

IMPERIAL

A fine-scale spatiotemporal air quality modeling framework by combining big but noisy data

by

Hongwei Fan

Centre for Environmental Policy

Imperial College London

Submitted for the degree of Doctor of Philosophy

London, May 2026

Declaration of Originality

I, Hongwei Fan, declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been clearly and appropriately indicated and referenced.

This thesis includes material that has been submitted for publication or published in academic journals and conference proceedings. Where such material has been included, it represents my own work and contribution, and has been appropriately referenced.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Ambient air pollution is a major public health crisis. Traditional air quality monitoring relies on sparse reference-grade stations, which are costly and unable to capture fine-scale spatial variability. Although low-cost sensor networks provide extensive spatial coverage, their measurements often contain substantial noise and uncertainty. This dissertation addresses the challenge of transforming “big but noisy” spatiotemporal datasets—including low-cost sensor observations, land-use, traffic, and meteorological data—into accurate, high-resolution hourly maps of $\text{PM}_{2.5}$ and NO_2 concentrations.

Firstly, a scalable calibration methodology was developed to significantly enhance data reliability for networks encompassing hundreds of low-cost sensors. A novel transformer-based calibration approach was introduced, uniquely leveraging temporal dependencies (e.g., diurnal variations) in sensor data, thereby reducing the root mean squared error (RMSE) by 10–20% compared to traditional methods.

Secondly, an innovative deep learning framework based on the Vision Transformer (ViT) architecture was introduced, capitalising on its attention mechanism to model intricate spatial patterns across the urban landscape. Applied to calibrated sensor data and auxiliary inputs, the ViT-based model generates $\text{PM}_{2.5}$ and NO_2 predictions at $100 \text{ m} \times 100 \text{ m}$ hourly resolution, consistently outperforming three established baselines.

Finally, comprehensive uncertainty quantification is implemented. Uncertainty arising from calibrated low-cost sensors and proposed ViT-based model predictions are systematically estimated. To effectively integrate these uncertainties, a Bayesian Maximum Entropy (BME) data fusion framework was developed, providing robust probabilistic predictions and rigorous uncertainty quantification for fine-scale urban air quality estimations.

Collectively, this research addresses critical challenges in large-scale calibration of low-cost sensors, development of fine-scale air quality models, and rigorous uncertainty quantification through data fusion. The proposed framework enables reliable urban air quality estimation at fine spatiotemporal scales, supporting evidence-based environmental management and public health policy.

Keywords: Air pollution, low-cost sensors, calibration, deep learning, Vision Transformer, Bayesian Maximum Entropy, spatiotemporal mapping, uncertainty quantification.

Acknowledgements

As I look back on this doctoral journey, my heart is filled with a deep and inexpressible sense of gratitude.

To have the opportunity—after years of full-time work—to set everything aside and devote myself entirely to four years of academic exploration at one of the world’s leading institutions, Imperial College London, has been a rare and extraordinary privilege. This chapter of my life, though marked by sleepless nights, relentless revisions, and moments of uncertainty, now draws to a close. What remains is not the exhaustion, but a quiet sense of fulfilment, profound thankfulness, and enduring emotion.

First and foremost, I owe my deepest gratitude to my two supervisors, Associate Professor Audrey de Nazelle and Associate Professor Rossella Arcucci. You have not only guided me through the academic wilderness but have also stood as role models in life—brilliant, wise, and kind. Throughout my PhD, your presence has been a constant source of inspiration. Whether it was offering warmth and support during my personal challenges, or providing timely and crucial guidance when I felt lost in research—you were always there. Being your student has been one of the greatest blessings of my life. The midnight emails, the insightful discussions, your enduring patience, and your quiet encouragement—these are treasures I will carry with me always.

I would also like to express my sincere appreciation to Professor Benjamin Barratt and the entire Environmental Research Group at Imperial College London. The Breathe London air quality monitoring network, which you built with vision and dedication, laid a crucial foundation for my research. I am especially grateful to Professor Barratt for generously sharing this valuable dataset and for your openness and inclusivity as a scholar. Your contributions not only made this work possible but also profoundly shaped its trajectory. You are true pioneers in this field.

My deepest personal thanks go to my wife, Li Ying. Without your wholehearted support, unwavering encouragement, and countless sacrifices, I would never have had the strength or courage to pursue this dream. For much of my PhD, I was separated by continents and an eight-hour time difference. Yet you shouldered the burdens of daily life alone, never once asking for anything in return. Your quiet strength and enduring love have been the strongest force carrying me through loneliness and fatigue. You are the quiet hero behind every page of this thesis.

To my beloved daughter, Wenxin—your arrival during a pivotal stage of my PhD brought light and joy into a world filled only with code and data, and reminded me each day that life is so much more than research. You have been my sweetest comfort in difficult times, and my gentlest source of motivation throughout this journey.

I am also deeply grateful to my collaborator, Cheng Sibó. Our shared brainstorming, rigorous discussions, and close collaboration were instrumental to the progress of this work and made the publication of my research possible. The camaraderie I built through this academic journey is something I will always cherish.

Finally, I would like to thank my parents. Your unconditional love and unwavering support have been the lighthouse of my life. Through every high and low, you have stood behind me, offering strength, courage, and peace. There is no way I could ever repay all that you have given me, but I hope this milestone brings you pride and joy.

To all the supervisors, colleagues, friends, and loved ones who have supported me along the way—thank you. This dissertation is not only a record of academic pursuit, but also a testament to your kindness, encouragement, and belief in me. Your generosity has left an indelible mark on my life, and I will carry it with me always.

List of Tables

Table 2.1 PM _{2.5} and NO ₂ monitor stations in London Air Quality Network (2024)	51
Table 2.2 Each pair of co-located Breathe London nodes and London Air Quality Network sites for PM _{2.5} , covering the period from December 2020 to September 2024.....	59
Table 2.3 Each pair of co-located Breathe London nodes and London Air Quality Network sites for NO ₂ , covering the period from December 2020 to September 2024	60
Table 2.4 Environmental Covariates used in sensor calibration	62
Table 2.5 Training dataset for PM _{2.5} and NO ₂ colocation calibration	63
Table 2.6 Testing dataset for PM _{2.5} and NO ₂ from the nearby reference-node pairs	64
Table 2.7 The core hyperparameters and their impacts in random forest	67
Table 2.8 Performance comparison of multiple linear regression, random forest, and transformer models for PM _{2.5} calibration in Breathe London.	72
Table 2.9 Performance comparison of multiple linear regression, random forest, and transformer models for NO ₂ calibration in Breathe London.	72
Table 2.10 Calibration performance under varying numbers of reference co-location sites (5, 10, and 19)	83
Table 2.11 Calibration performance with and without temperature and relative humidity	84

Table 2.12 Calibration performance under different site configurations (roadside-only versus mixed site types).....	85
Table 3.1 Land use categories used in my study	105
Table 3.2 Road type used in my study	106
Table 3.3 Meteorological variables used in my study	113
Table 3.4 The NO ₂ and PM _{2.5} performance of proposed ViT-based model for each configuration (Reference-only, Raw Breathe London + Reference, Calibrated Breathe London + Reference).....	120
Table 3.5 Air quality model performance for PM _{2.5} and NO ₂	125
Table 3.6 Sensitivity of ViT-based NO ₂ and PM _{2.5} predictions to input variables. Only variables with $ \Delta \text{RMSE} \geq 0.1$ are reported.	136
Table 4.1 Uncertainty estimation for NO ₂ and PM _{2.5} using Monte Carlo Dropout in the proposed ViT-Based air quality model	172
Table 4.2 Evaluation metrics for ViT-MCDO and Bayesian Maximum Entropy based on validation reference stations.	174

List of Figures

Figure 2.1 The calibration framework for all low-cost sensors.....	36
Figure 2.2 Transformer architecture	43
Figure 2.3 Scaled Dot-Product Attention and Multi-Head Attention.....	46
Figure 2.4 Network framework of the Transformer model for low-cost sensors calibration.....	47
Figure 2.5 Map of PM _{2.5} monitoring stations in London Air Quality Network in 2024	52
Figure 2.6 Map of NO ₂ monitoring stations in London Air Quality Network in 2024	52
Figure 2.7 Low-cost sensor used in Breathe London network, produced by Clarity	54
Figure 2.8 The map of Breathe London network in 2024	55
Figure 2.9 Locations of Breathe London nodes co-located with reference London Air Quality Network monitors for NO ₂	57
Figure 2.10 Locations of Breathe London nodes co-located with reference London Air Quality Network monitors for PM _{2.5}	58
Figure 2.11 Scatter plots of raw (left) and transformer-calibrated (right) NO ₂ measurements from three selected Breathe London nodes (CLDP0153, CLDP0283, and CLDP0042), compared to London Air Quality Network stations (January 2024)	75
Figure 2.12 Scatter plots of raw (left) and transformer-calibrated (right) PM _{2.5} measurements from three selected Breathe London nodes (CLDP0153, CLDP0020,	

and CLDP0042), compared to London Air Quality Network stations (January 2024)	77
Figure 2.13 Distributions of raw(left) and transformer-calibrated (right) NO ₂ measurements from Breathe London, compared to London Air Quality Network, across the full test dataset (2020 – 2024).....	77
Figure 2.14 Distributions of Raw(left) and transformer-calibrated (right) PM _{2.5} measurements from Breathe London, compared to London Air Quality Network, across the full test dataset (2020 – 2024).....	78
Figure 3.1 The pipeline of proposed ViT-based air quality framework.....	99
Figure 3.2 Overall architecture of the ViT-based air quality model.	101
Figure 3.3 The building coverage in Greater London	107
Figure 3.4 The Green coverage in Greater London	108
Figure 3.5 Google Maps Snapshot of Traffic Congestion in Greater London – 9:00 a.m., February 22, 2024.	109
Figure 3.6 Scatter plots of ViT predicted vs. observed PM _{2.5} concentrations at four randomly selected London Air Quality Network stations: GN3, HV1, MR8, and SK8. The three columns represent different data configurations: (1) Reference-only (left), (2) Raw Breathe London + Reference (middle), (3) Calibrated Breathe London + Reference (right).	122
Figure 3.7 Scatter plots of ViT predicted vs. observed NO ₂ concentrations at four randomly selected London Air Quality Network stations: SK8, BT8, RI1, and GN4. The three columns represent different data configurations: (1) Reference-only (left), (2) Raw Breathe London + Reference (middle), (3) Calibrated Breathe London + Reference (right).	124

Figure 3.8 Estimated NO ₂ maps from the ViT model at 6 AM, 8 AM, 12 PM, 6 PM, 8 PM, and 11 PM on 14 February 2024.....	129
Figure 3.9 Estimated PM _{2.5} maps from the ViT model at 8 AM, 12 PM, 7 PM, 11 PM on 14 February 2024.	132
Figure 4.1 The proposed BME framework for uncertainty estimation	153
Figure 4.2 The proposed framework for LCS uncertainty estimation.....	154
Figure 4.3 The proposed framework for uncertainty estimation of proposed ViT-Based Air Quality Model by Monte Carlo Dropout	158
Figure 4.4 The BME framework for integrating data source with varying uncertainty	165
Figure 4.5 Calibrated Breathe London PM _{2.5} measurements with corresponding prediction interval at 3 random selected stations HP1, GN0, and SK6 over February 2024.....	170
Figure 4.6 Calibrated Breathe London NO ₂ measurements with corresponding prediction interval at 3 random selected stations HP1, GN0, and SK6, February 2024.	171
Figure 4.7 Estimated NO ₂ Concentration and Uncertainty Maps Using Monte Carlo Dropout in the Proposed ViT-Based Air Quality Model at 9 AM, 12 PM, 3 PM, and 6 PM on 14 February 2024 (from top to bottom).	174
Figure 4.8 NO ₂ estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, T=100) VS BME and at 5 PM, February 14, 2024	176
Figure 4.9 Relationship between estimated NO ₂ concentrations and predictive uncertainty from BME at 5 PM, February 14, 2024.....	177

Figure 4.10 Relationship between sensor density and predictive uncertainty from BME (3KM grid, 24-hour average, February 14, 2024, $r=-0.6611$).....	178
Figure 4.11 Map of Greater London depicting the distribution of relative deprivation by neighborhood (LSOA), based on IMD 2019.	183
Figure 4.12 Bivariate maps illustrating the overlay of the Index of Multiple Deprivation with NO ₂ pollution concentration (A) and predictive uncertainty (B) in Greater London at 5 PM on February 14, 2024.	184
Figure 4.13 Relationship between IMD and predictive uncertainty from BME at 5 PM, February 14, 2024	185

Contents

Declaration of Originality	1
Copyright Declaration	2
Abstract	3
Acknowledgements.....	5
List of Tables	7
List of Figures.....	9
Contents	13
Chapter 1: Introduction	17
1.1 Research Motivation	17
1.2 Problem Statement.....	19
1.3 Aims and Objectives of This Research.....	23
1.4 Thesis Overview	24
Chapter 2: Low-Cost Sensor Calibration.....	27
2.1 Literature Review	28
2.2 Methodology and Research Design	34
2.2.1 Multiple Linear Regression Calibration Method.....	36
2.2.2 Random Forest Calibration Method	39
2.2.3 Deep Learning Calibration Method.....	42

2.2.4 Dataset and Preprocessing	49
2.2.5 Model Optimization and Validation Method	65
2.3 Experiments and results	71
2.4 Discussion	78
Chapter 3: Development of Deep Learning Framework for Fine-scale Spatiotemporal Air Quality Modeling based on Low-Cost Sensors	88
3.1 Literature Review	88
3.2 Methodology and Research Design.....	98
3.2.1 Model Architecture.....	99
3.2.2 Dataset and Preprocessing	103
3.2.3 Model Optimization and Validation Method	115
3.2.4 Comparison Models: Land Use Regression, Random Forest, and Convolutional Neural Networks	118
3.3 Experiments and Results.....	120
3.4 Discussion	132
Chapter 4: Uncertainty Analysis in Air Quality Estimation	142
4.1 Literature Review	143
4.2 Methodology and Research Design	152
4.2.1 Quantifying Low-Cost Sensor Uncertainty as Inputs to the Data Fusion Framework.....	153

4.2.2 Quantifying ViT-based Air Quality Model Uncertainty as Inputs to the Data Fusion Framework	156
4.2.3 Bayesian Maximum Entropy Approach for Integrating Data with Varied Uncertainties	161
4.3 Experiments and Results.....	168
4.3.1 Aleatoric Uncertainty in Low-cost Sensor Measurements.....	168
4.3.2 Epistemic Uncertainty in Air Quality Model Estimation.....	172
4.3.3 The Performance of Bayesian Maximum Entropy Framework for Integrating Data with Varied Uncertainties.....	174
4.4 Discussion	179
Chapter 5: Conclusions and Future Research	187
5.1. Summary of Key Findings.....	187
5.2. Contributions to Knowledge.....	191
5.3. Limitations of the Study	194
5.4. Recommendations for Future Research.....	198
References	205
Appendix A. Publications.....	229
Appendix B. Data Collection and Data Preprocessing.....	230
1. Standard Grid Generation	230
2. Breathe London Monitoring Data.....	231

3. Reference Monitoring Station Data	233
4. Land Use Characteristic Data	234
5. Traffic Data	235
6. Meteorological Data	237
Appendix C. Low-Cost Sensor Calibration	240
Appendix D. Air Quality Model	248
Appendix E. Uncertainty Analysis in Air Quality Estimation	254

Chapter 1: Introduction

1.1 Research Motivation

Ambient air pollution remains one of the most pressing environmental health crises of the 21st century (Burnett et al., 2014; Cohen et al., 2005; Madrigano et al., 2013), responsible for approximately 4.2 million premature deaths globally each year (WHO, 2024). Researchers have shown that even modest increases in air pollutant concentrations can result in substantial negative health consequences at the population level (Cohen et al., 2017). Beyond its direct impact on mortality, ambient air pollution is strongly associated with a wide spectrum of acute and chronic diseases, including cardiovascular disorders (Andersen, Kristiansen, et al., 2012; Brook et al., 2010; Goldberg, 2008a), respiratory illnesses (Andersen, Bønnelykke, et al., 2012; Andersen et al., 2011; Gehring et al., 2010), various types of cancer (Goldberg, 2008b; Raaschou-Nielsen et al., 2011), and adverse effects on cognitive functioning and central nervous system health (Bos et al., 2014; Thurston et al., 2017; Volk et al., 2013; X. Zhang et al., 2018). Furthermore, air pollution represents a significant economic burden worldwide, with estimates suggesting that approximately 2–5% of global GDP is spent treating diseases linked to air pollution exposure (Lanzi, 2016). The severity of this crisis is further exacerbated by increasing global urbanization, with over 90% of the world's population now living in regions where air pollutant concentrations exceed recommended safe limits (WHO, 2021).

Given the extensive range of serious health impacts associated with ambient air pollution, accurately measuring and characterizing pollution exposure is essential for designing effective public health interventions. This need for detailed air quality information is driving the development of fine-grained spatiotemporal maps of air pollution. Such detailed maps can improve our understanding of complex air quality patterns, precisely identify pollution hotspots, enable accurate assessment of

exposure at individual or neighborhood levels, and ultimately facilitate targeted and effective health intervention strategies.

Air Pollutants

Urban air pollution comprises various pollutants, primarily resulting from human activities, which present significant challenges for public health and environmental management. Pollutants such as particulate matter (PM) and nitrogen dioxide (NO₂) are well-documented for their adverse effects on human health, posing considerable risks to public health (Hoek et al., 2010; Patel et al., 2010; Pope et al., 2009; WHO, 2006).

Particulate matter (PM) encompasses solid and liquid particles suspended in the air. Regulatory frameworks classify PM by aerodynamic diameter, most commonly into PM₁₀ (particles ≤ 10 μm in diameter) and PM_{2.5} (particles ≤ 2.5 μm). PM originates from diverse sources, both natural and anthropogenic, and can be categorized into primary and secondary types. Primary PM is emitted directly into the atmosphere from sources such as fossil-fuel combustion, industrial processes, dust resuspension, and biomass burning. Secondary PM is formed through atmospheric chemical reactions involving gaseous precursors, including nitrogen oxides (NO_x), sulfur dioxide (SO₂), volatile organic compounds (VOCs), and ammonia (NH₃) (Heal et al., 2012; Seinfeld & Pandis, 2016). Smaller particles (especially PM_{2.5}) can penetrate deeper into the respiratory tract and are associated with more severe health effects (WHO, 2006).

Nitrogen dioxide (NO₂), another major air pollutant, has been widely studied for its prevalence in urban areas and significant health impacts. NO₂ primarily originates from high-temperature combustion processes, such as those occurring in vehicle engines and power plants. These processes initially emit nitric oxide (NO), which undergoes rapid oxidation in the atmosphere to form NO₂ (Seinfeld & Pandis, 2016). Given its widespread sources and health implications, NO₂ is frequently used as a key indicator of urban air quality.

Due to their severe health impacts and frequent exceedances of regulatory guidelines in urban areas globally, this study selects PM_{2.5} and NO₂ as the primary pollutants.

1.2 Problem Statement

Developing reliable, high-resolution spatiotemporal maps of air pollution, particularly within densely populated urban environments, remains challenging due to numerous dynamic factors (e.g., meteorological conditions and traffic patterns) and local spatial factors (e.g., the rapid decay of pollutant concentrations away from sources), which cause substantial fluctuations over short spatial distances and brief time intervals (Apte et al., 2017a; Jacob & Winner, 2009).

Over the past few decades, governments have established extensive air pollution monitoring networks utilizing regulatory-grade monitoring stations. Although these stations provide relatively high accuracy, their significant installation and maintenance costs have limited the density of their deployment. Consequently, the resulting sparse distribution often proves inadequate for capturing fine-scale spatial variations in urban air pollution (Apte et al., 2017b).

To predict air quality in unmonitored areas, model-based techniques have been employed. Generally, there are two types of air quality models: physics-based and data-driven. Physics-based models simulate air pollution diffusion by physical and chemical processes, accounting for emissions from different sources (Appel et al., 2017; Byun & Schere, 2006a; Kwak et al., 2015; Righi et al., 2009). The advantages of bottom-up physics-based models are that they are grounded in fundamental atmospheric principles and transferable across locations and time periods, even in areas with sparse monitoring. Because they rely on physical laws rather than historical correlations, they can represent future or unseen conditions, enabling forecasting and scenario testing. For example, simulating the impacts of changes in emissions, land use, or policy interventions. Their ability to operate across multiple spatial and temporal scales, and their relative independence from observational datasets, explains why physics-based models remain central to regulatory planning

and long-term air quality assessment. However, the complexity of real-world air pollution events often exceeds the empirical assumptions used in physics-based models, leading to degraded performance. Furthermore, they are frequently associated with significant computational expenses, and acquiring sources of emission factors can also be challenging, which limits their practical use for real-time, fine-scale pollution estimation (Jerrett et al., 2005a).

In contrast, data-driven models directly learn patterns from data, which avoids computing the complex numerical model in physics-based methods and achieves a lower computational cost. For example, land-use regression (LUR) modeling is a widely used data-driven approach in epidemiologic research to estimate exposure to air pollution (Beelen et al., 2013; Hoek et al., 2008a; Jerrett et al., 2005b). By integrating both static and dynamic covariates, including land-use types, meteorological data, and traffic data, LUR models can predict air pollutant concentrations effectively (Hoek et al., 2008b; Jain et al., 2021).

However, prior studies indicate that the formation of air pollutants is complex and inherently nonlinear. Traditional LUR models, particularly the classical formulations based on linear regression, struggle to capture these nonlinear dynamics. Although modern extensions of LUR incorporate nonlinear components (e.g., splines or interaction terms), accurately modelling nonlinear relationships through multivariate regression remains a significant challenge, especially when pollutant concentrations depend on interacting meteorological, chemical, and spatial processes (Kwak et al., 2015).

In recent years, machine learning methods have significantly improved in many disciplines, primarily due to their capability to identify and capture complex patterns and relationships hidden within extensive datasets (Y. Cheng et al., 2014; Fukami et al., 2021; He et al., 2022). The dramatic expansion in the availability of large-scale datasets, commonly known as "big data", has been particularly instrumental in this progress, enabling machine learning models to achieve unprecedented predictive performance. Previous studies have demonstrated that machine learning models,

leveraging extensive air quality measurements, can substantially enhance air pollution modeling (Jain et al., 2021a; Zheng et al., 2013; Zimmerman et al., 2020). Complex machine learning architectures, such as artificial neural networks, are particularly effective in capturing intricate relationships among variables when trained on large, diverse, and high-dimensional datasets, highlighting the transformative impact of big data availability in environmental research. However, the sparse distribution of reference-grade monitoring stations limits the data available to these machine learning models, thereby constraining their capabilities.

