 matches that of MCD and is higher than that observed in California, indicating weaker uncertainty calibration in the Northeast. GCP produces notably narrower prediction in-

tervals compared to MCD across all confidence levels (e.g., $0.66 \mu\text{g}/\text{m}^3$ vs. $1.23 \mu\text{g}/\text{m}^3$ at 50%), except for the 99% confidence level ($6.01 \mu\text{g}/\text{m}^3$ vs. $4.74 \mu\text{g}/\text{m}^3$), suggesting that GCP expands its intervals more aggressively at higher confidence levels to mitigate its systematic undercoverage.

4.4. Uncertainty Decomposition

Predictive uncertainty was decomposed into epistemic and aleatoric components to examine how each contributes to the total uncertainty across spatial distances and wildfire conditions. Table 9 reports the mean epistemic and aleatoric uncertainty across the three distance thresholds for both regions. Aleatoric uncertainty dominates in all cases, exceeding epistemic uncertainty by a factor of roughly six in California and five in the Northeast. Both components increase modestly with sensor separation, with the largest values occurring at the 10 km threshold, indicating that greater spatial distance introduces additional irreducible variability rather than model uncertainty alone.

Table 9. Uncertainty decomposition of MCD prediction intervals, showing aleatoric, epistemic, and total uncertainties across study regions and distance thresholds.

Region	Distance	Epistemic ($\mu\text{g}/\text{m}^3$)	Aleatoric ($\mu\text{g}/\text{m}^3$)	Total ($\mu\text{g}/\text{m}^3$)
CA	1 km	0.32	1.97	2.28
CA	5 km	0.33	1.99	2.31
CA	10 km	0.33	2.19	2.52
NE	1 km	0.13	0.72	0.85
NE	5 km	0.14	0.67	0.81
NE	10 km	0.22	0.88	1.10

Figure 8 shows the decomposition of epistemic and aleatoric uncertainty under normal and wildfire conditions. In California (Figure 8a), both components rise sharply during wildfire events, with aleatoric uncertainty increasing from approximately 1.6 to $15.0 \mu\text{g}/\text{m}^3$ and epistemic uncertainty from 0.4 to $6.4 \mu\text{g}/\text{m}^3$. The dominance of aleatoric uncertainty indicates that the elevated uncertainty during wildfires is driven primarily by increased observational and atmospheric variability, although the rise in epistemic uncertainty also suggests greater model uncertainty under smoke-impacted conditions. In the Northeast, both epistemic and aleatoric uncertainties were lower than in California under both normal and wildfire conditions, with only a modest increase during the wildfire event.

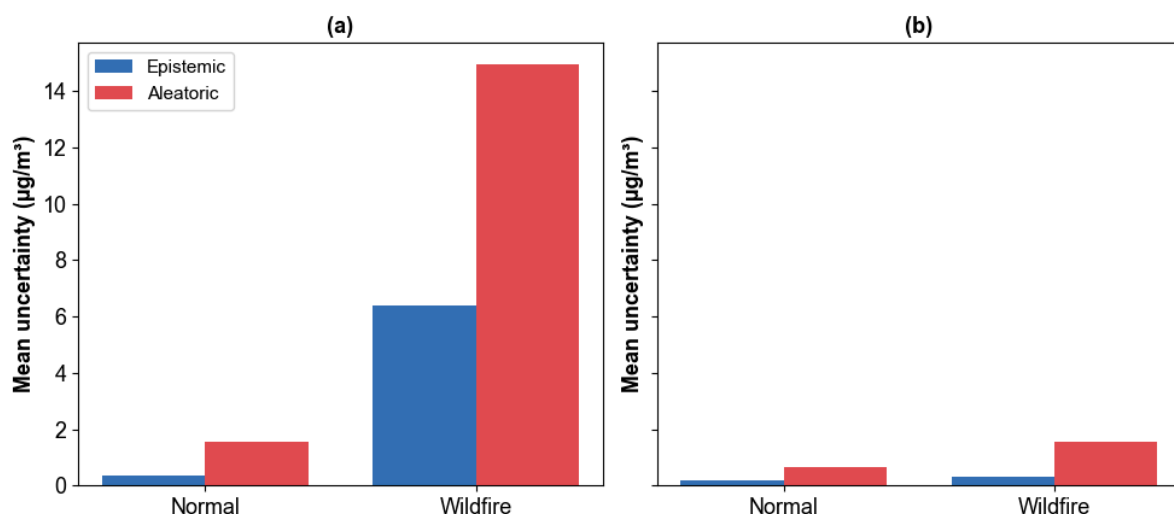


Figure 8. Epistemic and aleatoric uncertainties under normal and wildfire conditions in (a) California and (b) the Northeast.

4.5. Out-of-Distribution Detection

Figure 9 shows the distribution of log-transformed squared Mahalanobis distances for normal and wildfire observations. The OOD threshold was set at the empirical 97.5th percentile of the training-set distance distribution, so observations beyond this threshold were considered out-of-distribution, indicating that they lay unusually far from the feature space used for model training. In California (Figure 9a), wildfire observations are clearly separated from normal observations, with a median squared Mahalanobis distance of 252.98 compared with 17.36 under normal conditions. Accordingly, 59.0% of wildfire observations were flagged as OOD, compared with 4.4% under normal conditions, indicating a substantial distribution shift during the wildfire period. The bimodal wildfire distribution suggests that some wildfire-period observations remained similar to normal conditions, whereas smoke-impacted observations formed a distinct high-distance mode that deviated strongly from the training distribution. In the Northeast (Figure 9b), the separation was weaker, with median squared Mahalanobis distances of 54.88 for wildfire observations and 41.53 for normal observations. The OOD fractions were also closer, with 36.9% of wildfire observations and 29.4% of normal observations classified as OOD. This indicates that wildfire observations were more likely to fall outside the training feature distribution, but the contrast was less pronounced than in California. The weaker separation is consistent with the lower PM_{2.5} magnitudes during the Jennings Creek event and suggests a smaller wildfire-related distribution shift in the Northeast.

4.6. UQ During Wildfire vs. Normal Days

4.6.1. Wildfire Days

Figure 10a–c present the calibration results for a California wildfire event at 29.40 m sensor separation, characterized by a PM_{2.5} spike exceeding 350 µg/m³. The MCD uncertainty band expands substantially during the peak, achieving 71% coverage of the ground-truth observations during the event (Table 10). In contrast, the GCP interval remains comparatively narrow during the event, resulting in lower coverage.

Figure 10d–f present the corresponding results for a Northeast wildfire event at 6798.84 m sensor separation, with PM_{2.5} concentrations reaching approximately 60 µg/m³. In this case, the MCD uncertainty band remains relatively narrow, achieving only 44% coverage of ground-truth observations during the event (Table 10). Conversely, GCP expands its prediction intervals during the event, achieving 83% coverage (Table 10). These results demonstrate that the relative performance of MCD and GCP varies with the spatial and environmental characteristics of wildfire events.

Table 10. Summary of MCD and GCP uncertainty band behavior and ground truth coverage at the 95% confidence level under wildfire and normal conditions in California and the Northeast United States.

Condition	Region	Distance	MCD Band	MCD Coverage	GCP Band	GCP Coverage
Wildfire	California	29.40 m	Expands at spike	69/97 (71%)	Stays narrow	59/97 (61%)
Wildfire	Northeast	6798.84 m	Stays narrow	53/121 (44%)	Expands	100/121 (83%)
Normal	California	2.92 m	Narrow, efficient	117/145 (81%)	Wider throughout	140/145 (97%)
Normal	Northeast	83.69 m	Narrow, efficient	229/313 (73%)	Wider throughout	288/313 (92%)

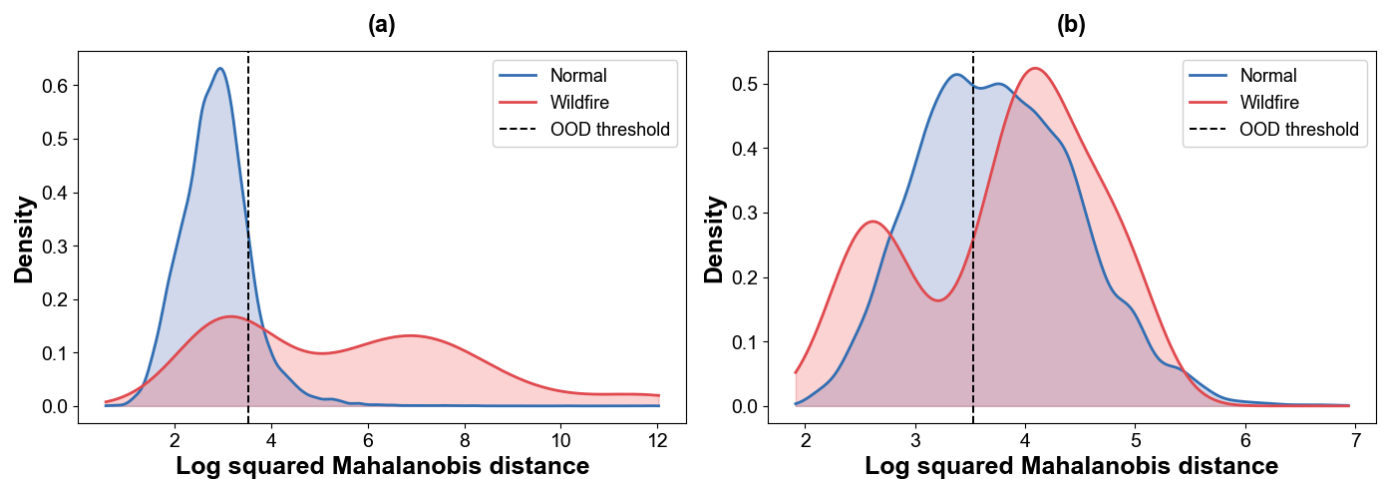


Figure 9. Distribution of log-transformed squared Mahalanobis distances for normal and wildfire observations in (a) California and (b) the Northeast. The dashed vertical line indicates the OOD threshold defined as the empirical 97.5th percentile of the training-set distance distribution.

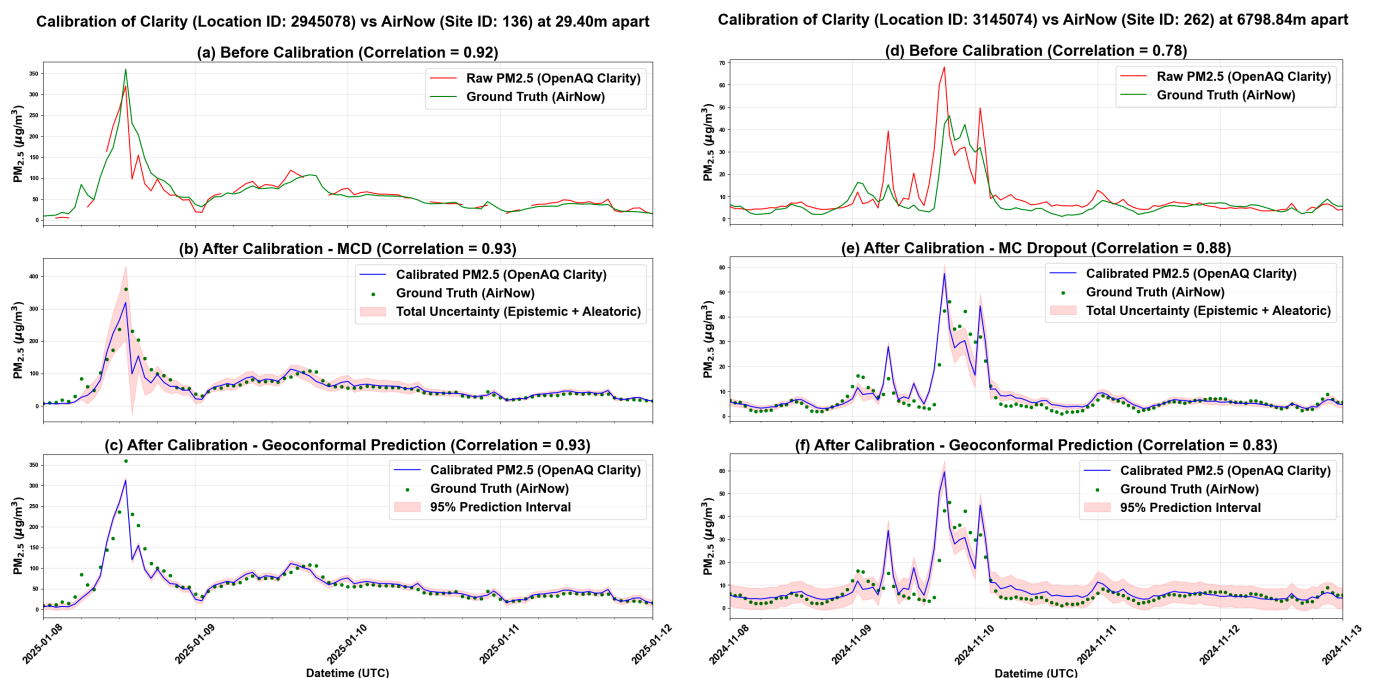


Figure 10. Time-series calibration of Clarity PM_{2.5} against AirNow ground truth for wildfire events in California (a–c) at 29.40 m and the Northeast (d–f) at 6798.84 m spatial separation at a 95% nominal prediction interval, showing (a,d) before calibration, (b,e) after calibration with MCD, and (c,f) after calibration with GCP. Gaps in the time series correspond to periods of missing PM_{2.5} observations.