To increase the spatial resolution of budget-constrained air quality measurements, low-cost sensors have been receiving increasing interest in recent years (Y. Cheng et al., 2014; Do et al., 2020; Giordano et al., 2021a; Jain et al., 2021a). A prominent example is Breathe London, an innovative urban air quality initiative coordinated by Imperial College London's Environmental Research Group in collaboration with local governmental bodies and community stakeholders (Breathe London, 2024). The initiative deploys an extensive network of affordable, low-cost sensors across Greater London to continuously measure key pollutants, including nitrogen dioxide (NO_2) and fine particulate matter ($\text{PM}_{2.5}$). Due to their compact size and relatively inexpensive cost, the sensors can be more easily deployed in urban areas than reference monitoring sites, thus increasing the monitoring coverage area. However, the integration of low-cost sensors within a machine learning framework to develop precise, fine-scale spatiotemporal air quality estimations introduces several new challenges:

1. Low-cost sensors (LCS) are less accurate than reference-grade monitors because their measurements are affected by both random noise and systematic bias. Noise arises from factors such as electronic instability or manufacturing variability. These effects increase measurement variability and reduce repeatability. In addition, LCS often exhibit systematic bias, where readings consistently overestimate or underestimate true concentrations due to imperfect calibration, temperature and humidity sensitivity, or drift over time. This bias is not constant: it can vary with environmental conditions, pollutant levels, or sensor degradation. Since machine

learning methods rely on high-quality data to provide trustworthy results, the large amount of noisy data might adversely affect model performance and introduce significant uncertainty into air quality estimates. Traditional calibration procedures, including laboratory calibration and long-term co-location with reference instruments, are designed to correct these biases, but their applicability in large-scale sensor networks is limited because they are time-consuming and resource-intensive. These limitations motivate the need for scalable, data-driven calibration frameworks that can accommodate dynamic, environment-dependent biases in low-cost sensors.

2. Traditional machine learning models have struggled to fully harness the rich data now available from dense low-cost sensor networks and diverse environmental data streams. These approaches impose simplifying assumptions and require manual intervention that limit their ability to capture fine-scale air quality patterns. For instance, many classical models implicitly assume spatial independence – treating measurements at different locations as unrelated – even though pollutant concentrations exhibit spatial dependency. Additionally, traditional models (e.g., land use regression models) rely on researchers to manually select and integrate heterogeneous data sources, such as meteorological conditions, land use, and traffic patterns. This process is often labor-intensive and tends to yield suboptimal results, particularly in complex urban environments. These shortcomings highlight the need for more advanced machine learning models (e.g. advanced deep learning models) that can better leverage high-density, multi-source datasets to achieve high spatiotemporal resolution and accuracy in urban air quality estimation.

3. Most existing machine learning approaches either lack explicit uncertainty quantification or fail to systematically incorporate measurement uncertainties into their predictive frameworks. One critical dimension of uncertainty quantification is the effective integration of data sources with varying levels of measurement accuracy. For example, addressing the inherently higher uncertainty in measurements obtained from low-cost sensors compared to reference-grade monitors. For decision makers, uncertainty must be communicated explicitly and transparently, typically through probabilistic outputs such as prediction intervals or

uncertainty bands over spatial maps. These representations help practitioners understand not only the estimated pollution levels but also the confidence associated with those estimates, enabling risk-aware decisions. Given that accurately quantifying uncertainty is essential for transparently communicating the reliability of air quality estimation and supporting informed decision-making, there is an urgent need to develop comprehensive modeling frameworks capable of:

- Quantifying data uncertainty associated with measurements from low-cost sensors, as well as the inherent uncertainty of the proposed air quality model.
- Systematically integrating data with varying levels of uncertainty, along with model uncertainty, provides transparent uncertainty quantification.

1.3 Aims and Objectives of This Research

The overarching aim of this research is to develop a fine-scale spatiotemporal air quality modeling framework that integrates large but noisy datasets from low-cost sensors along with other heterogeneous data sources. This framework is intended to improve urban air pollution estimation and support evidence-based environmental decision-making. To achieve this goal and address these challenges outlined in Section 1.2, the dissertation is structured around the following specific objectives:

1. To develop a deep learning calibration framework for low-cost sensor network, enabling their measurements to be reliably used for high-resolution air quality modeling.

2. To develop an advanced deep learning model that can integrate high-density, multi-source datasets (e.g., sensor readings, land use, traffic, and meteorological data) and estimate air pollutant concentrations with fine spatiotemporal granularity.

3. To develop a comprehensive uncertainty quantification framework capable of characterizing and integrating uncertainties from both sensor

measurements and model predictions, thereby enhancing the transparency and robustness of air quality estimates.

This research uses Greater London as a case study for implementing and evaluating the proposed air quality modeling framework. London is an ideal area for this research due to its extensive low-cost sensor networks, high-quality reference-grade monitoring infrastructure, and the availability of rich auxiliary data (GLA, 2021). Despite recent improvements in air quality, pollutants such as fine particulate matter (PM_{2.5}) and nitrogen dioxide (NO₂) were still responsible for an estimated 3,600 to 4,100 premature deaths in London in 2019 (Health Burden of Air Pollution in London | London City Hall, 2021). Furthermore, certain regions within London continue to exceed national air quality standards (GLA, 2020a). The methodological innovations developed in this research are not limited to the London context. They are broadly applicable and can be adapted to other urban areas facing similar challenges in air pollution monitoring and management, as discussed further in Chapter 5.

1.4 Thesis Overview

The thesis follows a structured methodological pipeline composed of several interconnected stages, each elaborated in a dedicated chapter. Specifically, the research progresses through three main components: sensor calibration, advanced air quality modeling, and comprehensive uncertainty quantification, systematically building towards robust and actionable air quality information.

Each chapter begins with a detailed literature review, explicitly outlining the current state-of-the-art methods and highlighting the key research gaps that motivate the development of novel approaches introduced in this study. Following each literature review, the chapters present the specific methods developed to address identified gaps, clearly explaining their theoretical foundations, implementation details, and practical implications.

Chapter 2 focuses on the calibration of low-cost air quality sensors. Recognizing the limitations of traditional calibration methods in terms of scalability and practicality, this chapter introduces a robust and scalable calibration framework specifically designed for urban-scale low-cost sensor networks. The calibration method developed in this chapter forms the foundational step, ensuring high-quality input data for subsequent air quality modeling stages.

Building upon the calibrated sensor data from Chapter 2, **Chapter 3** presents an advanced deep learning framework for modeling fine-scale spatiotemporal variations in urban air pollution. The chapter first reviews traditional air quality modeling approaches and their limitations, and then introduces a novel Vision Transformer (ViT)-based method to predict key pollutants, specifically NO_2 and $\text{PM}_{2.5}$, at high spatial and temporal resolutions. This methodological advancement overcomes previous challenges in fully leveraging rich, yet noisy, data from low-cost sensor networks.

To enhance transparency and reliability in air quality estimation, **Chapter 4** develops comprehensive uncertainty quantification methods. This chapter first reviews existing approaches to uncertainty assessment, identifying critical gaps in handling multiple uncertainty sources simultaneously. In response, it quantifies uncertainties arising from low-cost sensors and air quality models, and subsequently introduces a Bayesian Maximum Entropy (BME)-based data fusion framework explicitly designed to integrate uncertainties from sensor measurements, air quality model predictions, and reference monitoring data. This framework provides essential uncertainty metrics for air quality estimates, enhancing their interpretability and practical utility.

Finally, **Chapter 5** synthesizes and discusses the overall findings, explicitly outlines the contributions of this thesis to academic knowledge and urban air quality management practice, identifies methodological and practical limitations encountered during the research, and recommends targeted directions for future investigation. Readers can thus clearly understand how each chapter's

methodological advancements build upon each other, forming a cohesive approach to addressing the complex challenge of fine-scale air quality mapping.

Chapter 2: Low-Cost Sensor Calibration

Low-cost sensors pose several challenges related to data quality, such as sensor noise, cross-sensitivity, stability issues, and the influence of environmental conditions on their measurements (Masson et al., 2015a; Moltchanov et al., 2015; Pang et al., 2017; R. Williams et al., 2014). Many of these challenges arise from the fundamental sensing mechanisms used in low-cost devices. Electrochemical sensors, for example, measure pollutant concentrations by converting gas–electrode reactions into an electrical signal. Their outputs depend on diffusion rates and reaction kinetics, both of which are strongly influenced by temperature, relative humidity, and interfering gases. Masson et al. (2015a) demonstrated that relative humidity (RH) fluctuations substantially impact sensor readings, possibly due to condensation on potentiostat electronics. Furthermore, many electrochemical sensors have a limited operational lifetime (typically 1–2 years) due to gradual reagent consumption, leading to gradual drift and loss of accuracy, which adds significant complexity to calibration processes. Reference-grade monitors operate using well-established physical or chemical measurement principles with rigorous quality assurance. These instruments maintain high accuracy, stability, serve as the “ground truth” for calibrating LCS networks (Clements et al., 2022).

Calibration refers to the process of estimating a functional relationship that maps LCS outputs to the “true” pollutant concentrations measured by a reference-grade instrument, thereby reducing systematic bias in LCS measurements. In this study, I investigated several calibration models for the Breathe London network, a city-wide initiative measuring PM_{2.5} and NO₂ concentrations using low-cost sensors across Greater London. The objective is to identify calibration methods capable of addressing the aforementioned challenges and improve the accuracy and consistency of low-cost sensor measurements, thereby enabling their integration into high-resolution urban air quality models. It is important to note that even after calibration, residual uncertainty remains due to random noise and environmental dependencies. These uncertainties are explicitly characterised in Chapter 4.

2.1 Literature Review

One of the traditional methods for calibrating air quality sensors involves manual maintenance and laboratory-based calibration. This process typically includes cleaning the sensor, replacing degraded components, and periodically recalibrating the device in controlled laboratory settings against certified reference instruments. While this approach can be effective, it is both time-consuming and labor-intensive, limiting its scalability for large sensor networks (Concas et al., 2021a).

A widely adopted alternative is the co-location calibration method, where low-cost sensors are deployed adjacent to a reference monitoring station for a certain calibration period. During co-location, low-cost sensors record measurements in parallel with the reference instrument, and a calibration model is developed to align sensor readings (often alongside environmental covariates) with reference pollutant concentrations (Nalakurthi et al., 2024). Previous studies show that incorporating local environmental factors (e.g. temperature, humidity) into calibration models substantially improves accuracy (Kang & Choi, 2024). For instance, particulate matter sensors may report artificially high concentrations under high humidity due to hygroscopic particle growth, necessitating the inclusion of humidity correction functions (Zimmerman et al., 2017). In co-location calibration, a distinct calibration function is typically built for each sensor unit based on its localized training data.

Simple statistical models, particularly multiple linear regression (MLR), have been widely used to calibrate low-cost sensor data in co-location studies (Jain & Zimmerman, 2024). MLR is popular as a baseline due to its interpretability, computational efficiency, and ease of implementation. They assume that the reference value is linearly related to the low-cost sensor measurements and environmental covariates. Despite its simplicity and convenience, linear models come with notable limitations. They implicitly assume the data are independent and identically distributed and that relationships remain linear, assumptions that often do not hold in complex, real-world air pollution scenarios (Giordano et al., 2021b). Air quality data can exhibit complex, non-linear behavior and cross-sensitivity (e.g., a

metal-oxide gas sensor's response to NO_2 can be affected by the presence of O_3). Additionally, the performance of an MLR calibration may vary between sensors due to manufacturing variances and can degrade if the co-location period is too short to capture the full range of environmental conditions and pollutant levels. As a result, while MLR calibration serves as a useful benchmark, more advanced calibration techniques are essential for ensuring reliable performance across diverse environmental conditions (Castell et al., 2017a).

In recent years, researchers have increasingly adopted machine learning (ML) algorithms to enhance calibration of low-cost air quality sensors. ML methods excel at capturing non-linear and complex relationships between sensor readings, environmental factors, and reference pollutant concentrations, often outperforming traditional linear models (Abu-Hani et al., 2024). Among these methods, random forest (RF), a tree-based ensemble learning technique, has been widely applied. Random forest models have been applied to calibrate sensors for various pollutants, including carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), and particulate matter ($\text{PM}_{2.5}$, PM_{10}), demonstrating substantially improved accuracy over uncalibrated data and often surpassing linear models (Cordero et al., 2018; Zimmerman et al., 2018a). For instance, Zimmerman et al. (2018) showed that a random forest calibration model for low-cost sensors matched or exceeded the performance of a site-specific linear regression model in predicting particulate matter concentrations. Key advantages of RF include:

- **Ability to Model Non-Linear Relationships:** RF can naturally learn intricate, non-linear interactions among features and the target variable without requiring explicit feature transformations. This makes it well-suited for environmental sensor calibration, where pollutant concentrations may relate to covariates (temperature, humidity, etc.) in a non-linear way.
- **Robustness to Noise:** By averaging many trees trained on bootstrap-resampled data, RF dilutes the effect of noisy instances. Each tree is trained on a slightly different dataset and subset of features, so random noise in the

training data is less likely to consistently influence all trees, improving the model's robustness.

- **Reduced Overfitting:** Individual decision trees tend to overfit, but an ensemble of de-correlated trees reduces variance. The RF's use of bagging (bootstrap aggregating) and random feature selection means the final prediction is an average of many trees' outputs, which by the law of large numbers stabilizes the prediction and mitigates overfitting tendencies.

Despite these strengths, RF models have notable limitations. One well-documented issue is their poor ability to extrapolate beyond the range of training data (Jain et al., 2021). In practice, if a sensor subsequently encounters environmental conditions (e.g., temperature, humidity), emission patterns, or atmospheric regimes that fall outside the distributions observed during training, the RF model cannot reliably infer relationships beyond its learned feature space (Abu-Hani et al., 2024). Furthermore, RF does not capture temporal dependencies, despite the fact that air quality data often exhibit strong temporal correlations. Temporal structure is highly relevant to calibration because sensor biases, drift, and environmental influences typically evolve over time rather than occurring independently at each measurement. Calibration models that explicitly incorporate temporal dependency can therefore better distinguish true atmospheric variability from sensor-induced artefacts, resulting in more accurate and stable calibrated outputs. This limitation underscores the need for alternative modeling approaches that can explicitly incorporate temporal structure, such as recurrent neural networks (RNNs) or attention-based deep learning architectures.

While many calibration models use “traditional” machine learning (e.g. decision trees, random forests), there is a growing shift towards **deep learning** methods to capture complex, non-linear relationships inherent in sensor data (Abu-Hani et al., 2024). Deep learning offers flexible function approximation that can potentially achieve better accuracy (LeCun et al., 2015; Krizhevsky et al., 2017; Hadash et al., 2018). Researchers have explored neural networks ranging from simple multilayer

perceptrons to more advanced architectures (He et al., 2016; Vaswani et al., 2017). Neural networks can learn complex functional mappings and have been employed in several calibration studies (Y. Cheng et al., 2014; Cordero et al., 2018; Esposito et al., 2016). For instance, Cheng et al. (2014) calibrated low-cost PM_{2.5} and CO₂ sensors using a multilayer perceptron artificial neural network (ANN) achieving higher accuracy compared to linear baseline methods.

Early deep learning-based calibration models typically operated on feature vectors measured at a single time step, producing a calibrated output for that same time step, commonly referred to as one-to-one calibration. More recent approaches have advanced toward many-to-one calibration, in which feature sequences spanning previous time steps are used as input to predict the current calibrated value. This formulation better captures both the temporal variability of pollutant concentrations and the dynamic response characteristics of sensor materials. Studies adopting this paradigm have reported state-of-the-art performance (Ali et al., 2024). For example, Ali et al. (2024) implemented a one-dimensional convolutional neural network (1D-CNN) to calibrate a low-cost CO sensor and reported improved performance over conventional calibration techniques.

Incorporating historical sensor data can therefore enhance calibration accuracy, as consecutive data contain vital information regarding sensor dynamics and environmental fluctuations. While traditional methods, such as linear regression and Random Forest, have occasionally included recent measurements as input features, these models treat each measurement independently and lack the capacity to learn from sequential dependencies. This inherent limitation motivates adopting sequence modeling techniques, capable of effectively utilizing temporal context.

To address this need, recent studies have explored deep learning architectures designed to capture temporal dependencies. Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been employed to model temporal dependencies, as seen in Ali et al. (2024) and Bachechi et al. (2024). By processing input sequences, LSTM-based models could learn from time-

sequenced data to correct drift and time-varying biases. For example, Yu et al. (2020) proposed a sequence-to-point calibration model called AirNet, which uses a dual-sequence encoder network to integrate historical time-series data from both co-located low-cost sensor and reference station. By providing the network with temporal patterns, AirNet significantly outperformed methods that ignore sequence structure.

While RNNs like LSTM are well-suited for sequence learning, Transformer models, a more recent class of deep neural network, have proven exceptionally effective in capturing long-range temporal dependencies. Unlike RNNs, Transformers use attention mechanisms to model dependencies across time steps, regardless of their distance in the sequence. This parallel processing capability makes Transformers not only more computationally efficient but also more effective in learning temporal structures for sensor calibration tasks (Vaswani et al., 2017). The technical implementation and adaptation of Transformer-based calibration in this study are detailed in Section 2.2.3.

Limitations of Existing Models:

Scalability Challenges: While sophisticated calibration models can significantly enhance the accuracy of individual sensors, scaling such solutions across large networks remains a fundamental challenge. Most existing studies adopt localized calibration methods, where a separate calibration model is developed for each sensor after co-location with reference instrument (Zimmerman et al., 2018; Spinelle et al., 2015). Although effective on a small scale, this method requires extensive labor and time, rendering it impractical when calibrating hundreds of sensors across a city. Moreover, calibration models trained under specific environmental conditions often fail to generalize effectively when applied across diverse environments (Aula et al., 2022). Thus, individually calibrating every sensor across an entire city for extended durations may be impractical. The cost and logistical complexity of maintaining calibration across large-scale networks constitute a critical barrier to fully leveraging

low-cost sensors. In short, traditional sensor-by-sensor calibration does not scale effectively.

Research Gap and Contribution:

To overcome these limitations, recent research has begun to explore more efficient calibration frameworks. A particularly promising direction involves the development of global calibration models, which leverage data from multiple sensors (typically of the same make and model) across diverse environmental settings. Rather than fitting individual models for each sensor, global models aim to generalize across devices and locations, requiring minimal retraining to apply to newly deployed sensors (Abu-Hani et al., 2024; Vikram et al., 2019). These models have shown promising results in test environments, demonstrating improved performance when trained on multi-location datasets. However, current studies typically involve only small-scale experiments with a limited number of sensors (from a few to a dozen).

In addition, existing calibration approaches face several fundamental challenges that extend beyond nonlinear relationships or temporal variability. Sensor behaviour is highly environment-dependent, meaning the same pollutant concentration can yield different raw signals under varying humidity or temperature conditions. Calibration models that work well in one location may not transfer to another due to local emission characteristics, meteorological regimes, or sensor-to-sensor variability. Furthermore, drift arising from sensor ageing leads to gradual degradation in measurement accuracy, requiring recalibration at unknown and inconsistent intervals (Vajs et al., 2023b). These factors make it difficult to maintain stable measurement performance in long-term and wide-area sensor deployments, especially when hundreds of sensors operate under heterogeneous conditions.

Given these limitations, there is a clear research gap: the absence of a scalable, foundational calibration framework that can be deployed across dense low-cost sensor networks while continuously adapting to time-varying sensor behaviours and

diverse environmental conditions. To address this gap, I propose a novel deep-learning-based calibration framework designed to scale efficiently across large sensor networks and dynamically model nonlinear, environment-dependent, and temporally evolving error structures. This advancement aims to enable more robust and reliable large-scale air quality monitoring networks.

2.2 Methodology and Research Design

This section outlines the methodology employed for calibrating low-cost sensors (LCS), building upon approaches similar to those described by Abu-Hani et al. (2024) and Vikram et al. (2019). Specifically, I construct a global calibration model using a subset of sensors from the overall network and subsequently apply this model to calibrate the remaining sensors. To extend previous research, I introduce a comprehensive and robust experimental design that systematically evaluates network correction strategies:

Calibration Strategies

1. **Co-location Calibration:** Rather than developing individual calibration models for each sensor, I train a global model from a set of co-located sensors, which is then applied to the remaining sensor network.
2. **Environmental Corrections:** Environmental covariates, including site classification (e.g., roadside, urban background), meteorological variables (humidity and temperature), and temporal factors (hour of day), are incorporated to enhance calibration accuracy by accounting for external influences.
3. **Temporal Dependency Calibration:** Recognizing that historical sensor measurements contain valuable temporal information, I train calibration models using data from the preceding 24-hour period. The 24-hour window captures the full diurnal variability in air pollution levels, enabling the model to learn hour-of-day effects. Beyond diurnal patterns, this window also reflects other forms of temporal variability, such as changes driven by

evolving meteorological conditions (e.g., temperature, humidity) and shifts in emission activity (morning and evening traffic peaks). These factors directly influence sensor performance because temperature and humidity affect electrochemical reaction rates, cross-sensitivity behaviour, and baseline stability, while emission-driven concentration changes may interact with sensor non-linearities or limit-of-detection issues. By incorporating these time-dependent influences, the calibration model can better distinguish true atmospheric variability from sensor-induced artefacts, leading to more stable and accurate corrections. At the same time, the window is short enough to allow the model to quickly adapt to evolving conditions. This approach leverages temporal dependencies and offers more accurate sensor corrections compared to treating each measurement independently. This method aligns closely with the Dynamic Network Correction Algorithm utilized by the Breathe London project (Breathe London, 2024).

4. **Periodic Calibration:** Due to sensor drift over time, often caused by factors such as dust accumulation and LED emitter degradation, calibration must be regularly updated. To address this, I implement rolling calibration windows (e.g., hourly, daily, weekly, monthly) to periodically recalibrate sensors. Specifically, calibration models are trained on recent data windows (e.g., hourly models are trained using data from the hour; daily models utilize the day's measurements). Each updated model then recalibrates sensor readings collected during the corresponding time frame. Frequent recalibration effectively mitigates drift by adjusting for recent environmental changes and sensor behavior, thereby sustaining data accuracy.

The previous four strategies are then subsequently integrated into the following calibration methods:

Calibration Models

1. **Machine Learning-Based Calibration:** Two traditional machine learning models: linear regression and random forest were examined in the

experiments. Sections 2.2.1 and 2.2.2 applied linear regression and random forest models respectively.

2. **Deep-learning Based Calibration:** I proposed an advanced, transformer-based deep learning calibration model (detailed in Section 2.2.3). This novel method efficiently captures complex temporal dependencies and nonlinear interactions, further improving calibration accuracy beyond classical machine learning approaches.

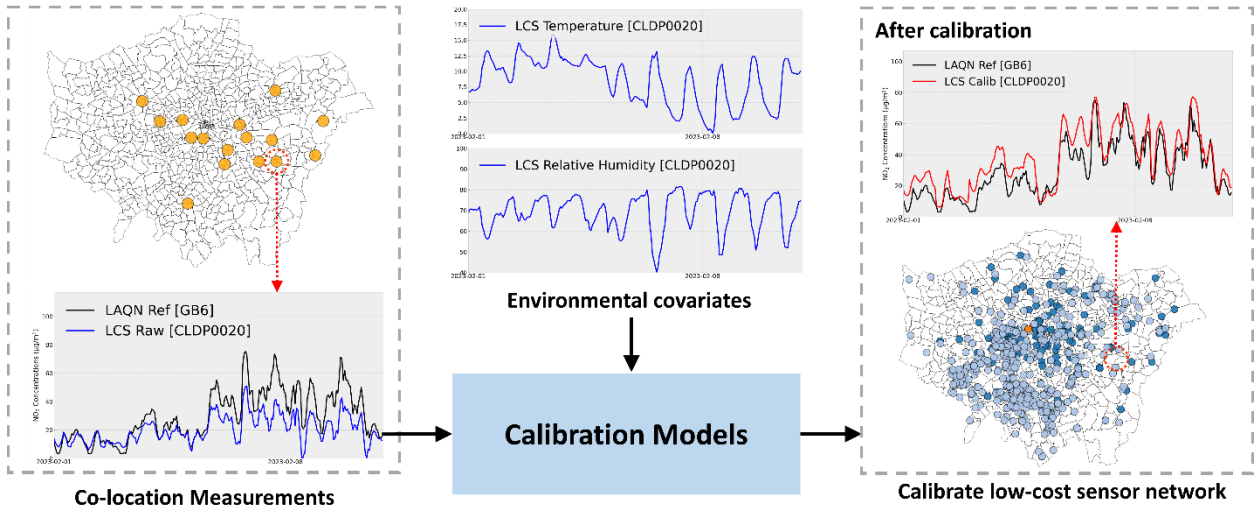


Figure 2.1 The calibration framework for all low-cost sensors

In summary, our methodology enhances the calibration of large-scale low-cost air quality sensor network by integrating global modeling, environmental and temporal feature integration, periodic recalibration, and advanced machine learning techniques. Each enhancement aims to improve accuracy and reliability of low-cost sensors. The following Sections (2.2–2.4) detail the implementation and performance of the linear regression, random forest, and deep learning calibration models, respectively, under this framework.

2.2.1 Multiple Linear Regression Calibration Method

Model Building

Multiple linear regression (MLR) serves as a conventional yet widely adopted approach for calibrating sensor data. Although it may not reach the accuracy levels of more advanced models such as random forests or neural networks, MLR provides several benefits: simplicity, ease of deployment, and interpretability. These characteristics make it a strong baseline and a practical option for routine sensor calibration in large networks. Accordingly, I implemented multiple linear regression (MLR) models to calibrate low-cost sensor measurements (PM_{2.5} and NO₂) against reference data.

Following the 4-step calibration strategies, the co-located measurements (introduced in Section 2.3) are used to train the linear calibration model, which uses the raw pollutant concentration data, which have been quality-controlled and formatted for analysis, as the primary independent variable and includes additional terms to adjust for influencing factors. In particular, I incorporate temperature and humidity as independent variables to account for environmental effects on sensor readings. I also add variables for the hour of day and for the type of sensor's location (e.g., residential, roadside) to capture temporal pollution cycles and site-specific differences, together with ambient temperature and relative humidity to account for meteorological influences on pollutant formation and sensor response. The general form of multiple linear regression model is given by the linear equation:

$$y = X\beta + \varepsilon \quad (2.1)$$

where y is the vector of true pollutant concentrations (from the reference monitor), X is the design matrix of predictors (including a column of ones for the intercept), β is the coefficient vector, and ε is the error term. For example, the full calibration model for PM_{2.5} can be expressed as:

$$\begin{aligned}
PM_{2.5,t} = & \beta_0 + \sum_{l=0}^{23} \beta_{1,l} \cdot (\text{SensorPM}_{2.5,t-l}) \\
& + \sum_{l=0}^{23} \beta_{2,l} \cdot (\text{Temperature}_{t-l}) + \sum_{l=0}^{23} \beta_{3,l} \cdot (\text{Humidity}_{t-l}) \\
& + \sum_{h=0}^{23} \beta_{4,h} \cdot I[\text{Hour}_t = h] + \sum_{k=1}^{K-1} \beta_{5,k} \cdot I[\text{SiteType} = k] + \epsilon_t
\end{aligned} \tag{2.2}$$

The definitions of the parameters in the equation are as follows:

- $PM_{2.5,t}$: The predicted $PM_{2.5}$ concentration at time t .
- $\text{SensorPM}_{2.5,t}$: The $PM_{2.5}$ concentration reported by the low-cost sensor at time t .
- Humidity_t : The measured humidity at time t .
- Temperature_t : The measured temperature at time t .
- Hour_t : Hour of the day at time t .
- SiteType : The site type of BL node.
- β : Coefficients to be estimated in the regression model.
- ϵ_t : The error term, assumed to follow a normal distribution.
- l : The temporal lag index (0–23), representing how many hours prior to time t the variable is measured.

Similar to $PM_{2.5}$, the full calibration model for NO_2 can be expressed as:

$$\begin{aligned}
NO_{2,t} = & \beta_0 + \sum_{l=0}^{23} \beta_{1,l} \cdot (\text{SensorNO}_{2,t-l}) + \sum_{l=0}^{23} \beta_{2,l} \cdot (\text{Temperature}_{t-l}) \\
& + \sum_{l=0}^{23} \beta_{3,l} \cdot (\text{Humidity}_{t-l}) + \sum_{h=0}^{23} \beta_{4,h} \cdot I[\text{Hour}_t = h] \\
& + \sum_{k=1}^{K-1} \beta_{5,k} \cdot I[\text{SiteType} = k] + \epsilon_t
\end{aligned} \tag{2.3}$$

Additionally, low-cost sensors tend to drift over time, necessitating periodic recalibration. To address this, I employ a rolling calibration strategy, where models are retrained on rolling time windows (e.g., hourly, daily, or monthly) and then applied to recalibrate all sensors within the corresponding time interval. The low computational cost of linear regression enables frequent model updates (even on an hourly basis) to counteract drift without heavy resource demands.

2.2.2 Random Forest Calibration Method

Model Building

Random Forest (RF) is a powerful ensemble learning method that constructs multiple decision trees and aggregates their predictions to enhance accuracy and reduce overfitting (Breiman, 2001). It is particularly effective at capturing complex nonlinear relationships that linear models often fail to represent adequately (Zimmerman et al., 2018). In this study, I explore the use of RF for calibrating low-cost $PM_{2.5}$ and NO_2 sensors.

RF Algorithm Description: An RF model for regression is essentially an ensemble of K regression trees $f_1(x), f_2(x), \dots, f_K(x)$. The overall RF prediction $\hat{y}$ for a given input x is the average of the individual tree predictions:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K f_k(x) \quad (2.4)$$

Each tree $f_k(x)$ is grown using the CART (Classification and Regression Tree) algorithm with two key randomization steps: bootstrap sampling of the data and random feature selection at each split. The RF training procedure can be summarized as follows:

1. **Bootstrap Sampling:** From the full training dataset of size N , construct K new training sets $B_1, B_2, \dots, B_K$ by sampling N observations with replacement. Each tree f_k is trained on its bootstrap sample B_k . (On average, each B_k contains about 63% of the unique training instances due to replacements, with some repeated observations.) This bagging procedure introduces diversity among the trees.
2. **Random Feature Splits:** Grow each decision tree by recursively splitting nodes. At each node, instead of considering all M available features for the best split, RF selects a random subset of m features (with $m \ll M$, a hyperparameter often called `max_features`). Within this subset, the algorithm finds the optimal split point based on an impurity criterion. For a regression tree, the impurity can be quantified by the variance (or equivalently the mean squared error) of the responses in a node. Formally, for a node containing a subset of training data D , the impurity (variance) is:

$$I(D) = \frac{1}{|D|} \sum_{(x_i, y_i) \in D} (y_i - \bar{y}_D)^2 \quad (2.5)$$

where $\overline{y_D}$ is the mean of the y values. The algorithm evaluates candidate splits (feature j and threshold t) that partition D into left D_L and right D_R subsets. The split quality is measured by the reduction in impurity:

$$\Delta I = I(D) - \frac{|D_L|}{|D|} I(D_L) - \frac{|D_R|}{|D|} I(D_R) \quad (2.6)$$

The chosen split is the one that maximizes ΔI , i.e. yields the largest drop in variance. This is equivalent to minimizing the weighted variance within the child nodes. By restricting splits to a random subset of features at each node, RF further de-correlates the trees, since each tree must consider different feature candidates and thus can develop a different structure even if the data were the same.

3. **Tree Growth and Stopping:** Each decision tree in the forest is fully grown without pruning to maximize its fit on the bootstrap sample. Splitting continues until a stopping criterion is met. I specify a minimum node size of 5 observations for terminal leaves (i.e., each leaf node must contain at least 5 training samples) to prevent overly small, noisy leaves. This acts as a depth constraint: it implicitly limits how deep the tree can grow by requiring a minimum number of data points to continue splitting. Growing trees to maximum depth (subject to the minimum leaf size) helps each tree capture fine-grained patterns; the subsequent averaging of many such trees will then smooth out idiosyncratic behavior, balancing the bias-variance trade-off.
4. **Ensemble Prediction:** Once all K trees are grown, the RF makes predictions on new data by averaging the predictions from all individual trees. For a new input x_{new} , each tree yields a prediction $f_k(x_{\text{new}})$. The RF output is $\hat{y}(x_{\text{new}}) = \frac{1}{K} \sum_{k=1}^K f_k(x_{\text{new}})$. This aggregation significantly reduces the prediction variance of the model. The ensemble's theoretical foundation can be understood through the bias-variance decomposition: bagging and random feature selection reduce the correlation between trees and hence shrink the variance of the average

prediction, while each deep tree has low bias. Consequently, the RF achieves a lower generalization error than any individual tree, as supported by Breiman’s original analysis (Breiman, 2001).

In preparing the data for the RF calibration model, I include the same features used in the linear regression (sensor raw readings and environmental covariates like temperature, humidity, time-of-day, and site classification). Unlike the linear model, the RF inherently accounts for complex interactions (for example, it can split on humidity at one branch of the tree and on temperature in another branch, capturing a humidity-temperature interaction effect on sensor readings) without any explicit feature engineering.

2.2.3 Deep Learning Calibration Method

In this study, I propose a Transformer-based deep learning calibration method specifically designed for low-cost sensor data. To the best of my knowledge, this is the first application of Transformer architecture to the domain of low-cost sensor calibration. Unlike recurrence-dependent approaches, Transformer utilizes attention mechanisms, enabling parallel processing of sequence elements and directly modeling dependencies regardless of their temporal distance (relationships that remained uncaptured by previous methods), thereby improving calibration accuracy.

Transformer Architecture

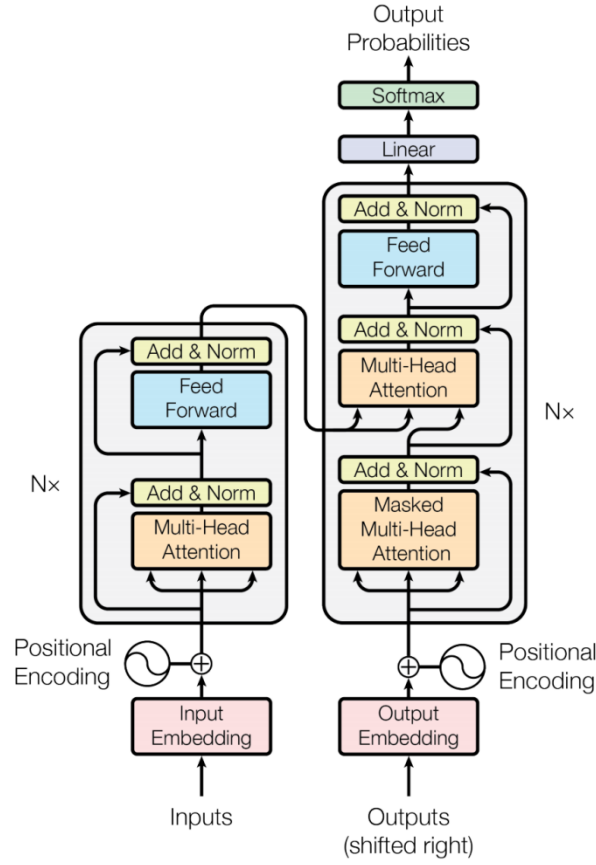


Figure 2.2 Transformer architecture

Transformer has an encoder-decoder architecture as depicted in Figure 2.2. The encoder processes an input sequence of symbol representation $(x_1, x_2, \dots, x_n)$ with time steps n into a continuous latent sequence $(z_1, z_2, \dots, z_n)$. Using the latent sequence, the decoder generates an output sequence $(y_1, y_2, \dots, y_n)$ incrementally, one symbol at a time. The decoder operates auto-regressively, using previously generated symbols as context to predict the next element in the sequence.

Encoder: The encoder of Transformer consists of a stack of N identical layers, I refer to N as a hyperparameter to control the number of encoder layers. Each layer has two sub-modules: (1) Multi-head attention layer: The input sequence is processed by a multi-head attention module (detailed later), which generates a transformed sequence of identical length. This mechanism enables each time step to

integrate contextual information from all positions in the sequence. (2) Position-wise feed-forward layer (FFN): Each time step then passes through an independent FFN (a small multilayer perceptron with nonlinear activation function), specifically the Rectified Linear Unit (ReLU), defined as $\text{ReLU}(x) = \max(0, x)$, to introduce nonlinearity into the model. The inclusion of ReLU enables the network to learn complex, nonlinear relationships between input features, rather than being restricted to linear transformations. After each sub-module, a residual connection adds the original input to the sub-module's output, with layer normalization applied to the result afterward. This design aids gradient propagation in deeper networks and prevents inter-layer distribution shifts.

Decoder: The decoder of Transformer mirrors the encoder with a N -layer stack, and includes identical operations: multi-head attention, FFN and residual connection. It adds a new sub-module, the encoder-decoder attention, which uses multi-head attention on the encoder's output to link encoder and decoder. This new sub-module lets the decoder focus on relevant encoder outputs during generation. It aligns generated content with input semantics via attention weights, while also enhancing long-range dependency modeling.

Scaled Dot-Product Attention: At the core of the Transformer is the scaled dot-product attention (Figure 2.3), which enables the model to learn weighted relationships among different time steps in the input sequence. Specifically, within each scaled dot-product attention layer, every input element can attend to (meaning selectively incorporate information from) all other element. This allows the model to dynamically emphasize on relevant historical readings when calibrating the current sensor reading.

Mathematically, the scaled dot-product attention operation maps a query and a set of key-value pairs to an output. In context, queries (Q), keys (K), and values (V) are all learned projections of the input sequence Z_{k-1} (the output of the previous layer, or the embeddings for the first layer). I compute Q , K , and V as:

$$\begin{aligned}
\text{Query: } Q &= Z_{k-1}W^q \\
\text{Key: } K &= Z_{k-1}W^k \\
\text{Value: } V &= Z_{k-1}W^v
\end{aligned} \tag{2.7}$$

where $W^q, W^k \in \mathbb{R}^{d_{in} \times d_k}, W^v \in \mathbb{R}^{d_{in} \times d_v}$ are learnable projection matrices (here d_{in} is the input dimension to this layer, and d_k is the dimension of Q and K , d_v is the dimension of V). The scaled dot-product attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.8}$$

The scaling factor $\frac{1}{\sqrt{d_k}}$ normalized the dot-product, preventing the dot-product from growing too large when dealing with high-dimensional Q and K . This avoids pushing the softmax into regions with extremely small gradients.

Multi-Head Attention: In practice, the Transformer uses multi-head attention to capture complex temporal dependencies in sequential data. Instead of relying a single attention function, Transformer employs h parallel heads, where h is treated as a tunable hyperparameter. Each head independently learns distinct projections matrices (W_i^q, W_i^k, W_i^v) , generates corresponding Q^i, K^i and V^i and computes head-specific attention head _{i} as described earlier. The attention of all heads are then concatenated and linearly transformed to produce the final output of the multi-head attention layer, as illustrated in Figure 2.4.

$$\begin{aligned}
\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o, \\
\text{where head}_i &= \text{Attention}(Q^i, K^i, V^i)
\end{aligned} \tag{2.9}$$

Where $W^O \in R^{hd_v \times d_{in}}$ is a learned projection matrix for the concatenated heads. Multi-head attention enables the Transformer to jointly attend to information across different representation subspaces. In other words, each head can focus on distinct patterns or time scales within the sequence. For example, some heads may learn short-term dependencies by strongly weighting recent readings, while other heads capture longer-term effects by attending to readings from earlier time steps that reflect diurnal trends. By combining these patterns, the Transformer captures a richer representation of the sensor's temporal behavior than a single attention mechanism could achieve.

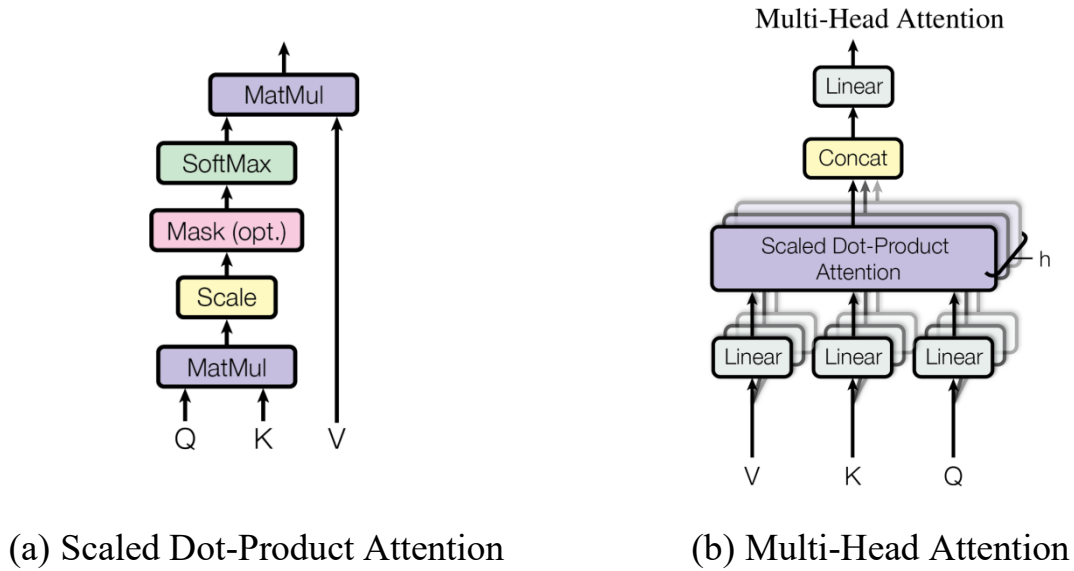


Figure 2.3 Scaled Dot-Product Attention and Multi-Head Attention

Proposed Transformer-Based Calibration Method

As illustrated in Figure 2.4, the proposed Transformer-based model takes as input a sequence of raw sensor readings and their associated covariates collected over a time window $[t - T, t]$, and outputs a calibrated sensor reading for the current time step t .

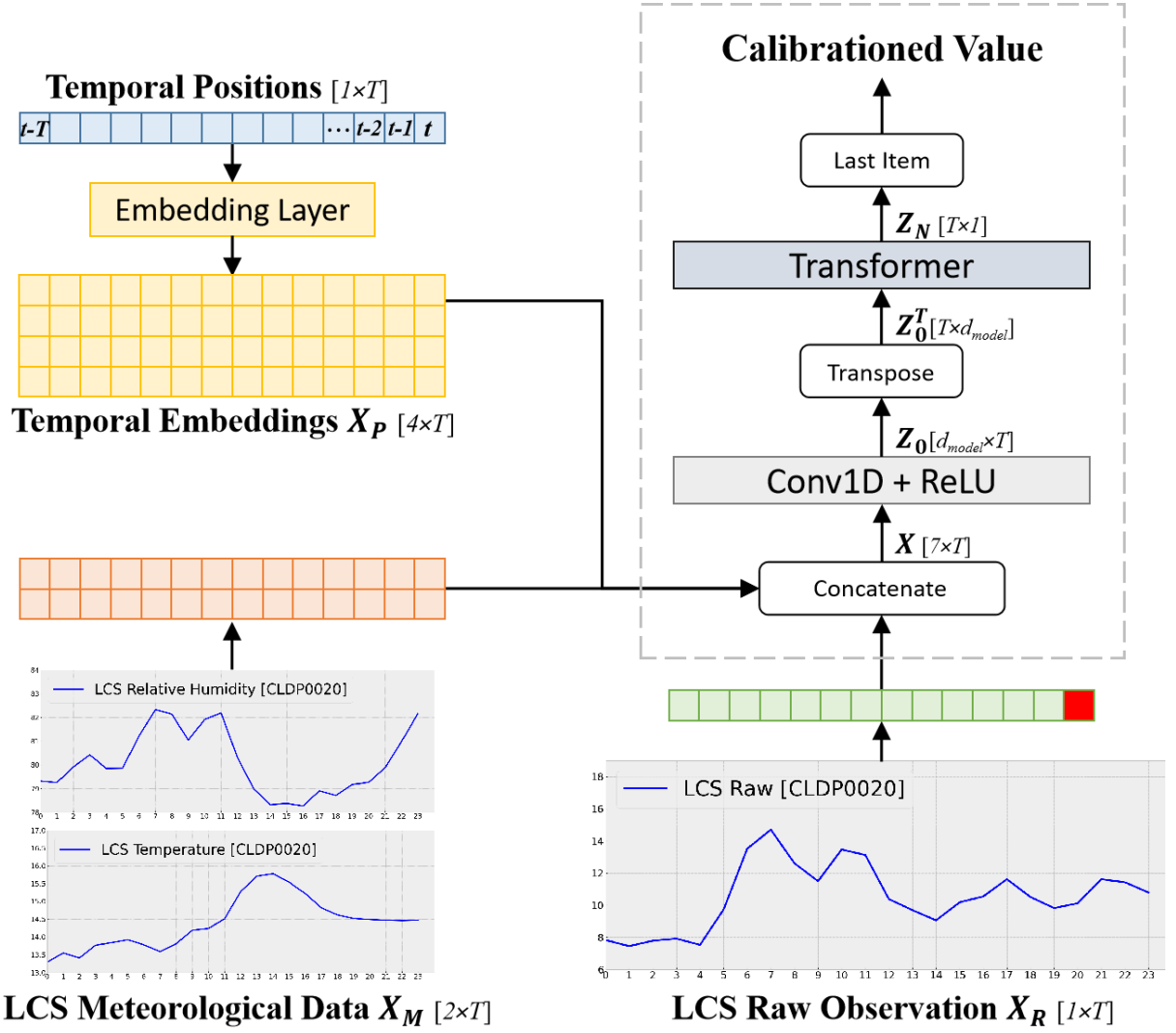


Figure 2.4 Network framework of the Transformer model for low-cost sensors calibration

Input Data Integration

The input data primarily includes following components:

LCS raw observation: The raw observation data $X_R \in R^{1 \times T}$ captured by low-cost sensors, which forms the core measurement sequence. This data is used directly as a feature, representing the direct readings from the sensor over the given time window. The last item in the sequence corresponds to the sensor reading at time t that requires calibration.