4.6.2. Normal Days

Figure 11 presents the time-series calibration of Clarity PM_{2.5} against AirNow ground truth under normal (non-wildfire) conditions in California (a–c) at 2.92 m and the Northeast (d–f) at 83.69 m spatial separation. Under normal conditions, both methods demonstrate strong calibration performance, with calibrated predictions closely tracking the ground truth across both regions. In California (Figure 11b), the MCD uncertainty band is narrow and efficient, capturing 81% of ground truth observations, while the GCP interval in Figure 11c is consistently wider throughout the time series, capturing 96% of observations.

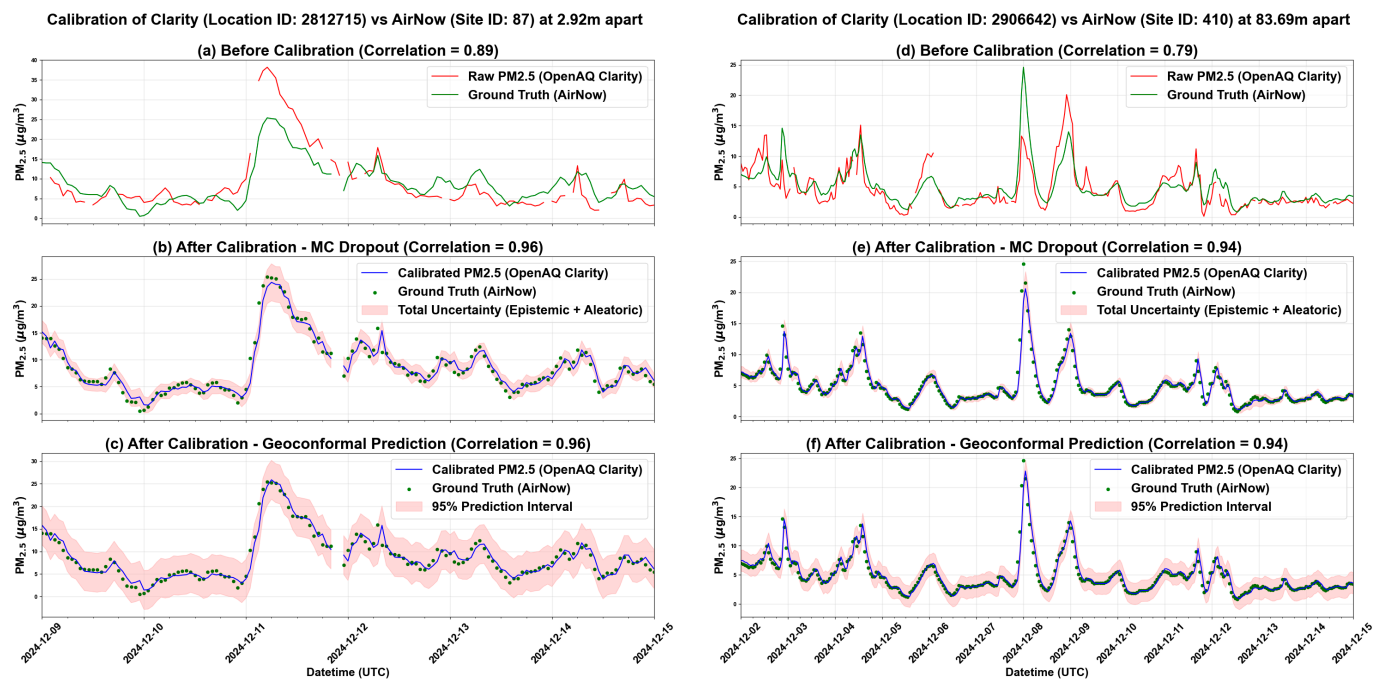


Figure 11. Time-series calibration of Clarity $PM_{2.5}$ against AirNow ground truth for a normal day in California (a–c) at 2.92 m and the Northeast (d–f) at 83.69 m spatial separation at a 95% nominal prediction interval, showing (a,d) before calibration, (b,e) after calibration with MCD, and (c,f) after calibration with GCP. Gaps in the time series correspond to periods of missing $PM_{2.5}$ observations.

In the Northeast (Figure 11e), MCD similarly produces narrow, efficient intervals with 73% coverage, while GCP in Figure 11f maintains wider intervals throughout, achieving 92% coverage. The wider GCP intervals reflect its residual-based calibration framework, which prioritizes coverage reliability over interval sharpness [62]. In contrast, MCD estimates uncertainty through stochastic forward passes, producing intervals that more closely follow local prediction variability. Overall, both methods track the ground truth well under normal conditions, with GCP providing higher coverage at the cost of wider intervals and MCD offering more efficient but less conservative uncertainty estimates.

4.7. Spatial Uncertainty

Figure 12 illustrates the spatial distribution of GCP-derived geospatial uncertainty across California at the 1 km distance threshold. The uncertainty varies considerably across sensor locations, ranging from approximately 1 to 6 $\mu g/m^3$, reflecting the spatially adaptive nature of the GCP framework. Sensors located in the San Francisco Bay Area and along the central coastal region exhibit lower uncertainty values (1–2 $\mu g/m^3$), suggesting that the model performs more consistently in these areas, likely due to higher sensor density and more homogeneous $PM_{2.5}$ conditions. In contrast, the sensor in the southern coastal region near San Diego exhibits the highest uncertainty (approximately 6 $\mu g/m^3$), indicating greater prediction variability, which may be attributed to sparse calibration data in the surrounding area or distinct local $PM_{2.5}$ dynamics [63]. Spatial autocorrelation of errors and uncertainties was assessed using Moran's I to verify whether residual uncertainty exhibits systematic spatial patterns. Moran's I was 0.61 ($p = 0.001$), indicating significant positive spatial autocorrelation. Residual errors are spatially clustered rather than randomly distributed, confirming that residual uncertainty exhibits systematic spatial patterns that could affect the validity of the prediction intervals.

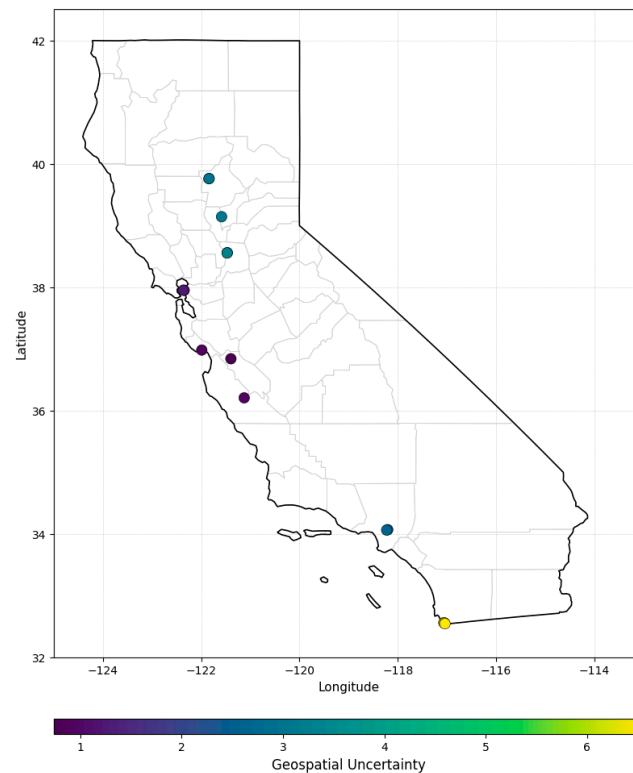


Figure 12. Spatial distribution of GCP-derived geospatial uncertainty ($\mu\text{g}/\text{m}^3$) across California sensor locations at the 1 km distance threshold.