LCS meteorological data: Auxiliary variables including temperature and relative humidity collected by the sensor, donated as $X_M \in \mathbb{R}^{2 \times T}$. These are also used directly as features in their original sequence form, with each variable aligned to the time window structure, providing environmental context that may influence the sensor readings.

Temporal positions: Information encoding the specific hour-of-day timing of each data point within the input time window. It preserves natural hourly progression to help the model distinguish time-dependent patterns. Unlike the fixed positional encoding used in the original Transformer, I employ an embedding layer to convert discrete hour values into continuous features. This design is motivated by the fact that the time window's endpoint t can take arbitrary values, meaning the hour values corresponding to each position within the window are not fixed. In this frame work, I set the dimension of embedding layer is 4, generating temporal position embedding $X_P \in \mathbb{R}^{4 \times T}$.

To form a unified input for the Transformer, these components are integrated by concatenation: X_R , X_M and X_P are concatenated along the feature dimension. This generates a merged feature map, which serves as the foundation for X , the input to the calibration model.

$$X = \text{Concat}(X_R, X_M, X_P) \quad (2.10)$$

Transformer-Based Calibration Model

The merged feature map X is fed into the first layer of the calibration model, which consists of a Conv1D layer followed by a ReLU activation function. This layer expands the number of feature maps to capture more complex input patterns. It also smooths the data and handles potential outliers by introducing non-linearity and filtering noise through the activation function. The output of this layer, denoted as Z_0 , is then transposed and fed into the first Transformer encoder block. Following

processing through the Transformer's encoder-decoder structure and auto-regressive generation, $Z_N \in \mathbb{R}^{T \times 1}$ is derived. The last element of Z_N is the calibrated result for the last element of X_R .

By using this architecture, the model learns a mapping from a sequence of raw sensor readings and auxiliary data to a calibrated output. It leverages multi-head attention to weigh the contributions of different past measurements. This attention-driven sequence modeling is what allows the Transformer approach to capture temporal dependencies that linear regression or random forest could not. Instead of assuming a fixed, static relationship between past and current values, the Transformer dynamically adjusts how much each past reading influences the current calibration based on learned patterns for example, accounting for environmental factor delays, or periodic behaviors.

2.2.4 Dataset and Preprocessing

This section describes the datasets in the London study case used to demonstrate the proposed calibration method, which include:

- **Low-Cost Sensor Air Quality Data:** Raw measurements collected from Breathe London sensor network. These measurements require calibration to enhance accuracy.
- **Reference Air Quality Data:** High-precision air quality measurements obtained from regulatory reference stations of the London Air Quality Network (LAQN), used as ground truth for calibration.
- **Environmental covariates:** Additional environmental and contextual variables, including site classification (roadside, urban background, etc.), meteorological data (temperature and humidity), and temporal factors (hour-of-day time indicators).

Reference Monitor Network

Urban air quality is typically monitored using networks of highly accurate, reference-grade instruments. These instruments provide precise pollutant measurements and continue to play a critical role in advancing understanding of air pollution dynamics. However, due to high costs of installation and maintenance, these reference networks commonly exhibit limited spatial coverage, posing challenges for capturing fine-scale variations in air pollution, a problem experienced by cities worldwide.

A notable example of such a reference monitoring network is the London Air Quality Network (LAQN), established in 1993 by researchers at King's College London and is currently managed by Imperial College London's Environmental Research Group (Londonair, 2024). LAQN is recognized as one of Europe's most comprehensive urban monitoring networks, significantly contributing to the understanding of urban air pollution dynamics. Each monitoring station within the LAQN is equipped with automated reference-grade sensors that provide hourly measurements of various air pollutants such as $\text{PM}_{2.5}$ and NO_2 . Similar to many urban monitoring networks, LAQN categorizes its monitoring stations based on their proximity to pollution sources: including kerbside, roadside, urban background, suburban and industrial. Roadside sites are located near major roads and primarily reflect traffic emissions; kerbside sites are positioned immediately adjacent to road edges; urban background sites represent general urban pollution away from direct sources; suburban sites are located in lower-density residential areas outside city centres; industrial sites capture pollution influenced by nearby industrial activities.

LAQN prioritizes measurement precision and reliability, which necessitates expensive instrumentation and rigorous maintenance. As a result, its monitors are relatively sparse across the city, providing high-quality long-term data but limited fine-scale spatial resolution. As of 2024, the LAQN operates 45 $\text{PM}_{2.5}$ and 81 NO_2 monitors across London. While substantial compared to many international cities, this coverage remains sparse, limiting detailed spatial characterization of urban air pollution. The number of LAQN monitors in London is summarized in Table 2.2,

while the corresponding monitor distribution maps are shown in Figure 2.5 and Figure 2.6.

Table 2.1 PM_{2.5} and NO₂ monitor stations in London Air Quality Network (2024)

Pollutant	Site Class	Number
PM _{2.5}	Roadside	25
	Suburban	2
	Urban background	11
	Kerbside	5
	Industrial	2
NO ₂	Roadside	50
	Suburban	6
	Urban background	15
	Kerbside	7
	Industrial	3

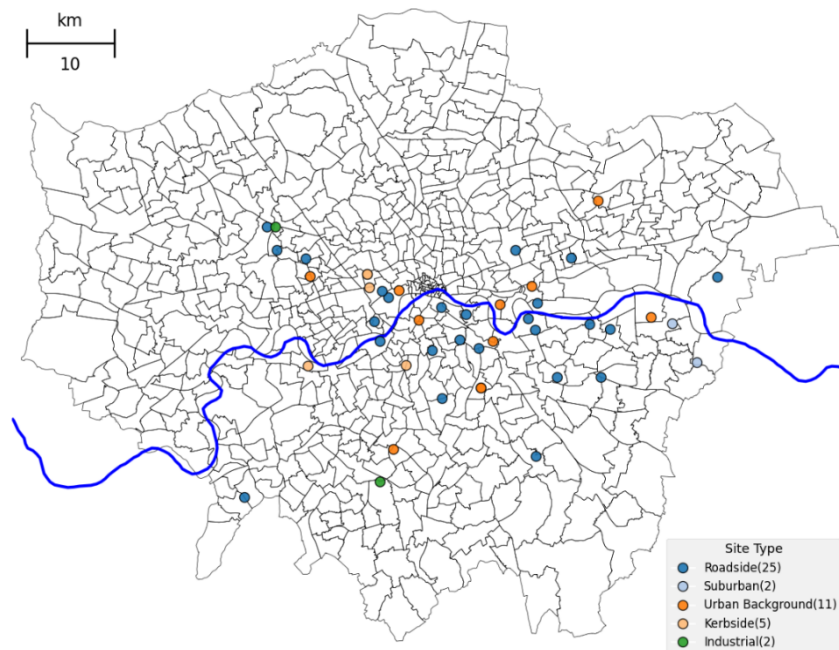


Figure 2.5 Map of PM_{2.5} monitoring stations in London Air Quality Network in 2024

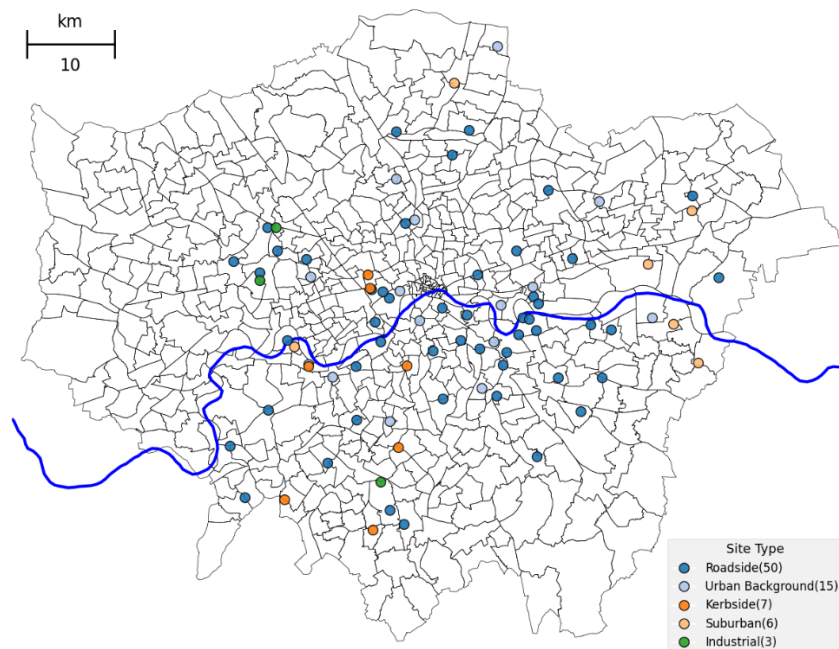


Figure 2.6 Map of NO₂ monitoring stations in London Air Quality Network in 2024

This balance of high accuracy and sparse coverage is typical of reference networks internationally, making LAQN an illustrative case of the common strengths and constraints inherent in urban air quality monitoring, and exemplifies the reference-grade systems deployed in major cities worldwide.

Low-cost Sensor Network

Recent advancements in low-cost sensor (LCS) technology offer affordable opportunities to bridge gaps in regulatory air quality monitoring and greatly improve spatial coverage (Bi et al., 2020; Minet et al., 2017; Piedrahita et al., 2014; Zimmerman et al., 2018b). LCS devices cost only a fraction of traditional reference stations, enabling the deployment of significantly denser monitoring networks. Their small size and low power requirements also enable placement in locations where reference monitors are impractical. In addition, LCSs can deliver high temporal resolution measurements (Malings et al., 2019a) and transmit data via cellular networks, ensuring real-time data accessibility. These features make LCS networks ideal for providing widespread, fine-grained air pollution data across urban areas.

The Breathe London network across Greater London is a pioneering example of a large-scale LCS deployment, designed to significantly enhance monitoring coverage. Administered by the Environmental Research Group at Imperial College London, the project actively involves citizens and community groups in air quality monitoring.

Breathe London sensor nodes (Figure 2.7), supplied by Clarity Movement Co., are equipped with electrochemical sensors for measuring nitrogen dioxide (NO_2) and optical particle counters for detecting fine particulate matter ($\text{PM}_{2.5}$). Additionally, each node continuously records ambient temperature and relative humidity to allow for corrections of environmental interference on pollutant measurements. These sensor units are solar-powered with integrated battery backups and transmit data in real-time via cellular networks. Due to their relatively low cost, portability, and compact size, the sensors can be installed in diverse locations typically unsuitable

for traditional reference-grade monitoring equipment. Consequently, these nodes enable wider spatial coverage and more flexible deployment, significantly enhancing accessibility and effectiveness in urban air quality monitoring.



Figure 2.7 Low-cost sensor used in Breathe London network, produced by Clarity

As of 2024, the Breathe London network comprises over 400 sensor nodes, and this number continues to expand. Sensor locations are classified based on proximity to local pollution sources. Kerbside nodes are located within approximately 1 meter of roads, where vehicle emissions are the predominant source of pollution. Roadside nodes, positioned slightly further away (generally within about 5 meters of the roadway), represent typical pedestrian exposure along urban streets. Urban background nodes are strategically placed away from immediate pollution sources, such as in schoolyards or residential neighborhoods, to capture representative ambient air quality conditions. Breathe London sensor nodes are typically mounted at heights of 3–4 meters, slightly higher than London’s reference monitoring stations, which are generally placed around 2 meters, to reduce interference and minimize vandalism risks. Figure 2.8 illustrates the distribution of the Breathe London nodes in 2024, clearly demonstrating significantly greater spatial coverage compared to the established London Air Quality Network (LAQN). This extensive sensor

deployment represents a substantial advancement in the spatial resolution and coverage of urban air quality monitoring data.

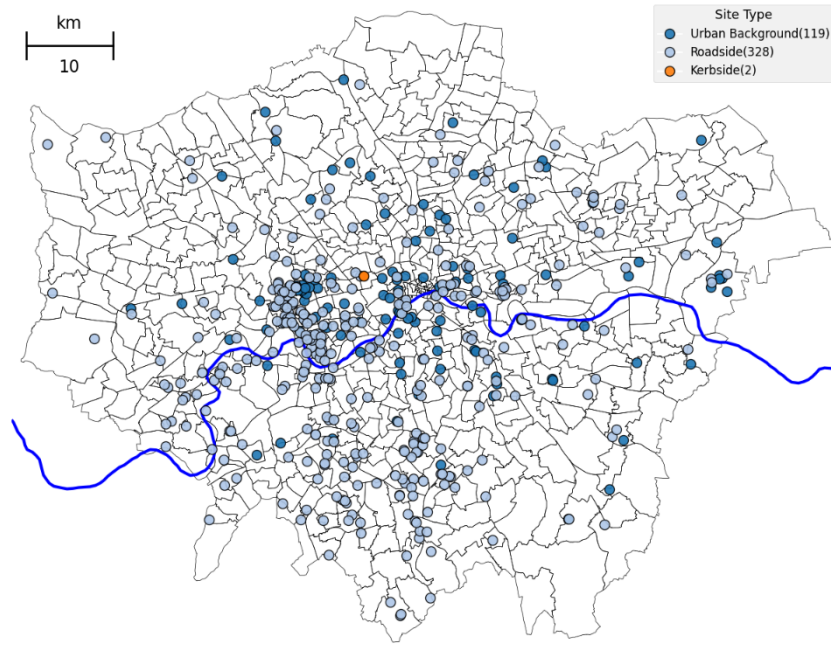


Figure 2.8 The map of Breathe London network in 2024

Despite these advantages, low-cost sensors continue to face significant challenges regarding data accuracy. Sensor readings often drift over time due to aging of sensor components or contamination of sensing surfaces (deSouza et al., 2023; Jiao et al., 2016; Malings et al., 2019b). Additionally, sensor measurements can be significantly influenced by environmental conditions, such as temperature and humidity fluctuations (Bai et al., 2020; Malings et al., 2020). Electrochemical sensors in particular are susceptible to cross-sensitivity issues, whereby their measurements may be affected by the presence of other pollutants (Cross et al., 2017; Masson et al., 2015b; Mead et al., 2013). The features and challenges of Breathe London mirror those encountered in many international efforts to modernize air quality monitoring using LCS networks.

Data Preparation Steps for Calibration Models

To develop and validate the calibration models, I constructed two separate datasets covering December 2020 to September 2024: a training dataset derived from 19 Breathe London nodes directly co-located with LAQN stations, and an independent testing dataset derived from BL nodes located near LAQN stations of matching site types.

The training dataset consists of paired measurements from BL sensors and LAQN reference stations collected simultaneously at co-located sites, alongside environmental covariates (temperature and humidity). For the independent testing dataset, BL nodes within 100 meters of LAQN stations and matching the same site type (e.g., roadside or urban background) were selected. As a result, I obtained 5 nearby node-reference pairs for PM_{2.5} and 5 pairs for NO₂.

Following dataset selection, the preprocessing pipeline included:

- (1) Hourly BL sensor measurements were matched precisely with colocation LAQN reference station data.
- (2) Integration of environmental covariates (temperature and humidity).
- (3) Data cleaning procedures to ensure high-quality data.
- (4) Feature engineering was applied to enhance calibration model performance.

The following sections describe these data and the procedures used to access and process the data including cleaning and feature engineering processes in detail.

Data Sources

To ensure rigorous quality control, the Environmental Research Group at Imperial College London continuously co-located 19 Breathe London nodes with monitoring stations from the London Air Quality Network for NO₂ and PM_{2.5} measurements (see Figure 2.8 and Figure 2.9). These co-location data underpin the calibration models aimed at improving accuracy across the wider BL sensor network.

LAQN data were retrieved via the openair R package (Carslaw & Ropkins, 2012), a tool designed for accessing UK air quality monitoring networks. I first identified LAQN stations operational from 2020 to 2024, downloaded their full data, and

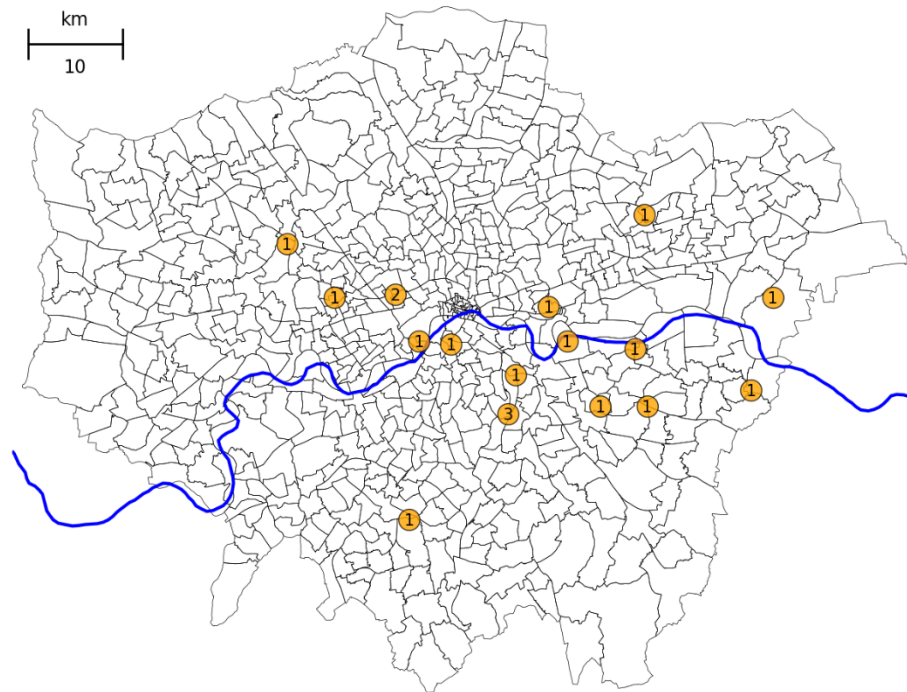


Figure 2.9 Locations of Breathe London Sensor nodes co-located with reference London Air Quality Network monitors for NO₂

extracted observations matching the BL dataset's time span.

Metadata for all low-cost sensor nodes, including node IDs and coordinates, were retrieved from the Breathe London Application Programming Interface (API) (Breathe London, 2024). These metadata were filtered to identify sensor nodes actively reporting data during the study period. Hourly temperature and humidity recorded by these nodes from December 2020 through September 2024 were subsequently obtained via the API. Hourly raw pollutant concentration data (PM_{2.5} and NO₂) for all Breathe London nodes over the same period were obtained from the Environmental Research Group, who manage the Breathe London sensor network.

For clarity, I used the raw dataset provided by the data manager, which had already been quality-controlled and formatted for analysis, rather than unprocessed sensor signals. This comprehensive dataset, spanning diverse seasonal patterns and varying pollution levels, provides a foundation for calibration analyses.

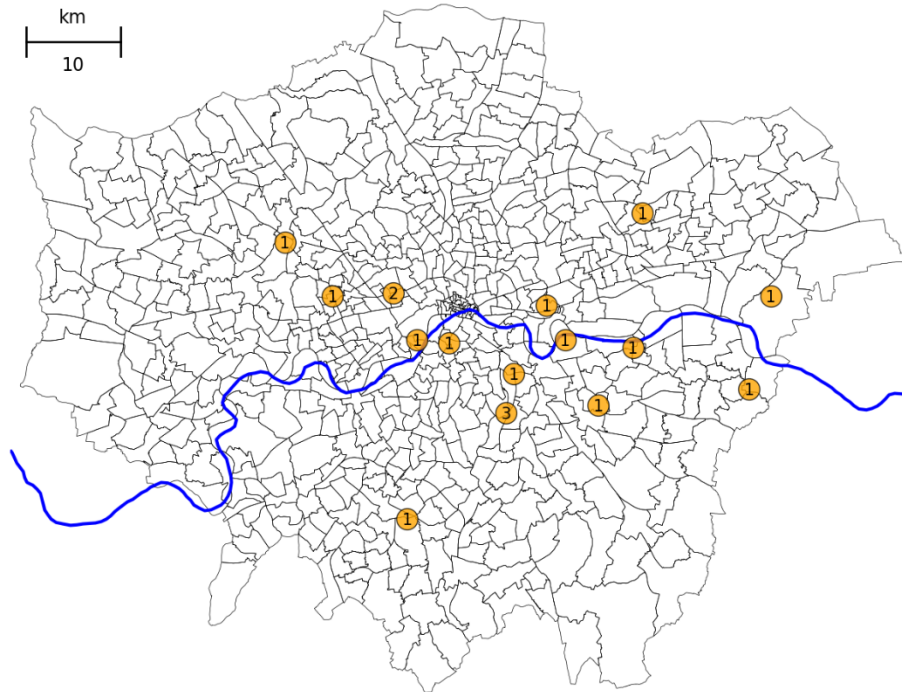


Figure 2.10 Locations of Breathe London Sensor nodes co-located with reference London Air Quality Network monitors for PM_{2.5}

Descriptive statistics (mean, minimum, and maximum) for raw PM_{2.5} and NO₂ concentrations at each co-located BL node and LAQN station are summarized in Tables 2.3 and 2.4. In some cases, the LAQN nodes reported implausible values such as negative concentrations, indicating sensor errors or outliers. These negative values might arise from instrumental noise, baseline correction procedures, or detection-limit effects, and are statistically valid in raw datasets. However, in this study the LAQN measurements are used as reference labels for training calibration models. Allowing negative reference concentrations to propagate into the calibration process may result in calibrated low-cost sensor outputs that are also negative, which

are physically implausible and inconsistent with the intended use of the calibrated data for subsequent air quality modelling. For this reason, negative LAQN values were excluded during the calibration dataset construction stage. Importantly, the proportion of negative values in the LAQN data is extremely small, accounting for 0.0871% of NO₂ observations and 0.7738% of PM_{2.5} observations over the study period. Given their rarity, the removal of these values has a negligible impact on the overall data distribution and does not materially affect the calibration results or subsequent analyses.

Table 2.2 Each pair of co-located Breathe London nodes and London Air Quality Network sites for PM_{2.5}, covering the period from December 2020 to September 2024

BL Node ID	LAQN sites	Latitude	Longitude	PM2.5 from BL ($\mu\text{g}/\text{m}^3$)		PM2.5 from LAQN ($\mu\text{g}/\text{m}^3$)	
				mean	min-max	mean	min-max
CLDP0010	LW2	51.47495	-0.03964	13.80	0.00~139.4	12.0	-16.50-201.28
CLDP0011	GR9	51.45636	0.04072	13.19	0.00~143.7	8.80	-10.65-218.03
CLDP0013	BX9	51.46598	0.18488	13.77	0.00~333.9	9.22	-4.30-145.59
CLDP0014	RB7	51.56948	0.08290	17.35	0.84~361.6	9.08	-5.00-186.00
CLDP0015	MR9	51.52254	-0.15459	14.92	0.00~191.9	9.63	0.31-181.25
CLDP0016	MR9	51.52254	-0.15459	16.65	0.00~153.6	9.63	0.31-181.25
CLDP0018	TH4	51.51505	-0.00841	14.19	0.75~136.5	10.4	-9.80-194.45
CLDP0019	GR4	51.45206	-0.04684	13.72	0.00~1238	8.23	-3.60-143.50
CLDP0021	GN6	51.49465	0.009208	11.94	0.00~126.8	9.00	-6.28-81.10
CLDP0022	SK6	51.49316	-0.10153	12.60	0.00~179.2	8.50	0.23-183.94
CLDP0023	ST5	51.38929	-0.14166	11.74	0.00~177.8	9.57	-4.00-309.00

CLDP0024	GN0	51.49053	0.074003	12.97	0.00~186.2	10.4	-16.90-185.48
CLDP0025	HV1	51.52079	0.20546	16.06	0.58~335.8	8.10	-9.69-169.21
CLDP0026	WM0	51.49468	-0.13194	12.05	0.00~168.2	9.93	0.00-265.00
CLDP0027	KF1	51.52048	-0.21265	14.00	0.00~206.0	8.57	0.00-121.56
CLDP0028	BT4	51.55247	-0.25809	17.37	0.00~4408	11.9	-15.00-131.00
CLDP0029	HP1	51.45206	-0.04684	15.73	0.00~219.9	7.91	0.18-189.17
CLDP0030	HP1	51.45206	-0.04684	12.08	0.00~185.7	7.91	0.18-189.17

Table 2.3 Each pair of co-located Breathe London nodes and London Air Quality Network sites for NO₂, covering the period from December 2020 to September 2024

BL Node ID	LAQN sites	Latitude	Longitude	NO ₂ from BL (µg/m ³)		NO ₂ from LAQN (µg/m ³)	
				mean	min-max	mean	min-max
CLDP0010	LW2	51.47495	-0.03964	39.2	-11.89~1386	27.3	-5.37~121.3
CLDP0011	GR9	51.45636	0.040725	20.2	-25.94~79.12	22.7	-2.64~143.0
CLDP0013	BX9	51.46598	0.18488	16.6	-83.53~83.41	17.0	-2.97~128.1
CLDP0014	RB7	51.56948	0.082907	12.6	-182.26~88.7	22.7	-0.80~233.9
CLDP0015	MR9	51.52254	-0.15459	30.2	-35.86~347.4	41.1	-3.05~191.2
CLDP0016	MR9	51.52254	-0.15459	21.5	-12.78~90.82	41.1	-3.05~191.2
CLDP0018	TH4	51.51505	-0.00841	19.2	-26.26~63.43	37.7	2.97~156.10
CLDP0019	GR4	51.45206	-0.04684	11.4	-167.86~50.4	12.7	-0.64~86.32

CLDP0020	GB6	51.45630	0.085606	14.0	-100.18~83.0	21.9	-2.08~139.3
CLDP0021	GN6	51.49465	0.009208	17.1	-75.26~61.68	23.1	-1.36~122.3
CLDP0022	SK6	51.49316	-0.10153	15.9	-27.71~66.53	19.9	-8.12~132.0
CLDP0023	ST5	51.38929	-0.14166	13.2	-65.65~88.33	21.9	0.20~133.04
CLDP0024	GN0	51.49053	0.074003	17.5	-93.74~106.6	24.5	-2.83~129.5
CLDP0025	HV1	51.52079	0.20546	13.5	-37.13~66.90	21.4	0.53~134.12
CLDP0026	WM0	51.49468	-0.13194	15.9	-66.29~71.34	23.9	-0.28~160.5
CLDP0027	KF1	51.52048	-0.21265	14.1	-30.36~61.66	17.1	0.00~191.41
CLDP0028	BT4	51.55247	-0.25809	24.7	-56.46~180.5	42.1	1.13~427.60
CLDP0029	HP1	51.45206	-0.04684	12.3	-62.48~77.00	15.2	-8.22~113.5
CLDP0030	HP1	51.45206	-0.04684	13.4	-34.60~62.48	15.2	-8.22~113.5

Integration of Environmental Covariates

I integrated key environmental covariates into the sensor dataset to account for external factors affecting sensor performance and pollutant concentrations. Specifically, I included meteorological variables(ambient temperature and relative humidity) because of their known effects on sensor measurements and pollutant dispersion. For example, particulate matter sensor accuracy can be significantly impacted by humidity. These covariates were obtained from concurrent meteorological measurements recorded by the Breathe London nodes and aligned with each hourly sensor observation using timestamps.

Temporal and site type were also included as covariates. Each measurement was tagged with hour-of-day and site type classification (roadside or urban background) to capture diurnal emission patterns and local environmental context. Incorporating

these factors enables the calibration models to account for variations in environmental conditions, enhancing their accuracy and reliability.

Table 2.4 Environmental Covariates used in sensor calibration

Covariate Type	Variables Included
Meteorological variables	Temperature
	Humidity
Site classification	Roadside or urban background
Temporal factor	Hour of day

Data Cleaning and Quality Assurance

Rigorous data cleaning procedures were implemented to ensure the integrity and reliability of the BL and LAQN measurements before calibration. Problematic data were identified and removed according to the following criteria:

- **Physical Plausibility:** Measurements falling outside realistic ranges for ambient air pollution were removed. Specifically, particulate matter (PM_{2.5}) readings exceeding 500 µg/m³ and nitrogen dioxide (NO₂) readings above 400 µg/m³ and negative values were considered implausible under typical urban conditions and thus excluded. This step mitigates the impact of spurious extreme values caused by sensor malfunction or environmental interference.
- **Outlier Removal (Z-score Method):** Extreme outliers in the sensor measurements were identified using the Z-score method. For each pollutant, I computed the Z-score for every hourly observation, defined as $z = \frac{x - \mu}{\sigma}$, where x is the individual reading, μ is the mean, and σ is the standard deviation of all valid measurements. Observations with $|Z| > 3$ were classified as outliers and removed. This threshold was selected based on the statistical property that approximately 99.7% of values in a normal distribution lie

within ± 3 standard deviations. The Z-score approach was selected because it provides a simple, transparent, and distribution-aware method for identifying statistically anomalous observations relative to the overall data distribution. Data beyond this range likely reflect anomalies such as sensor noise, transient spikes, or atypical environmental interference. Removing these outliers helps prevent distortion in model calibration.

- **Data Completeness:** For each pollutant (PM_{2.5} and NO₂), any hourly observation missing data from the BL sensor, LAQN reference measurement, or meteorological variables (ambient temperature and relative humidity) was removed from the dataset. This criterion ensured complete data coverage, as calibration models require simultaneous pollutant concentrations and environmental measurements to accurately adjust for external influences.

After applying these cleaning steps, the dataset was considered high-quality and suitable for model development. This cleaned training dataset (from the 19 co-located nodes) and the cleaned testing dataset (from the nearby reference-node pairs) are summarized in Table 2.5 to Table 2.6. The training data were subsequently used for calibration model fitting, while the testing data were reserved for performance validation.

Table 2.5 Training dataset for PM_{2.5} and NO₂ colocation calibration

LAQN Code	BL Code	Start Date	End Date	PM _{2.5} Hours	NO ₂ Hours
LW2	CLDP0010	2021-02-12	2024-09-30	17123	25088
GR9	CLDP0011	2021-02-10	2024-09-30	24614	28846
BX9	CLDP0013	2021-02-12	2022-01-01	7027	30249
RB7	CLDP0014	2021-02-19	2024-04-03	19688	18626
MR9	CLDP0015	2021-02-12	2024-09-30	28571	29152
MR9	CLDP0016	2021-02-12	2024-09-30	27972	29196
TH4	CLDP0018	2021-02-16	2022-08-16	8260	14174
GR4	CLDP0019	2021-02-16	2023-05-26	11901	15157
GB6	CLDP0020	2021-02-16	2024-09-03	0	26876
GN6	CLDP0021	2021-02-16	2024-09-30	23550	30538

SK6	CLDP0022	2021-02-17	2024-09-30	27274	27063
ST5	CLDP0023	2021-02-17	2024-06-20	26372	28421
GN0	CLDP0024	2021-02-23	2024-09-30	22164	30408
HV1	CLDP0025	2021-02-23	2024-09-15	24227	31083
WM0	CLDP0026	2021-02-19	2023-03-10	15491	20357
KF1	CLDP0027	2021-02-19	2022-01-20	6852	29423
BT4	CLDP0028	2021-02-19	2024-09-30	27564	28760
HP1	CLDP0029	2020-12-21	2024-09-30	29155	31434
HP1	CLDP0030	2021-04-16	2024-09-30	28303	29796

Table 2.6 Testing dataset for PM_{2.5} and NO₂ from the nearby reference-node pairs

Pollution	LAQN Code	BL Code	Start Date	End Date	Num Hours
PM _{2.5}	SK8	CLDP0042	2021-05-25	2024-09-30	27359
	WM5	CLDP0277	2023-12-07	2024-09-30	5824
	GB0	CLDP0020	2021-02-16	2024-09-30	18251
	KC1	CLDP0562	2024-01-18	2024-09-30	2881
	GN3	CLDP0153	2022-01-01	2024-09-30	16525
NO ₂	SK8	CLDP0042	2021-05-25	2024-09-30	26511
	GN3	CLDP0153	2021-09-23	2024-09-30	25477
	WA9	CLDP0472	2024-02-13	2024-09-30	5239
	ME9	CLDP0283	2022-04-15	2024-09-30	20790
	ME9	CLDP0445	2023-03-17	2024-09-30	10673

Feature Engineering:

In preparation for the calibration model training, several features were systematically engineered from the cleaned dataset to enhance predictive accuracy. Categorical variables were transformed using one-hot encoding, where each category is represented as a binary indicator variable, and continuous variables were normalized, as detailed below:

- **Categorical Variables:** Categorical descriptors, including the site type (roadside, urban background) and the hour of day, are transformed into binary

indicator variables (one-hot vectors). This approach enables the models to separately capture distinct effects associated with each site type and hourly interval. Such encoding effectively models diurnal emission patterns and site-specific sensor behaviors.

- **Normalization of Continuous Variables:** Continuous covariates, including ambient temperature and relative humidity, were scaled to a standard [0, 1] range using min-max normalization. This step ensures that variables with larger numerical ranges do not dominate the model's learning process. While normalization has limited effect on tree-based models such as Random Forest, it is critical for deep learning architectures like Transformers, which are sensitive to input scale and benefit from stabilized gradients during training.

These feature engineering steps produced a refined set of input variables optimized for calibration model development. After data cleaning and careful feature engineering, the final datasets consisted of a training set and an independent testing set. The training subset was employed to fit the model parameters, while independent testing set was held out (not used during calibration model training stage) for assessing the calibration models' generalization performance. These datasets were subsequently used to train and rigorously evaluate three calibration algorithms.

2.2.5 Model Optimization and Validation Method

The training and testing datasets were introduced in Section 2.3. All models training were carried out in Python on an NVIDIA RTX 3090 GPU. For both multiple linear regression and deep learning model, I employed the mean squared error (MSE) loss. The MSE loss is defined as:

$$L = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (2.11)$$

where y_i is the ground-truth reference value and $\hat{y}_i$ is the model prediction. By squaring the errors, MSE heavily penalizes larger discrepancies, which encourages the model to focus on reducing big calibration errors. This leads to improved overall calibration accuracy – the models learn to align the low-cost sensor outputs with the reference as closely as possible for all samples, without letting a few large errors dominate. Unlike linear regression and deep learning model, which minimizes a global loss function such as MSE, Random Forest does not optimize an explicit global objective. Instead, each decision tree within a Random Forest is trained in a greedy manner by minimizing node-level variance at each split. This local variance reduction criterion acts implicitly as a loss function, guiding the structure and performance of each tree.

Optimization for multiple linear regression model

The regression coefficients β in multiple linear regression were estimated using Ordinary Least Squares (OLS). OLS determines the optimal β values by minimizing the sum of squared residuals (defined as the differences between observed reference concentrations and model-predicted values). The closed-form solution for the OLS estimator is given by:

$$\beta = (X^T X)^{-1} X^T y \quad (2.12)$$

where $X^T X$ is the Gram matrix of the predictors and $X^T y$ is the covariance vector between predictors and observed reference measurements.

Under standard OLS assumptions, namely that residual errors (ϵ) have zero mean, constant variance (homoscedasticity), and are mutually uncorrelated, the Gauss–Markov theorem applies. This theorem states that OLS provides the Best Linear Unbiased Estimator (BLUE) for the regression coefficients – in other words, among all linear unbiased estimators, OLS yields the smallest variance for the coefficient estimates (Puntanen & Styan, 1989). Provided the underlying assumptions of linear regression are valid, OLS remains the optimal linear unbiased method for estimating pollutant concentrations during sensor calibration.

Optimization for Random Forest Model

The performance of Random Forest (RF) critically depends on the appropriate selection and configuration of hyperparameters, which governs the trade-off among model complexity, computational efficiency, and generalization capability. This section provides a comprehensive analysis of RF hyperparameters specific to low-cost sensor calibration, including theoretical foundations, empirical selection strategies, and operational constraints. In this study, I paid particular attention to the following hyperparameters: number of trees, maximum features per split and tree depth constraints (implemented via a minimum leaf size). Table 2.1 provides an overview of these core hyperparameters and their impact.

Table 2.7 The core hyperparameters and their impacts in random forest

Hyperparameter	Symbol	Impact Mechanism
Number of Trees	n_{tree}	The number of trees in the forest. A higher number of trees increases the model's stability and accuracy but also raises computational cost and time. The law of diminishing returns applies; beyond a certain point, increasing the number of trees yields minimal improvement on model performance.
Maximum Depth of Trees	d_{max}	The maximum depth of each tree. Deeper trees can model more complex patterns but are also more likely to overfit. Setting this parameter too high usually increases the model complexity unnecessarily
Minimum Samples Split	n_{split}	The minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns, thus lowering the risk of overfitting.

Minimum Samples Leaf	n_{leaf}	The minimum number of samples a leaf node must have. This parameter ensures that leaf nodes contain more than one sample, which helps in generalizing the model.
Maximum Features	m	The number of features to consider when looking for the best split. Using fewer features reduces overfitting but can also make the trees in the forest less diverse.

Number of Trees (n_{tree}): I tested values ranging from 100 to 500 and found that RMSE plateaued beyond 300 trees. Thus, I fixed $n_{\text{tree}} = 300$ to balance accuracy and computational cost.

Max Features per Split (m): Following Breiman’s guidance (Breiman, 2001), I set $m = M/3$, which reduced out-of-bag error by increasing inter-tree diversity while retaining strong individual tree performance.

Tree Depth / Leaf Size: Instead of explicitly limiting tree depth, I constrained the minimum leaf size to 5 samples, which prevented overfitting while preserving nonlinear flexibility. Smaller leaf sizes (e.g., 1) led to overfitting, while larger values (10 or 20) increased model bias.

Optimization for Transformer-based calibration model

Hyperparameter tuning: Instead of fixing hyperparameters to arbitrary values, I optimized the model’s key hyperparameters:

- **Input sequence length (T):** the size of the historical window. I used 24 hours of window lengths.
- **Number of attention heads (H):** I tried configurations with, for example, 2, 4, 8 heads in the multi-head attention. The model with 4 heads showed slightly improved accuracy, indicating that allowing more parallel attention mechanisms helps capture diverse relationships. However, beyond 8 heads I saw diminishing returns.

- **Embedding dimension (d_{in}) and hidden units:** I treated the size of the learned embedding (and the internal dimensionality of the model’s feed-forward layers) as a hyperparameter. Tested values included 64, 128, 256. $s = 128$ worked well, balancing model capacity and risk of overfitting. The MLP layers inside the network (both in the embedding module and the per-layer feed-forward networks) were set to have on the order of a few hundred hidden units (e.g., 128 or 256), tuned for each pollutant.
- **Number of Transformer layers (N):** I experimented with different depths. For the $PM_{2.5}$ sensor calibration, a slightly deeper model was beneficial (I selected $N = 4$ blocks for both encoder and decoder as optimal). For the NO_2 sensor calibration, the optimal complexity was lower (I used $N = 2$ blocks) – likely reflecting differences in the complexity of temporal patterns for these pollutants. In general, adding more than the optimal number of layers led to overfitting or no further improvement, so I chose the smallest N that gave the best validation error for each case.
- **Training configuration:** I optimized the model parameters using the AdamW optimizer, a variant of Adam that incorporates weight decay regularization. The learning rate was managed with a schedule that first increased linearly from 0.0 to 0.001 over the first 10 epochs (warmup phase) to stabilize training, then gradually decreased to a minimum of 0.0001 in subsequent epochs to allow for fine-grained parameter adjustments, balancing rapid convergence during early training with stable refinement in later stages. The model was trained for a maximum of 100 epochs, but I employed early stopping based on validation performance – if the validation MSE did not decrease for 10 consecutive epochs, I halted training to prevent overfitting. In practice, most models converged within ~ 50 epochs before early stopping kicked in. I used a mini-batch size of 32 (each batch consisting of 32 sequences of length T) during training. Batch normalization layers (as mentioned) further helped to stabilize training, allowing us to use a relatively larger learning rate without divergence.

Validation Method

I applied three complementary methods to evaluate the performance of low-cost sensors relative to reference-grade monitoring stations:

1. Data Distribution Analysis

This data distribution analysis enabled the assessment of the range, central tendency, and overall distribution patterns of the sensor measurements relative to reference-grade data, facilitating the identification of systematic biases or anomalies. These distribution plots provide a foundational overview of sensor behavior over the study period.

2. Correlation Coefficients and Error Metrics

I assessed the agreement between sensor and reference readings using a set of statistical measures. The **Pearson correlation coefficient** was used to quantify the strength of the linear relationship between paired measurements. Additionally, I computed the root mean square error (RMSE) between each Breathe London node and its corresponding reference station to represent the average magnitude of the low-cost sensor's error. Higher correlation values (closer to 1.0) and lower RMSE values indicate better sensor performance. These metrics provide an objective basis for determining how well the low-cost sensors track the reference monitors and whether further calibration or correction is necessary to improve their accuracy. The formulas are as follows:

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^n (M_t - O_t)^2}{n}} \quad (2.13)$$

Pearson Correlation Coefficient

$$r = \frac{1}{n-1} \sum_{t=1}^n \left(\frac{M_t - \bar{M}}{\sigma_M} \right) \left(\frac{O_t - \bar{O}}{\sigma_O} \right) \quad (2.14)$$

sensor measurements (M_t) at time t with reference station observations (O_t), where n represents the total number of observations, and σ_M and σ_O denote the standard deviations of LCS and reference observations, respectively

3. Scatter Plot Analysis

Lastly, I utilized scatter plots to visually examine the correlation and agreement between low-cost sensor measurements and corresponding reference-grade monitor readings. Each plot featured a 45-degree unity line, representing ideal agreement. Data points closely clustered along this line indicate strong agreement, while systematic deviations suggest consistent under- or overestimation by the sensors. Patterns observed in these plots can highlight potential calibration errors, sensor saturation, or interference from environmental conditions affecting sensor performance. This visual analysis thus offers essential insights into overall sensor reliability, informing targeted calibration adjustments and guiding further sensor optimization.

2.3 Experiments and results

I evaluated 4 periodic calibration settings: Hourly, Daily, Weekly and Monthly. For the Transformer model, only daily, weekly and monthly were selected due to trade-offs between performance and computational constraints.

Overview of Calibration Results

Tables 2.9 and 2.10 summarize the calibration performance of multiple linear regression (MLR), random forest (RF), and transformer-based models for $PM_{2.5}$ and NO_2 , respectively, under different periodic calibration settings (e.g., hourly, daily, weekly). Across both pollutants, almost all calibrated models outperform the raw,

uncalibrated sensor data in terms of reduced RMSE and increased Pearson correlation with reference measurements.

Table 2.8 Performance comparison of multiple linear regression, random forest, and transformer models for PM_{2.5} calibration in Breathe London.

Model type	periodic calibration	Pearson Correlation	RMSE
Raw value	/	0.7393	6.99
Linear	Hourly	0.5463	4.91
Linear	Daily	0.7839	3.13
Linear	Weekly	0.7737	3.18
Linear	Monthly	0.7659	3.23
RF	Hourly	0.7693	3.21
RF	Daily	0.7839	3.17
RF	Weekly	0.7959	3.09
RF	Monthly	0.8017	3.01
Transformer	Daily	0.8157	2.75
Transformer	Weekly	0.7965	2.90
Transformer	Monthly	0.8026	2.86

* All models use the same covariates: meteorological variables, site classification, and temporal factor. Bold highlights indicate the best performance.

Table 2.9 Performance comparison of multiple linear regression, random forest, and transformer models for NO₂ calibration in Breathe London.

Model type	Periodic calibration	Pearson Correlation	RMSE
Raw value	/	0.5722	14.17
Linear	Hourly	0.6235	14.12
Linear	Daily	0.7692	10.62
Linear	Weekly	0.7494	11.14
Linear	Monthly	0.7269	11.55
RF	Daily	0.7872	10.29
RF	Weekly	0.7790	10.60
RF	Monthly	0.7676	10.90
Transformer	Daily	0.8372	8.61

Transformer	Weekly	0.7766	9.97
Transformer	Monthly	0.7586	10.57

* All models use the same covariates: meteorological variables, site classification, and temporal factor. Bold highlights indicate the best performance.

For PM_{2.5} the uncalibrated sensor data had an RMSE of 6.99 $\mu\text{g}/\text{m}^3$. After applying multiple linear regression with selected covariates (introduced in Table 2.4), the RMSE was reduced to 3.13 $\mu\text{g}/\text{m}^3$ (daily model). Similarly, for NO₂, multiple linear regression reduced the RMSE from 14.17 $\mu\text{g}/\text{m}^3$ to 10.62 $\mu\text{g}/\text{m}^3$. These improvements confirm the effectiveness of baseline statistical calibration.

Nonlinear Calibration: Random Forest vs. Multiple Linear Regression

To evaluate the benefits of nonlinear modeling, I compared RF with the MLR baseline. For PM_{2.5}, the daily RF model achieved an RMSE of 3.01 and a Pearson r of 0.78, outperforming MLR (RMSE: 3.13, r : 0.76). For NO₂, the daily RF model achieved $r \approx 0.79$ vs. 0.76 for MLR.