5. Discussion

5.1. Comparison of Uncertainty Performance

GCP generally exhibited calibration curves that remained closer to the ideal diagonal than MCD, indicating more consistent calibration across confidence levels [15]. The contrasting overcoverage of MCD and undercoverage of GCP suggest that the two methods characterize predictive uncertainty differently [58]. Regional differences indicate that uncertainty calibration is influenced by local spatial characteristics, with both methods exhibiting weaker calibration performance in the Northeast than in California [64]. The Northeast results show that similar ECE values do not necessarily correspond to similar calibration performance. Miscalibration area and calibration curves provided additional insight into uncertainty reliability beyond that captured by ECE alone. These findings emphasize the importance of evaluating uncertainty calibration using multiple complementary metrics rather than relying on a single performance measure.

The event-based analysis revealed important differences in the behavior of MCD and GCP under wildfire and normal conditions. MCD and GCP exhibited a trade-off between interval efficiency and coverage reliability across the evaluated scenarios. Neither method consistently outperformed the other across all wildfire events, with MCD achieving higher coverage during the California wildfire event and GCP achieving substantially higher coverage during the Northeast wildfire event. These differences suggest that the relative performance of UQ methods depends strongly on the spatial and environmental characteristics of the event. The contrasting behavior of the two methods may reflect their different interval construction mechanisms, with MCD deriving uncertainty from model-estimated variance and GCP calibrating interval width using local residual information [58]. Under normal conditions, GCP consistently achieved higher coverage with wider prediction intervals, whereas MCD produced narrower and more efficient intervals.

These findings align with prior work on uncertainty quantification under distribution shift. Model-based uncertainty methods are known to show reduced calibration as distribution shift increases, producing overconfident or unreliable prediction intervals [16]. This is consistent with the MCD undercoverage in the Northeast wildfire event, where MCD captured only 44% of observations. Conformal methods construct intervals from empirical residuals rather than model-based variance estimates, allowing the intervals to reflect observed calibration errors. Under a distribution shift, coverage can decrease if calibration does not account for the new data conditions. Weighted and adaptive conformal methods address this by reweighting residuals toward the target distribution or updating calibration over time [65,66]. GCP applies this principle spatially by weighting residuals according to geographic proximity, producing wider intervals in the affected region and improving coverage during the same wildfire event, reaching 83%. Evidential deep learning offers an efficient single-pass alternative that decomposes aleatoric and epistemic uncertainty, but its estimates also become underdispersive under distribution shift [67], which is consistent with the undercoverage behavior observed for MCD here. Distance-based OOD diagnostics further characterize this shift, separating low-error from high-error regions through the Mahalanobis distance [68], consistent with the stronger out-of-distribution separation observed in California than in the Northeast.

The uncertainty decomposition and spatial autocorrelation analyses show that the observed differences between MCD and GCP are linked to both the uncertainty source and the spatial structure. Aleatoric uncertainty dominated across regions and distance thresholds, indicating that much of the uncertainty reflects atmospheric and observational variability rather than model uncertainty alone. Both aleatoric and epistemic uncertainty increased during wildfire conditions, especially in California, suggesting that smoke-impacted conditions increase both irreducible variability and model uncertainty. Moran's I also showed significant spatial clustering of residual uncertainty, supporting the need for spatially informed UQ methods when uncertainty varies systematically across locations.

5.2. Comparison with Existing Studies

Table 11 compares this study with recent LCS calibration studies in terms of study area, sensor type, model, performance metrics, wildfire evaluation, and UQ. Compared to existing studies, this study achieved competitive baseline performance ($R^2 = 0.89$ in California and 0.91 in the NE), comparable to high-performing urban calibration models and outperforming tree-based spatial approaches [27,69]. Previous calibration studies, such as Park, Yoo, Park and Lee [13] primarily evaluated calibration performance under standard environmental conditions and did not explicitly assess model reliability during wildfire-driven distribution shifts. Similarly, Delp and Singer [12] addressed wildfire smoke impacts using fixed empirical scaling relationships rather than adaptive deep learning calibration. Furthermore, this study addresses an important gap in the $PM_{2.5}$ calibration literature related to predictive reliability. Previous calibration studies either lacked formal UQ entirely or limited uncertainty analysis to descriptive post hoc variability assessments [69] or distance-dependent error analyses [27]. In contrast, this study introduces a comparative evaluation of model-based and model-agnostic uncertainty quantification frameworks for low-cost $PM_{2.5}$ sensor calibration. Another major contribution of this work is the systematic evaluation of uncertainty reliability under both normal and extreme pollution conditions, extending calibration assessment beyond predictive accuracy alone.

From a practical perspective, UQ method selection should depend on spatial separation and the expected degree of distribution shift. When sensors are close to reference sites and conditions remain similar to the training distribution, MCD may be suitable because it provides narrower intervals. When spatial separation or distribution shifts are large,

such as during wildfire smoke episodes, GCP may be preferable because it provides more reliable coverage through wider intervals. In low-cost sensor applications, calibrated PM_{2.5} estimates should be reported together with prediction intervals so that air quality agencies can identify when estimates are reliable and when caution is needed. These intervals can also help identify locations where additional sensor deployment or reference-site support may be needed, supporting more reliable exposure assessment, wildfire-smoke monitoring, and public-health decision-making.

Table 11. Comparison of this study with recent low-cost PM_{2.5} sensor calibration studies.

Study	Study Area	Sensor	Model	R ²	RMSE	Wildfire	UQ Applied?	Notes
This study	CA/NE	Clarity, AirNow	Transformer	0.87/0.91	3.04/1.38	Yes	Yes	Evaluates reliability under wildfire and spatial shift
[27]	Sofia, Bulgaria	AirThings + reference	Random Forest	0.75	6.3	No	Distance-based UQ	Spatial uncertainty increases with distance
[22]	India	CMOS LCS	GNN	0.93	4.2	No	NA	Strong ML calibration performance
[69]	Norway	Mobile LCS	XGBoost	0.78	3.97	No	Variability-based	Descriptive uncertainty only
[13]	South Korea	SPS 30	Hybrid LSTM	0.93	~3	No	No	High calibration accuracy only
[12]	USA (CA, UT wildfires)	PurpleAir + reference	Adjustment factors	NA	NA	Yes	No	Wildfire-specific calibration

5.3. Limitations

This study presents a systematic evaluation of UQ methods for PM_{2.5} calibration, but several limitations should guide the interpretation of the findings and future research. Wildfire events were represented through elevated PM_{2.5} concentrations rather than fire-specific physical variables such as distance from the fire source, burned area, air mass trajectories, or satellite-based smoke indicators, which limits the physical interpretation of the distribution shift and extrapolation to other fire scenarios. Other extreme conditions, such as dust storms, fireworks, winter stagnation episodes, and industrial pollution events, can also produce sharp PM_{2.5} increases, and their probability and spatial extent may vary by region, season, and local emission sources. These conditions were not considered in this study. The wildfire and normal-condition comparisons were based on individual sensor pairs per region, which constrains the generalizability of the event-specific coverage results. The calibration models did not account for sensor aging, drift, manufacturing variability, or reference-instrument uncertainty, which can affect calibration accuracy and predictive uncertainty, particularly in long-term deployments. The comparison was also limited to MCD and GCP, without evaluating other methods such as deep ensembles, quantile regression, and evidential deep learning, limiting positioning against the full range of available UQ approaches. Finally, GCP relies on a fixed validation-set residual calibration that does not adapt dynamically to changing conditions, contributing to its wider intervals under normal conditions.

5.4. Future Directions

Although demonstrated using PM_{2.5} calibration, the proposed framework is directly applicable to broader environmental sensing problems, including satellite retrievals, atmospheric composition, hydrology, water quality, weather forecasting, and Earth system digital twins. Several extensions of this work could further improve calibration robustness and broaden its applicability. Future work should examine the operational deployment of calibrated LCS, where the trustworthiness and reliability of their uncertainty estimates are critical for air-quality decision-making [70]. Such deployment would help translate calibrated estimates into actionable exposure information, extending dependable PM_{2.5} monitoring to underserved communities and supporting timely health advisories during extreme pollution episodes [71]. In this context, method selection should balance interval efficiency against coverage reliability, prioritizing coverage where missed exposures carry the greatest risk [72]. Second, data augmentation strategies, such as generating synthetic sensor observations for under-represented pollution regimes, could improve model robustness and reduce predictive uncertainty during extreme events like wildfires [73]. Targeted augmentation of high-PM_{2.5} episodes is particularly promising given the upper-tail underestimation observed in the wildfire-period results. Third, future work should incorporate sensor aging, drift, manufacturing variability, and reference-instrument uncertainty as calibration covariates to quantify their individual contributions to predictive uncertainty, particularly in long-term deployments. Fourth, calibrated LCS could function as a high-resolution observational layer within digital twin systems, densifying the spatial coverage of sparse regulatory networks [74]. Through near-real-time data assimilation, these observations could continuously update the digital twin state, enabling scenario-based forecasting and adaptive air-quality decision-making [75,76] towards a decision infrastructure [76]. Reliable uncertainty quantification is a foundational capability of atmospheric and environmental digital twins because confidence estimates determine when observations can be trusted for forecasting and decision-making. Together, these directions point toward a next generation of uncertainty-aware, spatiotemporally adaptive calibration frameworks that can deliver trustworthy PM_{2.5} estimates across diverse environmental conditions.

6. Conclusions

This study presented a systematic evaluation of UQ methods for PM_{2.5} calibration using LCS data, with a focus on reliability under spatial variability and wildfire-induced distribution shift. A transformer-based calibration model was combined with two complementary UQ methods, MCD and GCP, to assess predictive reliability across California and the Northeast United States. The calibration model achieved strong performance at short spatial distances, with test R^2 values of 0.89 and 0.91 at the 1 km threshold in California and the Northeast, respectively, and accuracy generally decreasing as sensor separation increases, highlighting the sensitivity of calibration to spatial heterogeneity. Both approaches produced reliable uncertainty estimates but differed in how uncertainty is represented based on spatial and temporal constraints. In terms of reliability, GCP generally exhibited calibration curves that more closely followed the ideal diagonal than MCD across the study regions. Under wildfire conditions, coverage depended on sensor separation. MCD held up when sensors were close (71% in California at 29 m) but dropped when they were far apart (44% in the Northeast at 6.8 km), while GCP stayed high (83%). This is because GCP widens its intervals from local residuals, whereas MCD's model-based variance does not adjust to the shift. In practice, MCD provides narrow, efficient intervals under normal conditions and scales effectively during highly localized pollution anomalies. At larger spatial separations, GCP generally achieved higher coverage through wider prediction intervals. These findings demonstrate that uncertainty behavior is highly context-dependent

and that event-specific analyses reveal patterns that may not be apparent from aggregate performance metrics alone.

A limitation of this study is that wildfire events were characterized through elevated $PM_{2.5}$ rather than fire-specific physical variables, which constrains extrapolation to other fire scenarios. To conclude, this study provides a practical basis for using low-cost sensors in operational air-quality monitoring by quantifying calibrated $PM_{2.5}$ and its uncertainty. This helps decision-makers judge when measurements are reliable enough for wildfire health advisories and guides sensor deployment in regions with sparse regulatory coverage. Future work should explore data augmentation strategies for high- $PM_{2.5}$ events, adaptive calibration approaches for GCP, systematic benchmarking against other UQ methods such as deep ensembles and evidential deep learning, and hybrid frameworks that combine the complementary strengths of model-based and model-agnostic uncertainty methods.

This study demonstrates that uncertainty reliability in environmental sensing depends jointly on model uncertainty, spatial transferability, and shifts in environmental distribution. The proposed spatial uncertainty evaluation framework for AI-enabled environmental sensing could also be applied to pollutants beyond $PM_{2.5}$, such as ozone, NO_2 , methane, temperature, precipitation, water quality, soil moisture, and satellite retrievals. This work contributes a foundation for trustworthy environmental AI by integrating spatially adaptive uncertainty quantification, out-of-distribution detection, and reliability assessment into operational environmental sensing workflows. These capabilities are expected to become fundamental components of future environmental digital twins and AI-enabled decision infrastructures.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/atmos17070692/s1>. Figure S1: Box plots of $PM_{2.5}$ concentrations for AirNow and Clarity sensor pairs at (a) 1 km, (b) 5 km, and (c) 10 km distance thresholds in California; Figure S2: Box plots of $PM_{2.5}$ concentrations for AirNow and Clarity sensor pairs at (a) 1 km, (b) 5 km, and (c) 10 km distance thresholds in the Northeast; Figure S3: $PM_{2.5}$ concentration distributions for AirNow at (a) 1 km, (b) 5 km, and (c) 10 km, and Clarity at (d) 1 km, (e) 5 km, and (f) 10 km distance thresholds in California; Figure S4: $PM_{2.5}$ concentration distributions for AirNow at (a) 1 km, (b) 5 km, and (c) 10 km, and Clarity at (d) 1 km, (e) 5 km, and (f) 10 km distance thresholds in the Northeast; Table S1: Calibration and uncertainty-interval metrics for MCD and GCP approaches across confidence levels at the 1 km distance threshold in California; Table S2: Calibration and uncertainty-interval metrics for MCD and GCP approaches across confidence levels at the 1 km distance threshold in the Northeast.