The performance of the proposed Transformer-based Model

I further assessed the Transformer-based model, which incorporates temporal dependencies absent in MLR and RF. On the PM_{2.5} task, the Transformer achieved the best performance: RMSE of 2.75 $\mu\text{g}/\text{m}^3$ and Pearson $r \approx 0.84$, outperforming RF (3.01, $r \approx 0.78$) and MLR (3.13, $r \approx 0.76$). For NO₂, similar improvements were observed: the Transformer yielded lower RMSE and higher correlation than both baselines across all time windows.

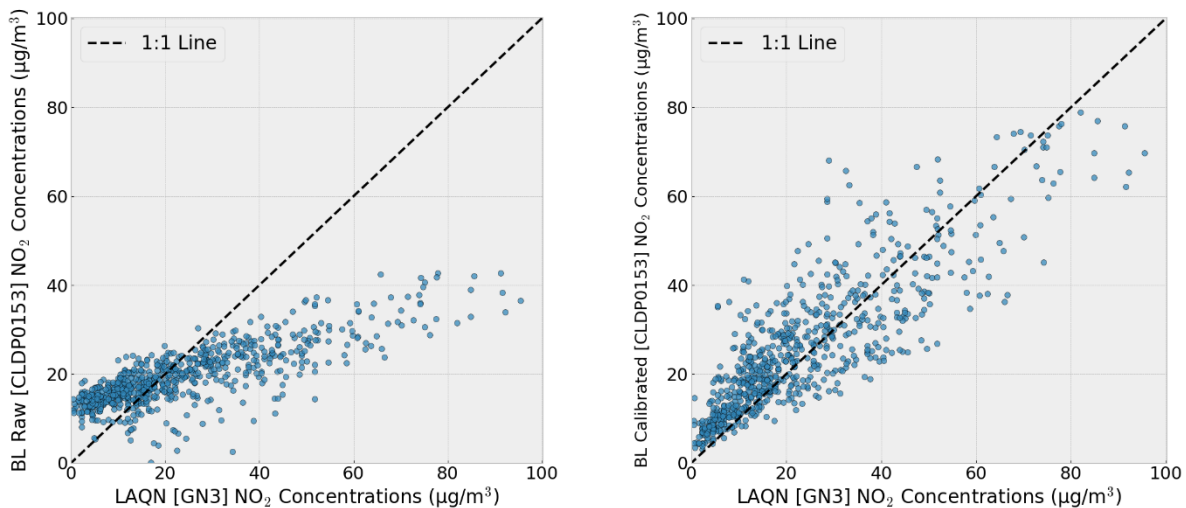
Visual Comparison

Figure 2.11 illustrates scatter plots of NO₂ measurements before and after Transformer model calibration for three random selected Breathe London (BL) nodes, including CLDP0153, CLDP0283, and CLDP0042, drawn from the test dataset, alongside their co-located reference-grade monitors in January 2024. The

scatter plots demonstrate that, after calibration, the NO₂ observations show improved alignment with reference values and reduced systematic bias.

A similar pattern is observed in the calibrated PM_{2.5} data, as shown in Figure 2.12, which displays scatter plots before and after calibration for three random selected Breathe London (BL) nodes, including CLDP0153, CLDP0020, and CLDP0042 from the test dataset, alongside their co-located reference-grade monitors. Following Transformer-based calibration, the PM_{2.5} measurements from BL nodes exhibit tighter clustering and reduced bias relative to the reference data.

Overall, these visualizations complement the quantitative results, confirming that that calibration not only improves accuracy metrics but also reduces systematic biases. More detailed results are provided in Appendix B.



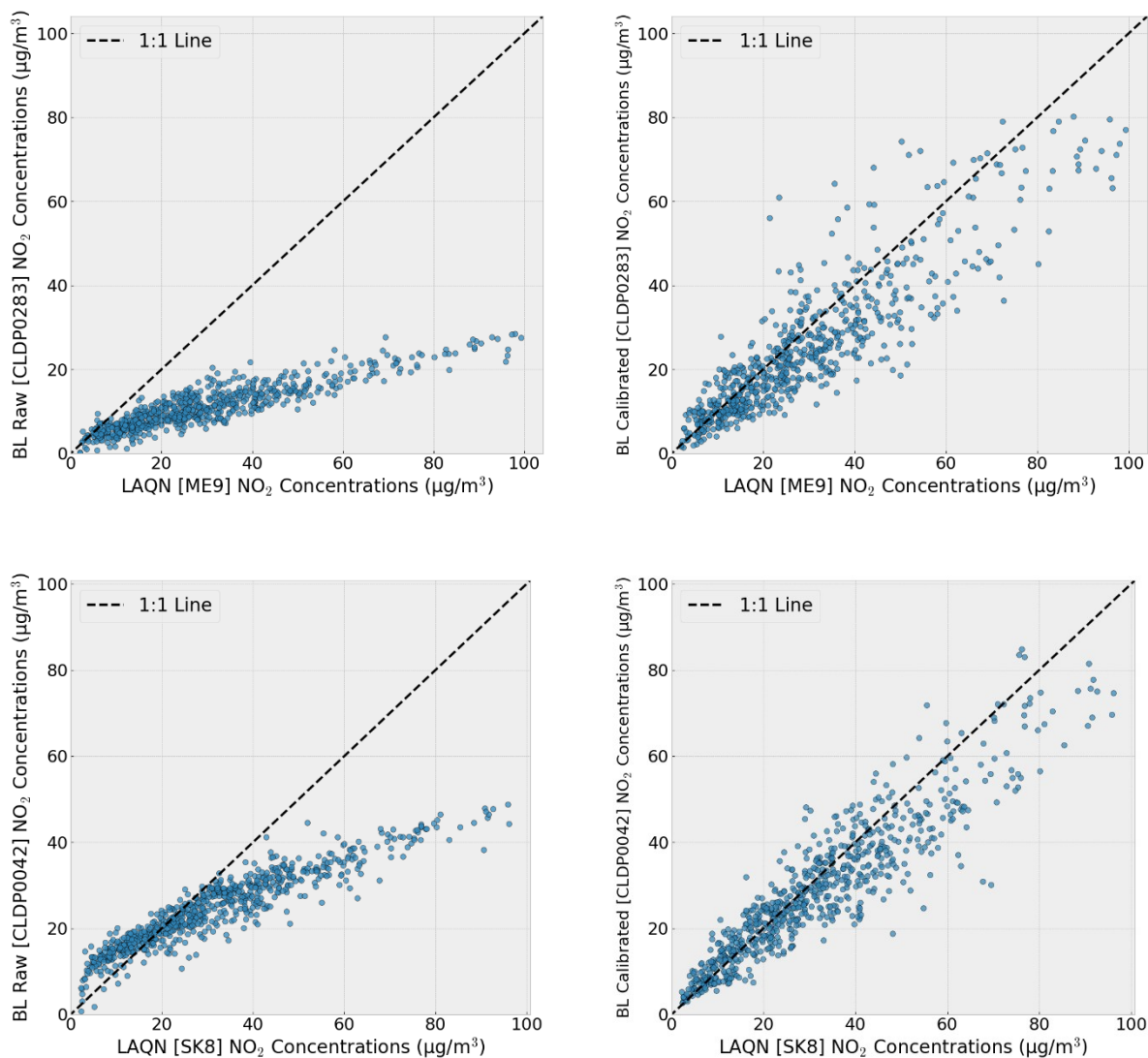
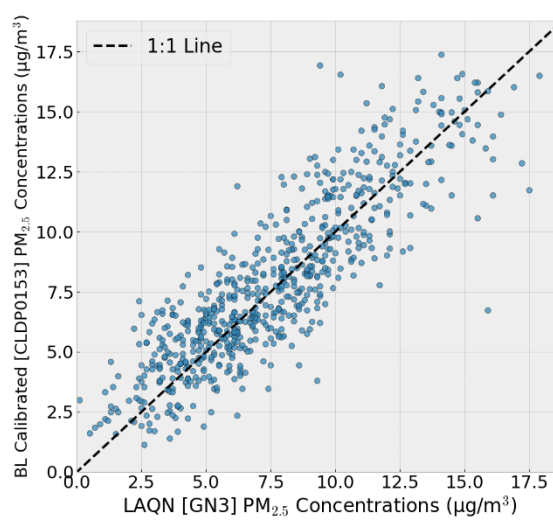
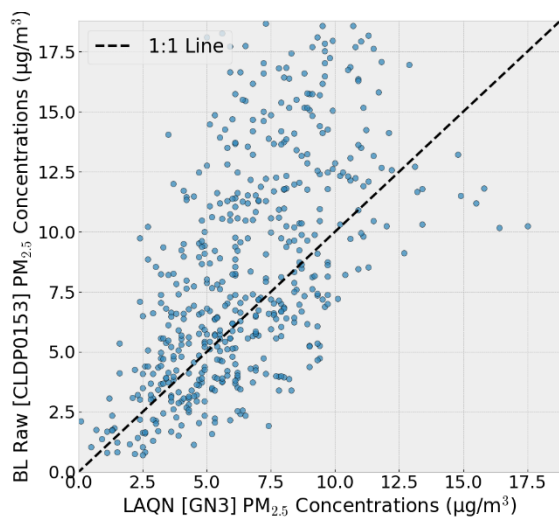
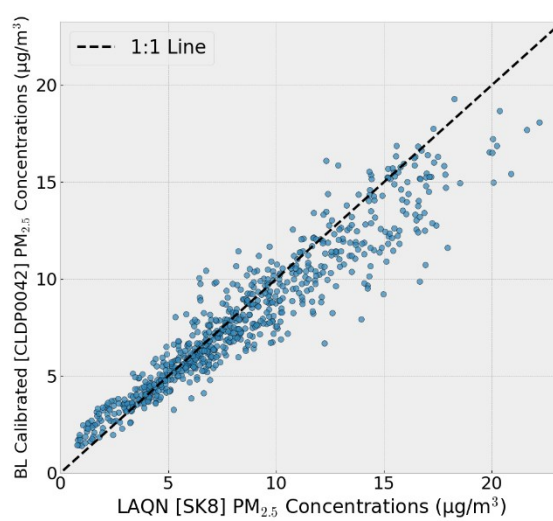
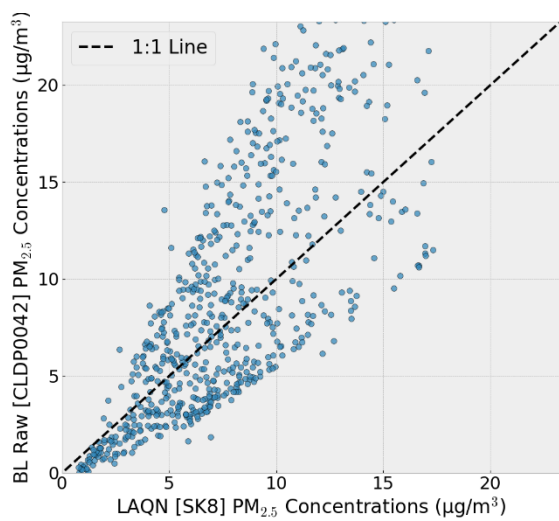


Figure 2.11 Scatter plots of raw (left) and daily transformer-calibrated (right) NO₂ measurements from three selected Breathe London nodes (CLDP0153, CLDP0283, and CLDP0042), compared to London Air Quality Network stations (January 2024)



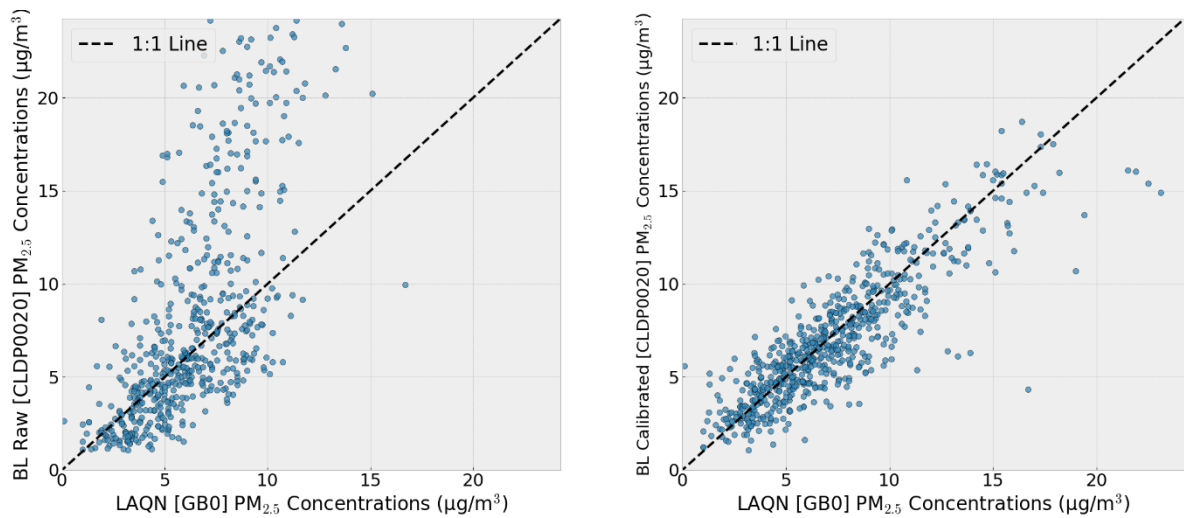


Figure 2.12 Scatter plots of raw (left) and daily transformer-calibrated (right) $\text{PM}_{2.5}$ measurements from three selected Breathe London nodes (CLDP0153, CLDP0020, and CLDP0042), compared to London Air Quality Network stations (January 2024)

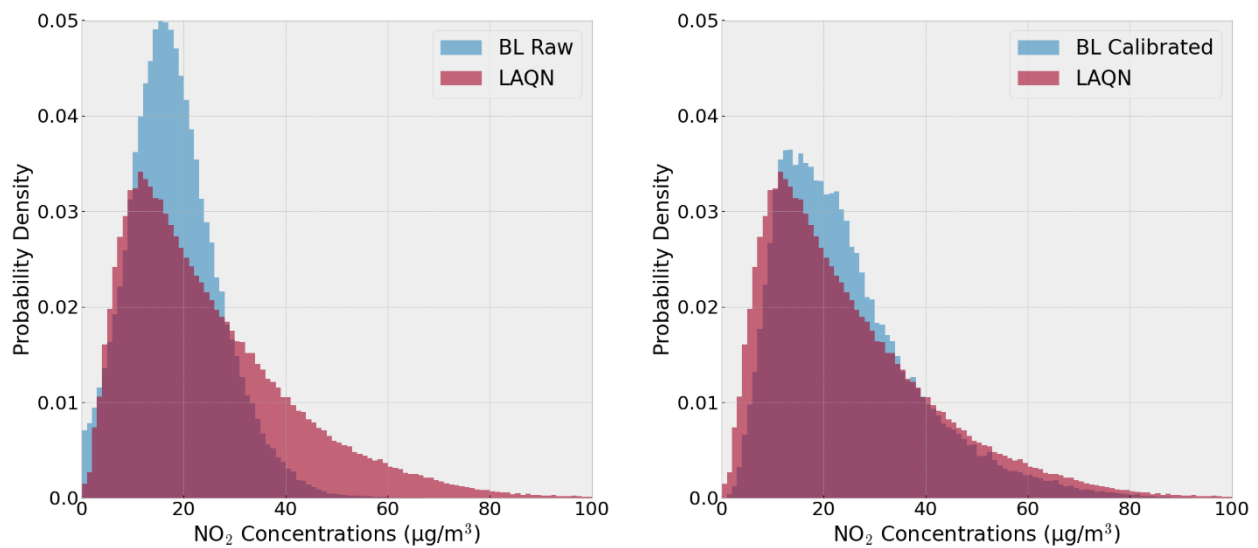


Figure 2.13 Distributions of raw(left) and daily transformer-calibrated (right) NO_2 measurements from Breathe London, compared to London Air Quality Network, across the full test dataset (2020–2024)

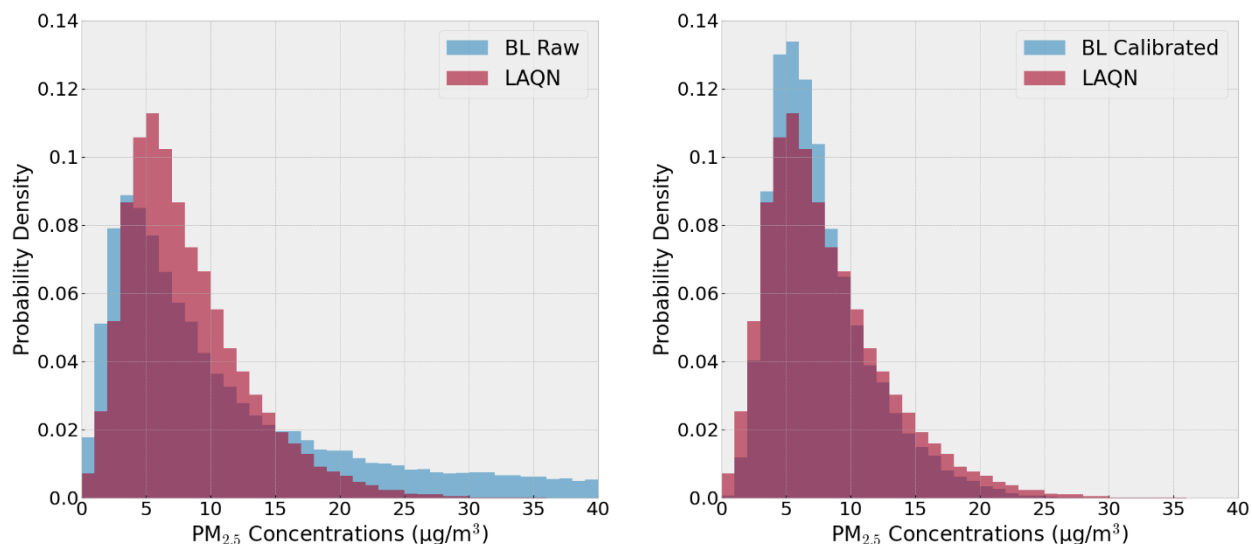


Figure 2.14 Distributions of raw(left) and daily transformer-calibrated (right) PM_{2.5} measurements from Breathe London, compared to London Air Quality Network, across the full test dataset (2020–2024)

I further compared the distributions of raw and calibrated Breathe London (BL) sensor measurements with those from near-by reference monitoring stations using the test dataset spanning from 2020 to 2024. As illustrated in Figures 2.13 and 2.14, after calibration, the distributions of PM_{2.5} and NO₂ measurements from the BL nodes closely align with those from the reference monitors, indicating substantial improvement in measurement accuracy and consistency following calibration.

2.4 Discussion

Overview of Key Findings and Contributions

This chapter addressed the critical challenge of calibrating large-scale low-cost air quality sensor network. I focused on PM_{2.5} and NO₂ measurements from Breathe London network, comparing calibration approaches ranging from a traditional multiple linear regression (MLR) to a machine learning ensemble (random forest) and a novel transformer-based deep learning model. By developing a scalable

calibration framework, I aimed to improve data reliability across a large urban sensor network. My results demonstrate that proper calibration can elevate low-cost sensors to near reference-grade performance, with calibrated sensor outputs showing high correlation with regulatory-grade monitors. The proposed transformer-based model, in particular, emerged as a standout contribution of this chapter – it leverages temporal patterns in the sensor data via sequence modeling, yielding more accurate and stable calibrations over time than the baseline methods.

The key contributions of this work include: (1) A periodic re-calibration framework that can be scaled to hundreds of low-cost sensors using co-location data, and (2) the introduction of a transformer-driven calibration approach that captures temporal dependencies (e.g. diurnal cycles) to enhance calibration accuracy beyond conventional models.

Comparison of MLR, RF, and Transformer Models

All three calibration methods substantially improved sensor agreement with co-located reference stations, but with clear differences in performance.

- **MLR** offered baseline adjustment using linear regression. It yielded moderate error reductions and correlation improvements, increasing Pearson correlation coefficients to the 0.70–0.80 range. However, MLR struggled to capture nonlinear effects, limiting its overall effectiveness.
- **Random Forest (RF)** consistently outperformed MLR, achieving lower RMSE and higher correlation across pollutants. Its performance advantage stems from its ability to model complex, nonlinear interactions and hierarchical dependencies among features. These findings align with prior work (Zimmerman et al., 2018), which also highlighted the superiority of RF over linear calibration techniques.
- **Transformer-based models** achieved the best overall performance. By incorporating historical sensor data as input sequences, the model effectively

captured time-varying dynamics. The transformer consistently achieved the **lowest RMSE and highest correlation**, improving RMSE by approximately 10–20% relative to RF and MLR.

The transformer's attention mechanism allowed for dynamic weighting of prior observations or features, enabling the model to adaptively calibrate based on learned temporal patterns. In contrast, both MLR and RF, which rely on instantaneous input features, lacked the capacity to model time-dependent changes in sensor behavior.

Strengths and Limitations of the Proposed Calibration Approach

The strength is the scalability and generalizability of my approach. The calibration models were trained on data from reference-collocated sensors and then applied across the citywide network. After calibration, the overall distribution of calibrated low-cost sensor measurements closely matched that of the reference-grade observations, indicating that systematic discrepancies were substantially reduced. Moreover, when evaluated on an independent testing dataset, the proposed calibration approach achieved the lowest RMSE and the highest correlation among the tested calibration approaches. These results demonstrate that a well-trained model can generalize calibration to sensors beyond the training locations. This is crucial for real-world scenarios where only a limited subset of low-cost sensors can be co-located with reference instruments.

The methodology is also pollutant-agnostic in principle: I showed it works for both NO_2 and $\text{PM}_{2.5}$, and it could be extended to other pollutants or sensor types with appropriate training data. Moreover, by incorporating environmental features (such as humidity and temperature), my approach inherently accounts for known interference factors, enhancing accuracy.

Despite these strengths, several limitations must be acknowledged. 1. First, the field deployment was conducted in a limited range of environments, primarily urban locations, lacking coverage in heavy industrial zones. This means the calibration models were not exposed to the full diversity of air pollutant sources (e.g. industrial

emissions), and their performance in such conditions remains unverified. The generalizability of the results in industrial or other extreme environments is therefore uncertain. 2. Second, the **co-location density and strategy** used for calibration present additional limitations. It remains an open question how the calibration accuracy would scale with fewer reference sites or how it would benefit from a denser network of collocations. These considerations highlight the need for further research to optimize reference network design for large-scale deployments. In addition, because the calibration models are trained using data from sensors co-located with reference monitors, they are inherently limited in their ability to capture sensor-specific drift occurring in devices that are not co-located. Low-cost sensors deployed away from reference sites may experience gradual performance degradation or changes in sensitivity over time that are not directly observable in the training data. As a result, such drift may remain partially uncorrected when applying a globally trained calibration model across the wider sensor network. 3. Third, and more fundamentally, calibration does not eliminate all sources of measurement variability in low-cost sensor data. While the proposed framework substantially reduces systematic discrepancies between low-cost sensors and reference-grade instruments, it can only correct error components that are explainable given the available input variables and the adopted calibration model structure. Residual variance remains due to factors such as sensor-intrinsic noise, short-term environmental fluctuations, unobserved micro-scale atmospheric processes, and sensor ageing effects. Importantly, this remaining variability may differ across sensors, locations, and time periods, meaning that calibrated measurements should not be interpreted as deterministic representations of true pollutant concentrations. The unexplained variance remaining after calibration constitutes a non-negligible source of uncertainty that cannot be ignored when calibrated sensor data are used for subsequent modelling or policy-relevant analyses. Failure to account for this residual uncertainty may lead to overconfident estimates and misinterpretation of fine-scale pollution patterns, which is further introduced in Chapter 4.

To further evaluate the robustness of the proposed calibration framework under different conditions, a set of sensitivity analyses was conducted, as described in the following section.

Sensitivity Analysis

Although the proposed calibration framework exhibits strong performance under the deployment conditions considered in this study, its practical utility cannot be assumed to be invariant across monitoring contexts. In real-world urban deployments, calibration performance is intrinsically shaped by constraints on reference monitor availability, auxiliary data completeness, and the representativeness of sensor configurations. Examining how the framework responds to alternative deployment scenarios is therefore essential for understanding its robustness, transferability, and limitations beyond the London case study.

Effect of Reference Site Availability

To examine how calibration performance depends on the availability of reference co-location sites, I conducted a targeted sensitivity analysis in which the calibration model was trained using progressively smaller subsets of reference-located sensors. Specifically, three configurations were evaluated, using 5, 10, and all 19 available co-location sites, respectively. This experiment was designed to assess the robustness of the proposed framework under reference-constrained deployment scenarios commonly encountered in cities with limited regulatory monitoring infrastructure.

To maintain computational feasibility, a single calibration model was trained using the full 2020–2024 training dataset, rather than employing the full periodic recalibration strategy described earlier in the chapter. This simplified training regime provides a consistent and controlled basis for evaluating how reductions in reference-site coverage affect overall calibration performance.

Table 2.10 Calibration performance under varying numbers of reference co-location sites (5, 10, and 19)

Pollution	Number	Pearson Correlation	RMSE
NO ₂	5	0.6736	12.16
	10	0.7195	12.08
	19	0.7279	11.69
PM _{2.5}	5	0.7774	3.2312
	10	0.7844	3.2201
	19	0.7986	3.1011

The calibration performance under the three configurations is summarized in Table 2.10. The results show a clear but gradual improvement in model performance as additional reference sites are incorporated. For NO₂, the Pearson correlation increases from 0.67 when using 5 sites to 0.72 with 10 sites, and further to 0.73 when all 19 sites are included, while RMSE decreases from 12.16 $\mu\text{g}/\text{m}^3$ to 11.69 $\mu\text{g}/\text{m}^3$. A similar trend is observed for PM_{2.5}, where correlation improves from 0.78 to 0.80, accompanied by a reduction in RMSE from 3.23 $\mu\text{g}/\text{m}^3$ to 3.10 $\mu\text{g}/\text{m}^3$.

These results indicate that calibration accuracy benefits from increased reference-site availability, but with diminishing marginal returns beyond a moderate number of co-locations. The most substantial performance gains occur when increasing the number of reference sites from 5 to 10, whereas the improvement from 10 to 19 sites is comparatively modest. This suggests that once the calibration model has been exposed to a sufficiently diverse set of environmental conditions and site characteristics, additional reference sites primarily serve to refine the learned calibration relationships rather than fundamentally altering them.

Importantly, reasonable calibration performance is achieved even with a limited number of reference sites, particularly for PM_{2.5}. This finding has direct implications for real-world deployment: while dense reference networks remain desirable, the proposed framework does not require exhaustive co-location coverage to function effectively. Instead, a relatively small number of strategically selected reference sites

may be sufficient to support large-scale calibration. Such behaviour is essential for transferability to other urban environments globally, where the number of regulatory monitoring stations is often limited by cost, infrastructure, and logistical constraints.

Effect of Meteorological Data Availability (Temperature and Relative Humidity)

To further assess the robustness of the proposed calibration framework under data-constrained deployment scenarios, I conducted an additional sensitivity analysis examining the impact of excluding key meteorological covariates, namely temperature (T) and relative humidity (RH), from the calibration inputs. This scenario reflects realistic conditions in many cities globally, where collocated or high-quality meteorological measurements may be unavailable, incomplete, or temporally misaligned with low-cost sensor observations.

For this experiment, calibration models were trained using the same aggregated 2020–2024 dataset as in the previous sensitivity analysis, once including T and RH as input features and once excluding them entirely. The resulting calibration performance is summarized in Table 2.11.

Table 2.11 Calibration performance with and without temperature and relative humidity

Pollution	T and RH	Pearson Correlation	RMSE
NO ₂	No	0.61	13.28
	Yes	0.72	11.69
PM _{2.5}	No	0.79	3.10
	Yes	0.75	3.31

The results reveal distinct pollutant-specific sensitivities to meteorological data availability. For NO₂, excluding temperature and relative humidity leads to a pronounced degradation in calibration performance: the Pearson correlation decreases from 0.72 to 0.61, while RMSE increases from 11.69 $\mu\text{g}/\text{m}^3$ to 13.28 $\mu\text{g}/\text{m}^3$. This substantial performance loss highlights the strong dependence of NO₂ sensor response on environmental conditions, consistent with the known sensitivity

of electrochemical NO₂ sensors to temperature-driven diffusion processes and humidity-related interference effects. In contrast, the impact of removing meteorological inputs on PM_{2.5} calibration is comparatively modest and exhibits a different pattern. While the Pearson correlation remains high (0.79 without T and RH versus 0.75 with T and RH), RMSE shows only a marginal change (3.10 µg/m³ without T and RH versus 3.31 µg/m³ with T and RH).

These findings have important practical implications. For NO₂, the inclusion of meteorological variables is critical for achieving reliable calibration, and omitting them can lead to substantial residual bias and uncertainty. For PM_{2.5}, however, acceptable calibration performance may still be achievable under limited meteorological data availability. Overall, this sensitivity analysis demonstrates that the proposed calibration framework exhibits heterogeneous dependence on auxiliary data availability across pollutants.

Effect of Reference Site Type (Roadside-only versus Mixed Site Types)

A further sensitivity analysis was conducted to examine the impact of reference site composition on calibration performance, focusing on a scenario in which only roadside reference sites are available for model training. This configuration reflects a realistic deployment constraint in many cities, where regulatory monitors are disproportionately located near major roads to capture peak traffic-related pollution, while background or suburban reference sites are sparse or absent.

Using the same aggregated 2020–2024 dataset and training strategy as in the previous sensitivity experiments, calibration models were trained either with a mixed set of reference sites (including roadside, urban background, and other site types) or with roadside sites only. The resulting performance metrics are reported in Table 2.12.

Table 2.12 Calibration performance under different site configurations (roadside-only versus mixed site types)

Pollution	Site Type	Pearson Correlation	RMSE
NO ₂	Mixed Site Types	0.72	11.69
	Only Roadside Sites	0.71	11.56
PM _{2.5}	Mixed Site Types	0.79	3.10
	Only Roadside Sites	0.78	3.28

The results indicate that restricting the calibration to roadside reference sites leads to only a modest change in overall performance. For NO₂, the Pearson correlation decreases slightly from 0.72 to 0.71, while RMSE marginally improves from 11.69 $\mu\text{g}/\text{m}^3$ to 11.56 $\mu\text{g}/\text{m}^3$. For PM_{2.5}, the correlation declines from 0.79 to 0.78, accompanied by a small increase in RMSE from 3.10 $\mu\text{g}/\text{m}^3$ to 3.28 $\mu\text{g}/\text{m}^3$.

From a deployment perspective, these results indicate that roadside-only reference networks are not inherently prohibitive for calibration, particularly for traffic-dominated pollutants such as NO₂. Nevertheless, incorporating a diversity of site types remains advantageous for achieving balanced performance across pollutants and environments. Mixed site-type coverage helps ensure that the calibration model is exposed to a broader range of concentration regimes and sensor response conditions, thereby reducing the risk of systematic bias when deployed citywide. Importantly, the performance variations observed in these sensitivity analyses are relatively modest compared with the improvements achieved by the proposed transformer-based calibration framework. This suggests that while deployment constraints such as reference-site availability and auxiliary data completeness can influence calibration performance, the modelling framework itself plays a more dominant role in determining overall calibration accuracy.

Summary

In summary, this chapter introduces a scalable periodic recalibration framework specifically designed for large-scale low-cost air quality sensor networks. It also presents a novel transformer-based deep learning model tailored to sensor calibration, capable of capturing temporal dependencies and enhancing data quality. While the framework was demonstrated in the context of Greater London, which was selected

for its extensive availability of low-cost and reference-grade monitoring infrastructure, the proposed approach is readily transferable to other urban environments. By enhancing measurement accuracy without incurring additional equipment costs, this approach facilitates the acquisition of dense, high-quality air pollution data, thus advancing the potential for urban air quality monitoring and research.

The calibration results obtained in this chapter establish a reliable foundation for subsequent fine-scale air quality modelling. By improving the accuracy and consistency of low-cost sensor measurements, the calibrated dataset enables the development of a deep learning framework that can more effectively integrate heterogeneous spatial and temporal information without being compromised by significant sensor bias. In the next chapter, I build upon these calibrated observations to develop an advanced Vision Transformer (ViT)-based model capable of estimating hourly $\text{PM}_{2.5}$ and NO_2 concentrations at 100 m spatial resolution across Greater London. This transition marks a shift from improving data quality to leveraging these enhanced datasets for high-resolution spatiotemporal air pollution mapping.

Chapter 3: Development of Deep Learning Framework for Fine-scale Spatiotemporal Air Quality Modeling based on Low-Cost Sensors

Urban air pollution exhibits substantial spatial variability at fine scales, primarily driven by localized emission sources such as traffic and industrial activities (Marshall et al., 2008; Tan et al., 2014). Additionally, air pollution concentrations can fluctuate considerably within short time intervals, displaying pronounced hourly variations (Sokhi et al., 2021). Understanding these fine-scale spatiotemporal pollution patterns is therefore crucial for identifying pollution hotspots and implementing targeted strategies to protect vulnerable urban populations from adverse health effects.

Building upon the calibration work presented in Chapter 2, which aligned low-cost sensor (LCS) measurements with reference-grade monitors to produce high-quality data, this chapter introduces a novel deep learning framework based on the Vision Transformer (ViT) architecture. The framework is designed to exploit dense sensor data from LCS networks for modeling air pollution at high spatial and temporal resolutions. Specifically, the proposed approach integrates calibrated LCS data with supplementary datasets, including reference monitoring stations, meteorological data, traffic data and land-use information, to train a ViT-based deep learning model capable of generating accurate air pollution maps at high spatial and temporal resolutions.

3.1 Literature Review

Physics-based Air Quality Modeling Approaches

Physics-based models simulate air pollution diffusion, physical and chemical processes, accounting for emissions from different sources. For example, Gaussian plume models are among the simplest: they assume a steady-state Gaussian

distribution of pollutant concentration downwind of a source, enabling quick estimates of near-source dispersion (Holmes & Morawska, 2006). These models are computationally efficient and have been widely used for regulatory applications and screening studies due to their rapid results (Leelőssy et al., 2014). However, Gaussian models are limited by strong assumptions – they perform poorly in calm or non-steady meteorological conditions and cannot easily account for complex terrain or chemical reactions (Holmes & Morawska, 2006; Leelőssy et al., 2014) .

For larger spatial scales and chemically reactive pollutants, Eulerian chemical transport models (CTMs) such as the Community Multiscale Air Quality (CMAQ) model are employed. These models solve the atmospheric diffusion–advection–reaction equations on a 3D grid, simulating pollutant transport, dispersion, deposition, and chemical transformations over regional to continental scales (Byun & Schere, 2006a). CTMs can capture complex photochemical interactions (e.g., ozone and secondary aerosol formation) and time-evolving meteorology, providing a detailed representation of air quality dynamics (Holmes & Morawska, 2006). The enhanced realism of CTMs comes with significant limitations. They are computationally intensive, often requiring high-performance computing to run fine-resolution or long-term simulations (Yang et al., 2024). The necessity of extensive input data (high-resolution meteorological fields and up-to-date emission inventories for industry, traffic, etc.) is another challenge – inaccuracies or outdated data in these inputs can lead to substantial prediction errors (Sokhi et al., 2022). Moreover, CTMs produce results on a large grid scale, which can dilute extreme local variations (Leelőssy et al., 2014).

At the micro-scale (urban street or neighborhood level), Computational Fluid Dynamics (CFD) models are used to resolve detailed flow patterns and pollutant dispersion (Leelőssy et al., 2014). This physics-based approach allows representation of complex geometries such as buildings and street canyons, giving high-resolution pollutant concentration distributions that simpler models cannot provide. Strengths of CFD models include their flexibility and fidelity in describing local airflow phenomena – for example, they can predict how building layouts

induce circulation or how street-level turbulence affects pollutant dispersion, which is critical for urban air quality assessments (Holmes & Morawska, 2006). However, CFD simulations are notoriously expensive computationally: to resolve fine-scale turbulence and flow features, they require very fine grids and small-time steps, leading to long runtimes and high memory usage (Leelőssy et al., 2014). This computational expense limits their practical domain size and simulation duration – CFD is typically applied to localized areas or short episodes rather than regional, long-term studies. Additionally, CFD models require detailed input data (such as accurate building geometry and high-resolution boundary conditions from meteorological models), and their results can be sensitive to the chosen turbulence parameterizations.

Despite these limitations, bottom-up physics-based models offer several important advantages that remain difficult for purely data-driven approaches to replicate. Because they are grounded in fundamental atmospheric processes, physical models are inherently transferable across locations and time periods, even where monitoring networks are sparse (Appel et al., 2020). Their reliance on physical laws rather than statistical correlations enables them to represent conditions outside the range of historical observations (Byun & Schere, 2006b). In addition, physics-based models uniquely enable future forecasting and scenario testing: CTMs can evaluate the impacts of hypothetical changes in emissions, land use, or policy interventions (such as ULEZ expansions) by systematically adjusting inputs or boundary conditions (M. L. Williams et al., 2018). These strengths explain why physical models remain central to regulatory decision-making and long-term air quality planning.

Overall, real-world air pollution events are complex and some empirical assumptions in physics-based models may not reflect them accurately, which degrades the model's performance. Furthermore, they are frequently associated with significant computational expenses, and acquiring sources of emission factors can also be challenging. These limitations motivate the development of complementary data-driven deep learning approaches capable of exploiting large, heterogeneous datasets to achieve high-resolution, near-real-time air quality modelling.

Data-Driven and Machine Learning Approaches

Unlike physics-based methods that require a comprehensive understanding of the inner mechanisms of air pollution, Data-driven and Machine Learning models are able to capture the intrinsic physical and chemical rules by automatically learning from a large amount of historical data, showing great promise in various air quality analytical tasks with a lower computational budget.

Early data-driven approaches included land-use regression (LUR), a method that leverages geographic characteristics to estimate pollutant concentrations at unmonitored locations (Araki et al., 2015; Hoek et al., 2008b). Specifically, LUR models utilize spatial covariates, such as land use type, proximity to roads, or population density, as predictor variables in multivariate regression frameworks to estimate air pollutant concentrations (Beelen et al., 2013; Hankey & Marshall, 2015; Hatzopoulou et al., 2017; Hoek et al., 2008b; Ryan et al., 2008). Recent advances in data availability (e.g. low-cost sensors, satellite data) and statistical methods have spurred the evolution of LUR and related approaches, enabling larger study domains and more dynamic predictors (W. Wang et al., 2022). A comprehensive review of 155 LUR studies from 2011–2023 shows that integrating multi-source observations and adopting advanced modeling techniques significantly improved predictive performance compared to classic linear models (T. Li et al., 2025a). In summary, early data-driven models like LUR established a foundation for statistical air quality prediction, and their recent enhancements illustrate the value of incorporating richer data and more flexible algorithms to address complex pollution patterns. However, the LUR models, which typically rely on linear regression, have inherent limitations in fully exploiting diverse existing and novel data sources, as they struggle to adequately capture the complex, nonlinear relationships among pollutants and predictor variables.

Building on this foundation, a growing number of studies have turned to machine learning (ML) algorithms for air quality modeling. Techniques such as random forests, gradient-boosted trees, and other ensemble learners can automatically

capture non-linear relationships among a multitude of environmental predictors (traffic, land use, meteorology, etc.), often yielding higher accuracy than conventional regressions (Brokamp et al., 2017). For example, a random forest model trained to predict particulate matter in Delhi was shown to vastly outperform simpler regression models, demonstrating its efficacy in modeling the complex interactions of air quality drivers (Singh et al., 2025). In general, tree-based ML models have proven especially effective at mapping air pollution; a recent systematic review found that these methods consistently outperform linear models in predicting ambient pollutant concentrations (Agbehadji & Obagbuwa, 2024). Overall, the adoption of machine learning techniques has markedly advanced data-driven air quality modeling, yielding robust predictive systems that can handle complex, non-linear dependencies and large-scale inputs in ways traditional models could not.

Other geostatistical frameworks, such as Kriging and Bayesian Maximum Entropy (BME), emphasize spatial correlation to improve pollutant estimations (Oliver & Webster, 1990; Xu et al., 2016a). Kriging, for instance, predicts pollutant concentrations at unmeasured locations by explicitly modeling spatial autocorrelation among sampled points. BME further extends this capability by integrating multiple, diverse data sources and capturing complex, nonlinear space–time relationships. For instance, Xu et al. (2016) employed BME to fuse CTM ozone predictions with observational data, effectively modeling their nonlinear and heteroscedastic relationships, significantly enhancing the accuracy of regional ozone maps. Audrey et al. (2014) used BME to integrate LUR and CTM outputs with observations, substantially improving NO₂ concentration estimations in the region of Catalunya.

Despite advances made in air quality modeling through data-driven methods, there remains significant challenges and key limitations, including:

- **Spatial non-stationarity:** Global models (e.g., a single LUR or RF for a large region) often cannot capture varying pollutant–predictor relationships across space. For instance, a factor that is a strong pollution driver in one area might be

less influential in another (such as urban core and suburban areas), due to unique local sources or meteorological conditions, violating model stationarity assumptions (Bertazzon et al., 2015; Milà et al., 2024). Researchers have observed that adding spatial coordinates or proxy variables into models like random forests only partially addresses this issue: such fixes can improve interpolation within well-monitored regions but tend to perform poorly when extrapolating to new areas (Milà et al., 2024). This suggests that more fundamentally spatially-aware modeling techniques may be needed to truly capture heterogeneous regional patterns.

- **Integration of heterogeneous data:** Modern air quality datasets originate from diverse sources (including ground-based monitoring stations, traffic data, meteorological models, and low-cost sensor networks), each characterized by distinct data formats, scales, and temporal-spatial resolutions. Traditional air quality models typically face difficulties assimilating these heterogeneous data streams effectively. Although ML models theoretically offer greater flexibility for integrating multiple data sources, practical data fusion remains nontrivial due to the substantial diversity in data structure and granularity. Traditional spatial methods such as LUR and kriging usually depend on pre-engineered features, which often fail to fully exploit the information available within complex datasets. Moreover, combining disparate input types (such as categorical land-use information, continuous meteorological fields, and low-cost sensor measurements) poses significant methodological challenges that exceed the capabilities of standard regression techniques and single-algorithm ML approaches, often leading to suboptimal model performance (T. Li et al., 2025b).
- **Complex spatial dependencies:** Another critical challenge in air quality modeling is effectively capturing the complex spatial dependencies inherent to pollutant distributions. Air pollutants exhibit intricate spatial patterns driven by atmospheric chemical reactions and pollutant transport processes. For instance, prevailing winds can significantly enhance spatial correlations by transporting pollutants between locations, whereas shifts in wind direction can rapidly alter or diminish these relationships. However, traditional machine learning methods

often rely heavily on extensive manual feature engineering or oversimplified assumptions, such as assuming each location's pollutant concentrations are independently influenced solely by local input features (e.g., land use characteristics within a fixed radius), thus failing to accurately represent dynamic spatial dependencies. As a result, these conventional approaches frequently yield suboptimal predictions, especially in environments characterized by strong spatial autocorrelation (Di et al., 2020).

In conclusion, while traditional data-driven methods like LUR, kriging, and RF have substantially improved exposure estimation and provided insights into pollution determinants, they each face shortcomings in modeling the full complexity of air quality dynamics. Persistent issues of spatial heterogeneity, multi-source data integration, and spatial dependencies limit the performance of these traditional approaches, especially as the demand grows for high-resolution (both spatial and temporal) pollution maps for health and policy applications. This has spurred a new generation of modeling efforts using deep learning to overcome these limitations. The motivation for these advanced methods is clear: by leveraging rich data sources and powerful learning algorithms, I aim to achieve more accurate and granular spatiotemporal air quality predictions, providing a stronger foundation for exposure assessment and air pollution mitigation strategies. Such deep learning models represent the frontier in air quality modeling, building on the strengths of earlier methods while addressing their critical gaps.

Advanced Deep Learning Methods for Air Quality Modeling

These challenges have motivated the development of advanced methodologies for high-resolution spatiotemporal air pollution modeling. In recent years, deep learning has emerged as a powerful approach in this domain, substantially extending traditional machine learning techniques. Deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other variants, provide flexible frameworks capable of automatically extracting intricate, nonlinear patterns directly from raw input data (Z. Zhang et al., 2024).

These models leverage hierarchical representations and complex nonlinear mappings to facilitate the fusion of heterogeneous data streams (such as meteorology, land use, and air quality observations), while capturing their intricate interactions.

Notably, CNNs and graph neural networks (GNNs) explicitly encode spatial dependencies: CNNs treat heterogeneous data as gridded images, while GNNs represent air quality monitoring networks as graphs with edges capturing spatial or dynamic relationships (e.g., proximity, prevailing winds). RNNs, especially Long Short-Term Memory (LSTM) networks, are well-suited for modeling temporal dependencies.