Author Contributions: Conceptualization, A.S.M., K.S., J.T.R., G.C. and C.Y.; methodology, A.S.M. and K.S.; software, A.S.M.; validation, A.S.M. and K.S.; formal analysis, A.S.M. and K.S.; investigation, A.S.M. and K.S.; resources, A.S.M.; data curation, S.S.; writing—original draft preparation, A.S.M. and K.S.; writing—review and editing, A.S.M.; visualization, A.S.M. and K.S.; supervision, C.Y.; project administration, C.Y.; funding acquisition, C.Y., G.C. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the NSF I/UCRC Program [1841520, 2232846] and the NASA AIST Program [80NSSC23K1023] and NASA Goddard CISTO [80NSSC21P2373], NOAA, and the many members of the Spatiotemporal I/UCRC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The low-cost sensor data used in this study were obtained from the OpenAQ platform (<https://openaq.org> (accessed on 1 June 2026)). Reference-grade $PM_{2.5}$ measurements were obtained from the U.S. EPA AirNow network via the AirNow API (<https://docs.airnowapi.org/> (accessed on 1 June 2026)).

Acknowledgments: During the preparation of this manuscript/study, the authors used OpenAI ChatGPT 5.5 for the purpose of language checking. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PM _{2.5}	Fine Particulate Matter (≤ 2.5 micrometers)
EPA	Environmental Protection Agency
LCS	Low-Cost Sensors
UQ	Uncertainty Quantification
BNNs	Bayesian Neural Networks
MCD	Monte Carlo Dropout
DE	Deep Ensembles
GCP	GeoConformal Prediction
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
R ²	Coefficient of Determination
MSE	Mean Squared Error
NE	Northeast (United States)
CA	California
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
LULC	Land Use and Land Cover
NLCD	National Land Cover Database
USGS	United States Geological Survey
FIPS	Federal Information Processing Standards
DoY	Day of Year
DoW	Day of Week
WoY	Week of Year
NOAA	National Oceanic and Atmospheric Administration
HRRR	High-Resolution Rapid Refresh
AWS	Amazon Web Services
S3	Simple Storage Service
API	Application Programming Interface
NLL	Negative Log-Likelihood
ECE	Expected Calibration Error
PICP	Prediction Interval Coverage Probability
PIW	Prediction Interval Width
GNN	Graph Neural Network

References

1. Reid, C.E.; Brauer, M.; Johnston, F.H.; Jerrett, M.; Balme, J.R.; Elliott, C.T. Critical Review of Health Impacts of Wildfire Smoke Exposure. *Environ. Health Perspect.* **2016**, *124*, 1334–1343. [[CrossRef](#)] [[PubMed](#)]
2. Smith, S.; Trefonides, T.; Srenganathan Malarvizhi, A.; LaGarde, S.; Liu, J.; Jia, X.; Wang, Z.; Cain, J.; Huang, T.; Pourhomayoun, M.; 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. *Sensors* **2025**, *25*, 1028. [[CrossRef](#)] [[PubMed](#)]
3. He, J.; Huang, C.H.; Yuan, N.; Austin, E.; Seto, E.; Novosselov, I. Network of low-cost air quality sensors for monitoring indoor, outdoor, and personal PM_{2.5} exposure in Seattle during the 2020 wildfire season. *Atmos. Environ.* **2022**, *285*, 119244. [[CrossRef](#)]
4. Kramer, A.L.; Liu, J.; Li, L.; Connolly, R.; Barbato, M.; Zhu, Y. Environmental justice analysis of wildfire-related PM_{2.5} exposure using low-cost sensors in California. *Sci. Total Environ.* **2023**, *856*, 159218. [[CrossRef](#)] [[PubMed](#)]
5. Concas, F.; Mineraud, J.; Lagerspetz, E.; Varjonen, S.; Liu, X.; Puolamäki, K.; Nurmi, P.; Tarkoma, S. Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration. *ACM Trans. Sens. Netw.* **2021**, *17*, 20. [[CrossRef](#)]

6. Chu, H.-J.; Ali, M.Z.; He, Y.-C. Spatial calibration and PM2.5 mapping of low-cost air quality sensors. *Sci. Rep.* **2020**, *10*, 22079. [[CrossRef](#)] [[PubMed](#)]
7. Barkjohn, K.K.; Holder, A.L.; Frederick, S.G.; Clements, A.L. Correction and Accuracy of PurpleAir PM2.5 Measurements for Extreme Wildfire Smoke. *Sensors* **2022**, *22*, 9669. [[CrossRef](#)] [[PubMed](#)]
8. Raheja, G.; Nimo, J.; Appoh, E.K.E.; Essien, B.; Sunu, M.; Nyante, J.; Amegah, M.; Quansah, R.; Arku, R.E.; Penn, S.L.; et al. Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient PM2.5 Monitoring in Accra, Ghana. *Environ. Sci. Technol.* **2023**, *57*, 10708–10720. [[CrossRef](#)] [[PubMed](#)]
9. Bi, J.; Wildani, A.; Chang, H.H.; Liu, Y. Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale. *Environ. Sci. Technol.* **2020**, *54*, 2152–2162. [[CrossRef](#)] [[PubMed](#)]
10. Wang, Y.; Du, Y.; Wang, J.; Li, T. Calibration of a low-cost PM2.5 monitor using a random forest model. *Environ. Int.* **2019**, *133*, 105161. [[CrossRef](#)] [[PubMed](#)]
11. Berrocal, V.J.; Guan, Y.; Muyskens, A.; Wang, H.; Reich, B.J.; Mulholland, J.A.; Chang, H.H. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration. *Atmos. Environ.* **2020**, *222*, 117130. [[CrossRef](#)] [[PubMed](#)]
12. Delp, W.W.; Singer, B.C. Wildfire Smoke Adjustment Factors for Low-Cost and Professional PM2.5 Monitors with Optical Sensors. *Sensors* **2020**, *20*, 3683. [[CrossRef](#)] [[PubMed](#)]
13. Park, D.; Yoo, G.W.; Park, S.H.; Lee, J.H. Assessment and Calibration of a Low-Cost PM2.5 Sensor Using Machine Learning (HybridLSTM Neural Network): Feasibility Study to Build an Air Quality Monitoring System. *Atmosphere* **2021**, *12*, 1306. [[CrossRef](#)]
14. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5580–5590.
15. Lou, X.; Luo, P.; Meng, L. GeoConformal Prediction: A Model-Agnostic Framework for Measuring the Uncertainty of Spatial Prediction. *Ann. Am. Assoc. Geogr.* **2025**, *115*, 1971–1998. [[CrossRef](#)]
16. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019; p. 1254.
17. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [[CrossRef](#)]
18. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized quantile regression. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 3543–3553.
19. Lee, C.H.; Wang, Y.B.; Yu, H.L. An efficient spatiotemporal data calibration approach for the low-cost PM2.5 sensing network: A case study in Taiwan. *Environ. Int.* **2019**, *130*, 104838. [[CrossRef](#)] [[PubMed](#)]
20. Aula, K.; Lagerspetz, E.; Nurmi, P.; Tarkoma, S. Evaluation of Low-cost Air Quality Sensor Calibration Models. *ACM Trans. Sens. Netw.* **2022**, *18*, 72. [[CrossRef](#)]
21. Wang, A.; Machida, Y.; deSouza, P.; Mora, S.; Duhl, T.; Hudda, N.; Durant, J.L.; Duarte, F.; Ratti, C. Leveraging machine learning algorithms to advance low-cost air sensor calibration in stationary and mobile settings. *Atmos. Environ.* **2023**, *301*, 119692. [[CrossRef](#)]
22. Guruprakash, B.; Chouhan, P.S. Graph Neural Network-Enhanced CMOS-Based Low-Cost Air Pollution Monitoring for Scalable Environmental Sensing Systems. *ICTACT J. Microelectron.* **2025**, *11*, 2189–2195. [[CrossRef](#)]
23. Sousan, S.; Wu, R.; Popoviciu, C.; Fresquez, S.; Park, Y.M. Advancing low-cost air quality monitor calibration with machine learning methods. *Environ. Pollut.* **2025**, *374*, 126191. [[CrossRef](#)] [[PubMed](#)]
24. Hua, J.; Zhang, Y.; de Foy, B.; Mei, X.; Shang, J.; Zhang, Y.; Sulaymon, I.D.; Zhou, D. Improved PM2.5 concentration estimates from low-cost sensors using calibration models categorized by relative humidity. *Aerosol Sci. Technol.* **2021**, *55*, 600–613. [[CrossRef](#)]
25. Adong, P.; Bainomugisha, E.; Okure, D.; Sserunjogi, R. Applying machine learning for large scale field calibration of low-cost PM2.5 and PM10 air pollution sensors. *Appl. AI Lett.* **2022**, *3*, e76. [[CrossRef](#)]
26. Bush, T.; Papaioannou, N.; Leach, F.; Pope, F.D.; Singh, A.; Thomas, G.N.; Stacey, B.; Bartington, S. Machine learning techniques to improve the field performance of low-cost air quality sensors. *Atmos. Meas. Tech.* **2022**, *15*, 3261–3278. [[CrossRef](#)]
27. Zhivkov, P.; Fidanova, S. Machine Learning Calibration Transfer for Low-Cost Air Quality Sensors: Distance-Based Uncertainty Quantification in a Hybrid Urban Monitoring Network. *Atmosphere* **2026**, *17*, 335. [[CrossRef](#)]
28. Gómez, R.; Rodríguez, J.; Ferro, R. Calibration of Low-Cost Sensors for PM₁₀ and PM_{2.5} Based on Artificial Intelligence for Smart Cities. *Sensors* **2026**, *26*, 796. [[CrossRef](#)] [[PubMed](#)]
29. Liu, J.C.; Mickley, L.J.; Sulprizio, M.P.; Dominici, F.; Yue, X.; Ebisu, K.; Anderson, G.B.; Khan, R.F.A.; Bravo, M.A.; Bell, M.L. Particulate Air Pollution from Wildfires in the Western US under Climate Change. *Clim. Change* **2016**, *138*, 655–666. [[CrossRef](#)] [[PubMed](#)]

30. Holder, A.L.; Mebust, A.K.; Maghran, L.A.; McGown, M.R.; Stewart, K.E.; Vallano, D.M.; Elleman, R.A.; Baker, K.R. Field Evaluation of Low-Cost Particulate Matter Sensors for Measuring Wildfire Smoke. *Sensors* **2020**, *20*, 4796. [CrossRef] [PubMed]
31. Ahangar, F.E.; Cobian-Iñiguez, J.; Cisneros, R. Combining Regulatory Instruments and Low-Cost Sensors to Quantify the Effects of 2020 California Wildfires on PM_{2.5} in San Joaquin Valley. *Fire* **2022**, *5*, 64. [CrossRef]
32. Koldasbayeva, D.; Tregubova, P.; Gasanov, M.; Zaytsev, A.; Petrovskaya, A.; Burnaev, E. Challenges in data-driven geospatial modeling for environmental research and practice. *Nat. Commun.* **2024**, *15*, 10700. [CrossRef] [PubMed]
33. Barkjohn, K.K.; Gantt, B.; Clements, A.L. Development and application of a United States-wide correction for PM_{2.5} data collected with the PurpleAir sensor. *Atmos. Meas. Tech.* **2021**, *14*, 4617–4637. [CrossRef] [PubMed]
34. Malings, C.; Tanzer, R.; Hauryliuk, A.; Saha, P.K.; Robinson, A.L.; Presto, A.A.; Subramanian, R. Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Sci. Technol.* **2020**, *54*, 160–174. [CrossRef]
35. Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Model. Softw.* **2018**, *101*, 1–9. [CrossRef]
36. Lin, Y.C.; Chi, W.J.; Lin, Y.Q. The improvement of spatial-temporal resolution of PM_{2.5} estimation based on micro-air quality sensors by using data fusion technique. *Environ. Int.* **2020**, *134*, 105305. [CrossRef] [PubMed]
37. Wallace, L.; Zhao, T. Spatial Variation of PM_{2.5} Indoors and Outdoors: Results from 261 Regulatory Monitors Compared to 14,000 Low-Cost Monitors in Three Western States over 4.7 Years. *Sensors* **2023**, *23*, 4387. [CrossRef] [PubMed]
38. Kumar, V.; Senarathna, D.; Gurajala, S.; Dhaniyala, S.; Mondal, S.; Sur, S. Effects of distance and number of PM_{2.5} low-cost sensors on correction models. *Int. J. Data Sci. Anal.* **2025**, *20*, 6781–6796. [CrossRef]
39. Malarvizhi, A.S.; Smith, K.; Yang, C. Uncertainty quantification in geospatial AI/ML applications: Methods, metrics, and open-source support with an air quality use case. *Big Earth Data* **2026**, 1–34. [CrossRef]
40. Starr, C.; Christensen, A.J.; Shirah, G. Spread of the Palisades and Eaton Fires—January 2025. Available online: <https://doi.org/10.5281/zenodo.16782807> (accessed on 11 July 2026). [CrossRef]
41. Rafferty, J.P. Los Angeles Wildfires of 2025. Available online: <https://www.britannica.com/event/Los-Angeles-wildfires-of-2025> (accessed on 2 July 2026).
42. Mann, B. Dry Weather in New York and New Jersey Leads to a Rash of Dangerous Wildfires. Available online: <https://www.npr.org/2024/11/12/nx-s1-5186994/dry-weather-in-new-york-and-new-jersey-leads-to-a-rash-of-dangerous-wildfires> (accessed on 5 May 2026).
43. EPA Airnow. AirNow API Documentation. Available online: <https://docs.airnowapi.org/> (accessed on 2 July 2026).
44. OpenAQ Clarity. Fighting Air Inequality Through Open Data. Available online: <https://openaq.org> (accessed on 2 July 2026).
45. NOAA. High-Resolution Rapid Refresh (HRRR) Model Archive. Available online: <https://registry.opendata.aws/noaa-hrrr-pds/> (accessed on 2 July 2026).
46. Jayaratne, R.; Liu, X.; Thai, P.; Dunbabin, M.; Morawska, L. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmos. Meas. Tech.* **2018**, *11*, 4883–4890. [CrossRef]
47. Mathieu-Campbell, M.E.; Guo, C.; Grieshop, A.P.; Richmond-Bryant, J. Calibration of PurpleAir low-cost particulate matter sensors: Model development for air quality under high relative humidity conditions. *Atmos. Meas. Tech.* **2024**, *17*, 6735–6749. [CrossRef] [PubMed]
48. Dewitz, J. *National Land Cover Database (NLCD) 2021 Products*; U.S. Geological Survey: Reston, VA, USA, 2023. [CrossRef]
49. U. S. Census Bureau. Federal Information Processing Standards (FIPS) Codes for States and Counties. Available online: <https://www.census.gov/library/reference/code-lists/ansi.html> (accessed on 2 July 2026).
50. Malarvizhi, A.S.; Pan, P.; Stover, T.; Sun, D.; Yang, C. Optimizing GAIN model to improve AOD imputation using MODIS MAIAC data and multi-source data fusion as an example. *GISci. Remote Sens.* **2025**, *62*, 2571244. [CrossRef]
51. Zahmatkesh, S.; Zech, P. Spatio-Temporal Missing Data Imputation: A Systematic Literature Review with a Focus on Statistical and Machine Learning-Based Approaches. *ACM Comput. Surv.* **2026**, *58*, 267. [CrossRef]
52. Karl, T.R.; Koss, W.J. *Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, 1895–1983*; National Climatic Data Center: Asheville, NC, USA, 1984.
53. Yang, G.; Lee, H.; Lee, G. A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea. *Atmosphere* **2020**, *11*, 348. [CrossRef]
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30.
55. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
56. Huber, P.J. Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*; Springer: New York, NY, USA, 1992; pp. 492–518.
57. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

58. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
59. Jain, P.; Barber, Q.E.; Taylor, S.W.; Whitman, E.; Castellanos Acuna, D.; Boulanger, Y.; Chavardès, R.D.; Chen, J.; Englefield, P.; Flannigan, M.; et al. Drivers and Impacts of the Record-Breaking 2023 Wildfire Season in Canada. *Nat. Commun.* **2024**, *15*, 6764. [[CrossRef](#)] [[PubMed](#)]
60. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM_{2.5} Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ. Sci. Technol.* **2016**, *50*, 4712–4721. [[CrossRef](#)] [[PubMed](#)]
61. Li, L.; Girguis, M.; Lurmann, F.; Pavlovic, N.; McClure, C.; Franklin, M.; Wu, J.; Oman, L.D.; Breton, C.; Gilliland, F.; et al. Ensemble-based deep learning for estimating PM_{2.5} over California with multisource big data including wildfire smoke. *Environ. Int.* **2020**, *145*, 106143. [[CrossRef](#)] [[PubMed](#)]
62. Zaffran, M.; Feron, O.; Goude, Y.; Josse, J.; Dieuleveut, A. Adaptive Conformal Predictions for Time Series. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 25834–25866.
63. Bashardoost, A.; Mesgari, M.S.; Karimi, M. Quantifying uncertainty in the spatial prediction of PM_{2.5} using a deep learning algorithm. *Ann. GIS* **2025**, *31*, 679–712. [[CrossRef](#)]
64. Zusman, M.; Schumacher, C.S.; Gasset, A.J.; Spalt, E.W.; Austin, E.; Larson, T.V.; Carvlin, G.; Seto, E.; Kaufman, J.D.; Sheppard, L. Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. *Environ. Int.* **2020**, *134*, 105329. [[CrossRef](#)] [[PubMed](#)]
65. Gibbs, I.; Candes, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 1660–1672.
66. Tibshirani, R.J.; Foygel Barber, R.; Candes, E.; Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019; Volume 32.
67. Schreck, J.S.; Gagne, D.J.; Becker, C.; Chapman, W.E.; Elmore, K.; Fan, D.; Gantos, G.; Kim, E.; Kimpara, D.; Martin, T.; et al. Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications. *Artif. Intell. Earth Syst.* **2024**, *3*, 230093. [[CrossRef](#)]
68. Flovik, V. Quantifying distribution shifts and uncertainties for enhanced model robustness in machine learning applications. *arXiv* **2024**, arXiv:2405.01978.
69. Hassani, A.; Castell, N.; Watne, Å.K.; Schneider, P. Citizen-operated mobile low-cost sensors for urban PM_{2.5} monitoring: Field calibration, uncertainty estimation, and application. *Sustain. Cities Soc.* **2023**, *95*, 104607. [[CrossRef](#)]
70. Yang, C.; Malarvizhi, A.S.; Yu, M.; Huang, Q.; Liu, L.; Wang, Z.; Duffy, D.Q.; Wang, S.; Smith, S.; Bao, S.; et al. Spatiotemporal Data Science. *Encyclopedia* **2026**, *6*, 84. [[CrossRef](#)]
71. Lu, T.; Liu, Y.; Garcia, A.; Wang, M.; Li, Y.; Bravo-villasenor, G.; Campos, K.; Xu, J.; Han, B. Leveraging Citizen Science and Low-Cost Sensors to Characterize Air Pollution Exposure of Disadvantaged Communities in Southern California. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8777. [[CrossRef](#)] [[PubMed](#)]
72. Chen, H.; Huang, Z.; Lam, H.; Qian, H.; Zhang, H. Learning Prediction Intervals for Regression: Generalization and Calibration. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, Virtual, 13–15 April 2021; pp. 820–828.
73. Pan, P.; Malarvizhi, A.S.; Yang, C. Data Augmentation Strategies for Improved PM_{2.5} Forecasting Using Transformer Architectures. *Atmosphere* **2025**, *16*, 127. [[CrossRef](#)]
74. Cowell, N.H.; Chapman, L.; Topping, D.; James, P.; Bell, D.; Bannan, T.; Murabito, E.; Evans, J.; Birkin, M. Moving from monitoring to real-time interventions for air quality: Are low-cost sensor networks ready to support urban digital twins? *Front. Sustain. Cities* **2025**, *6*, 1500516. [[CrossRef](#)]
75. Yang, C.; Lan, H.; Sriranganathan, A.; Trefonides, T.; Guan, W.; Huang, Q.; Su, Y.; Yu, M.; Zhang, M.; Zhang, S. Digital Twins. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2025; pp. 1–8. [[CrossRef](#)]
76. Yang, C.; Malarvizhi, A.S.; Masri, Y.; Smith, J.S.; Li, Z.; Huang, Q.; Liu, L.; Kim, J. Digital Twins as Decision Infrastructure: Evolution, Architecture, and Research Roadmap. *Big Earth Data* **2026**, 1–46. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.