Recent research has demonstrated the strength of these models: for instance, deep learning frameworks have successfully integrated satellite imagery and meteorological data into spatial embeddings, subsequently predicting ground-level pollutant concentrations with unprecedented accuracy at resolutions as fine as 1×1 km (Ma et al., 2024). In the “DeepAir” model, Yi et al. (2018) utilized deep neural networks to fuse data from multiple air quality monitoring stations, achieving significant improvements in both short-term and long-term forecasting accuracy. CNN-based models extract spatial features from gridded or image-based environmental data (L. Liang et al., 2023; Yan et al., 2021). GNN-based methods, by integrating domain-specific knowledge about the sensor topology, have demonstrated improved predictive performance for pollutants like $PM_{2.5}$ (Wang et al., 2021).

Limitations of Existing Deep Learning Models:

While deep learning has significantly advanced air quality modeling, current architectures still face several notable limitations that constrain their effectiveness in large-scale, high-resolution applications.

CNNs are limited by their local receptive fields, where each neuron in a given layer only processes information from a small, localized region of the input. As a result, capturing long-range spatial dependencies typically requires stacking many layers,

which increases model depth and computational cost. This limits CNNs' efficiency and scalability for modeling widespread pollutant transport patterns in urban environments.

Graph-based models (e.g., GNNs) rely on a predefined graph structure, typically determined by geographic proximity or sensor connectivity. However, these predefined connections may not fully reflect complex spatial interactions, such as wind-driven pollutant dispersion or topographic effects. Inaccurate or rigid graph definitions can therefore hinder model performance, especially in heterogeneous urban settings.

Additionally, many advanced deep learning models, including RNNs, LSTMs, and GNN variants, tend to exhibit high computational complexity and reduced efficiency when scaled to networks with hundreds or thousands of sensors. Applying such models across large low-cost sensor networks can be resource-intensive, and their performance may degrade when used for large-scale and high-resolution tasks. Furthermore, these methods require fixed numbers and positions of measurements as input data and cannot accommodate time-varying sensors, rendering them impractical in real-world, where the numbers and positions of sensors typically change over time.

In summary, existing approaches do not yet fully address the need for a model that can directly learn fine-scale spatial air quality patterns from dense sensor data in a flexible and computationally feasible manner. This gap in the literature points to the need for new modeling techniques that can leverage the rich data from low-cost sensor networks to produce high-resolution air quality maps - without relying on rigid input structures or excessive computational overhead.

Vision Transformers for Fine-Scale Air Quality Modeling

To address the aforementioned limitations, I propose applying Vision Transformer (ViT) architecture to fine-scale air quality modeling. Originally developed for image recognition, Vision Transformers offer a powerful mechanism for capturing

complex spatial relationships through a global attention framework (Han et al., 2023). Unlike CNNs, which are constrained by local receptive fields, ViTs enable the model to attend to all parts of the input simultaneously, allowing for the representation of both local structures and long-range spatial dependencies within a unified modeling architecture (Dosovitskiy et al., 2021).

In the context of air quality, a city's pollution distribution can be conceptualized as a two-dimensional “image,” where each sub-region (or grid cell) serves as a patch containing environmental and sensor-derived features. The ViT processes these patches as a sequence and uses attention mechanisms to capture intricate patterns and long-range spatial dependencies across the entire urban area. This global modeling capability allows the framework to account for distant influences, such as the impact of upwind industrial emissions on downwind neighborhoods, that traditional CNNs may struggle to represent effectively.

Vision Transformers have demonstrated state-of-the-art performance in capturing spatial structures within image data, indicating their potential for modeling the complex spatial variability inherent in urban air pollution. By using positional encodings, the ViT can inherently account for the spatial location of input data (e.g. sensors or grid cells) and learn how pollution at one location relates to pollution at another. This approach is particularly appropriate for fine-scale air quality modeling because it can ingest high-dimensional inputs (such as multi-sensor readings, land use maps, meteorological fields) and model their interactions across space. Recent studies have begun to explore transformer-based models in air quality forecasting – for example, a transformer model by Liang et al. (2023) achieved improved accuracy in nationwide air quality predictions across thousands of monitoring stations. However, that and similar efforts primarily focus on temporal sequence prediction at monitoring sites, whereas my interest lies in generating or predicting a fully spatially continuous pollution map at fine resolution. To the best of my knowledge, no prior work has specifically leveraged a Vision Transformer for constructing high-resolution air quality distributions using data from low-cost sensor networks.

Research Gap and Contribution: Given the limitations of traditional machine learning and deep learning approaches in handling large-scale and heterogeneous data, the application of Vision Transformers represents a novel solution for air quality research. The ViT addresses these challenges by harnessing the strength of attention to consider both nearby and far-range influences concurrently. By introducing a ViT-based framework for fine-scale air quality modeling, I expect to advance this field by more clearly capturing spatial dependencies and effectively utilizing large-scale, heterogeneous data, ultimately leading to more accurate and high-resolution air quality maps for urban environments.

3.2 Methodology and Research Design

Building upon the potential of Vision Transformer, I propose a novel deep learning model for air quality mapping. The model aims to infer fine-scale pollutant concentration surfaces (e.g., NO_2 , $\text{PM}_{2.5}$) from a combination of low-cost sensor measurements and auxiliary covariates, including land use, traffic congestion, and meteorological variables. These data are described in detail in Section 3.2.2.

To address data heterogeneity, all input features are first projected onto a uniform spatial grid, effectively transforming the multi-source data into image-like tensors. This regularization enables compatibility with the Vision Transformer architecture. The resulting tensors are then passed to the ViT encoder, which partitions it into fixed-size patches and applies multi-head attention to model long-range spatial dependencies across the urban landscape. The encoded patch representations are subsequently concatenated and reshaped into a structured grid format, preserving spatial topology.

Finally, a lightweight decoding module refines the latent representations into a dense pollution map, delivering high-resolution predictions at each grid cell. The overall pipeline of proposed framework is illustrated in Figure 3.1.

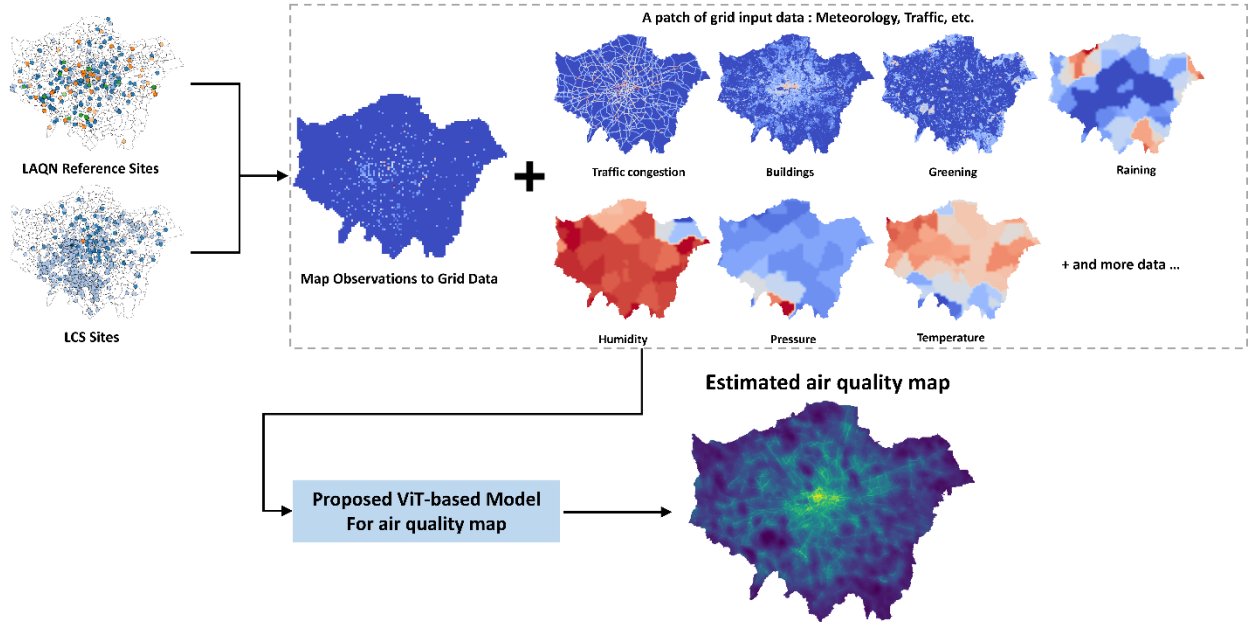


Figure 3.1 The pipeline of proposed ViT-based air quality framework

Each stage of the pipeline is described in detail in the following subsections to ensure full transparency and reproducibility.

3.2.1 Model Architecture

Problem Formulation:

My objective is to reconstruct the air pollution grid map at time t , denoted as $M_t \in R^{n_x \times n_y}$, using local sensor measurements $X_t \in R^n$ and auxiliary geographic features $G_t \in \mathbb{R}^{n_x \times n_y \times k}$. Here, n denotes the number of low-cost sensor measurements available at time t , n_x and n_y represent the number of grid points in the horizontal and vertical directions, respectively, and k corresponds to the number of covariates that provide additional spatial context. My goal is to estimate the mapping function:

$$f: (X_t, G_t) \mapsto M_t \quad (3.1)$$

that best reconstructs M_t , leveraging both sensor observations and spatially distributed auxiliary information.

Model Architecture

I propose a ViT-based air quality model, a novel integration of CNN and Vision Transformer, to generate fine-scale air pollution map. Leveraging the local feature extraction capabilities of CNN and the global context-modeling strengths of Vision Transformer, the ViT-based air quality model effectively processes diverse input data.

As illustrated in Figure 3.2, the ViT-based model architecture is composed of five major components: a CNN Encoder, a Vision Transformer Encoder, a CNN Decoder, a Feature Fusion stage, and an Output module. This design integrates convolutional neural network (CNN) layers for local feature extraction with a ViT for global context modeling. These encoding and attention modules are followed by a symmetric decoding path that reconstructs the output field. Each component plays a distinct role. The CNN Encoder extracts hierarchical feature maps from the input; the ViT Encoder captures dependencies across the entire spatial domain; the CNN Decoder restores the spatial resolution of the feature maps; the Feature Fusion stage combines encoder and decoder features; and the Output module produces the final field with appropriate scaling.

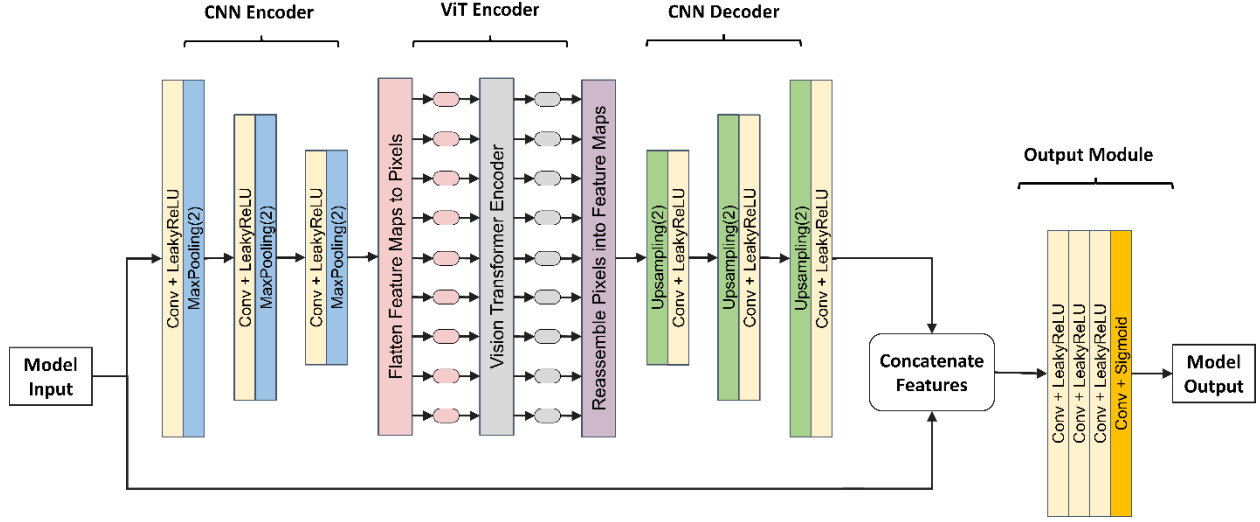


Figure 3.2 Overall architecture of the ViT-based air quality model, illustrating its five main modules. The left side (yellow and blue blocks) depicts the CNN Encoder, consisting of repeated convolutional, LeakyReLU, and pooling operations for local feature extraction. The middle section (pink and gray blocks) represents the ViT Encoder, where the encoder's feature maps are flattened into a sequence of feature vectors and processed. The right side (green and yellow blocks) shows the CNN Decoder with upsampling and convolution layers. The final yellow block represents the Output module, consisting of additional convolutional layers and a sigmoid activation to produce the output field.

CNN Encoder

The CNN Encoder module is a convolutional downsampling network that encodes the input field into a set of latent feature maps. Starting from the input data, the encoder applies a sequence of convolutional layers, each immediately followed by a LeakyReLU activation (with negative slope coefficient $\alpha = 0.1$, i.e., $f(x) = \max(\alpha x, x)$) and a 2×2 max-pooling operation for downsampling. Each convolution layer increases the number of feature channels while the pooling reduces the spatial dimensions, yielding progressively deeper but smaller feature maps. After k such encoder blocks, the input of size $H \times W$ is transformed into a compact feature map of size $H' \times W' \times C'$, where $H' = H/2^k$ and $W' = W/2^k$ (assuming H and W are divisible by 2^k). This encoded feature map provides a

condensed representation of the input, capturing multi-scale local patterns, and is passed to the ViT Encoder for global context integration.

Vision Transformer Encoder

The Vision Transformer Encoder module is a ViT-based transformer that processes the encoded feature map as a sequence of tokens to model global dependencies. First, the 3D feature map produced by the CNN encoder (of shape $H' \times W' \times C'$) is flattened into a one-dimensional sequence of $N = H' \times W'$ feature vectors, each of length C' . The Vision Transformer Encoder then applies L layers of multi-head attention and position-wise feed-forward networks (with residual connections and layer normalization) to these tokens. This attention mechanism enables the model to capture long-range spatial relationships, effectively learning interactions between distant regions of the field that convolutional layers alone might not capture. The output of the Vision Transformer Encoder is a sequence of N context-enriched feature vectors, which are then reshaped back into a two-dimensional arrangement for the decoding stage.

CNN Decoder

The CNN Decoder module converts the ViT's output back into high-resolution feature maps, essentially mirroring the encoder's process. The sequence of N ViT output vectors is first reassembled into a feature map of dimensions $H' \times W' \times C'$ (matching the encoder's output feature map size), which restores the spatial context to the globally processed features. Next, the decoder applies a series of upsampling and convolution operations to progressively recover the original spatial resolution $H \times W$. At each decoder stage, a $2 \times$ upsampling operation doubles the height and width of the feature map, and a convolution layer (with LeakyReLU activation, $\alpha = 0.2$) then refines the features. Through successive upsampling stages, the CNN Decoder produces feature maps of increasing resolution, eventually reaching the full resolution of the output space.

Feature Fusion

To preserve fine-scale spatial information and enhance reconstruction accuracy, the original input is concatenated with the feature maps produced by the CNN decoder. This fusion enables the model to integrate globally contextualized representations from the Vision Transformer with localized details from the raw input data. This fused representation carries both the global structure and the local detail necessary for accurate air pollution surface reconstruction.

Output Module

The final output module maps the fused features back to the original output space. The combined feature maps are passed through a sequence of convolutional layers with LeakyReLU activations to further refine local spatial details. Subsequently, a final 1×1 convolution layer (essentially a linear combination across feature channels) generates the raw prediction map. A sigmoid activation function applied element-wise, bounds the output values within the $[0, 1]$ range, in accordance with the normalized target variable. This step ensures the model produces valid, interpretable air quality estimations, thereby completing the end-to-end architecture.

3.2.2 Dataset and Preprocessing

To evaluate the performance of my proposed ViT-based air quality model, I applied my approach to reconstruct high-resolution spatiotemporal $\text{PM}_{2.5}$ and NO_2 maps for Greater London. Specifically, my objective is to generate hourly standardized $\text{PM}_{2.5}$ and NO_2 maps at a 100 m resolution for Greater London. Similar to other machine learning-based air quality models (Jain et al., 2021b), I employed a combination of static and time-varying predictor variables to develop the model. The datasets used in this study are introduced below:

Dataset

This study employs several datasets to demonstrate the effectiveness of the proposed methodology. The utilized datasets include:

- **Low-Cost Sensor Air Quality Data (Calibrated Breathe London):** Air pollutant measurements ($\text{PM}_{2.5}$ and NO_2) from the Breathe London sensor network, calibrated using the optimal models detailed in Chapter 2.
- **Reference Air Quality Data (London Air Quality Network):** High-precision pollutant measurements ($\text{PM}_{2.5}$ and NO_2) from regulatory-grade reference monitoring stations operated by the London Air Quality Network.
- **Land Use Characteristics:** High-resolution land use data describing urban characteristics across the Greater London area.
- **Traffic Data (Google Traffic Congestion):** Real-time traffic congestion data sourced from Google Maps, included to represent the impact of vehicular traffic on local air quality conditions.
- **Meteorological Data:** Real-time meteorological observations (e.g., temperature, humidity, wind speed) collected across the study region.

Detailed Descriptions of Key Datasets:

Breathe London Monitoring Data

All $\text{PM}_{2.5}$ and NO_2 measurements from Breathe London were calibrated using the optimal models from Chapter 2 and then used as inputs for the air quality model.

Reference Monitoring Station Data

All $\text{PM}_{2.5}$ and NO_2 LAQN measurements were served as ground truth for model training and validation, as described in Section 2.3.

Land Use Characteristic Data

Land-use data provides critical context for air quality models by capturing local emission sources and dispersion conditions (Johansson et al., 2022). Geographic Information System (GIS) datasets from OpenStreetMap (OSM) are especially valuable, as OSM offers detailed, up-to-date maps of urban features. OSM has become a foundation for high-resolution land-use mapping in modeling, supplying rich detail on areal features (e.g. land-use polygons and building footprints) and linear features (e.g. road networks) (Haklay & Weber, 2008; Mooney & Minghini, 2017). The following describes the OSM data preprocessing steps and how key land-use features – buildings, greening, and pavements – are incorporated into air quality predictions.

OSM’s raw map data must be processed into quantitative land-use variables for modeling. First, relevant features are extracted using OSM’s tagging system and converted into GIS layers. Areal features (polygon data) such as building footprints, parks, and other land-use areas are filtered and saved as separate layers, while linear features (polyline data) like roads and highways are extracted into another layer. Each feature is assigned standard coordinate reference systems so that all data align spatially with the study grid. Next, the data are cleaned and categorized: for example, road classes from OSM (e.g. motorway, primary, residential) are grouped into major vs. minor roads to simplify traffic predictors, as shown in Table 3.2. Similarly, land-use polygons are reclassified into broader categories (Building, residential, commercial, industrial, green space, etc.) for easier interpretation, as shown in Table 3.1.

Table 3.1 Land use categories used in my study

Categories	Type in OSM	Description in OSM
Building	Building	Building outlines
Industrial	Industrial	An industrial area
Residential	Residential	A residential area

Commercial	Commercial	A commercial area
Green Space	Forest	A forest or woodland
	Farmland	Agricultural land (areas where crops are grown)
	Meadow	A meadow
	Park	A park
	Nature Reserve	A nature reserve
	Recreation Ground	An open green space for general recreation
	Grass	An area where grass grows

Table 3.2 Road type used in my study

Categories	Type in OSM	Description in OSM
Major Road	Trunk	Important roads, typically divided
	Primary	Primary roads, typically national.
	Secondary	Secondary roads, typically regional.
	Tertiary	Tertiary roads, typically local.
	Motorway	Motorway
Minor Road	Residential	Roads in residential areas
	Living street	Streets where pedestrians have priority
	Pedestrian	Pedestrian only streets
	Busway	Dedicated roads for bus, usually closed for any mode of transport except public transport

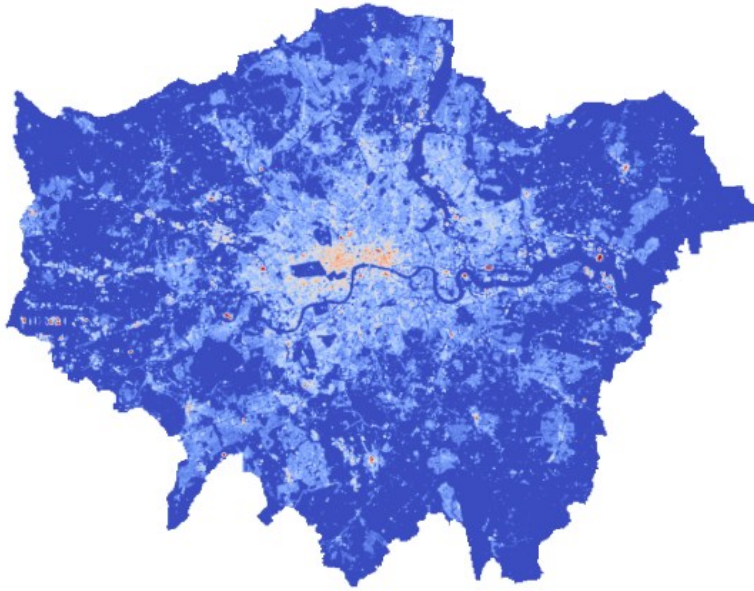


Figure 3.3 The building coverage in Greater London

Linear geographic features, primarily the road network (and other paved transport links), are extracted from OSM and integrated as key predictors of pollution. Roads are the most direct proxies for traffic-related emissions – they represent where vehicles travel, and thus where tailpipe emissions and resuspended road dust originate. In preprocessing, the total length of major roads and minor roads within a grid cell is computed. In summary, the road networks are translated into quantitative measures of traffic infrastructure for each grid cell. The detailed OSM data processing workflow is provided in Appendix B.4.



Figure 3.4 The Green coverage in Greater London

Traffic Data

In 2007, Google commenced the provision of traffic data in the form of coloured overlays on roads and motorways, indicating the speed of vehicles on specific roads. Crowdsourcing is employed to obtain the global positioning system (GPS) locations of a substantial number of mobile phone users, which are then used to create real-time traffic maps (Muñoz-Villamizar et al., 2021). I obtain real-time traffic congestion data based on the R package 'googletraffic' (Robert Marty, 2024), which is obtained through real-time data retrieval from the Google Maps Javascript API (Googletraffic, 2024). The traffic congestion defined by Google characterizes the rate of traffic flow: No traffic delay (Value 1): No congestion with traffic flowing at a normal speed; Light to moderate traffic (Value 2 and 3): Increasing congestion, with traffic flowing at a reduced speed; Heavy traffic (Value 4): Severe congestion, with heavily reduced traffic speed. Four categories of traffic conditions are colour

coded in Figure 3.4, as: 1. No traffic delays (green); 2. Light traffic (orange); 3. Moderate traffic (red); 4. Heavy traffic (dark red).

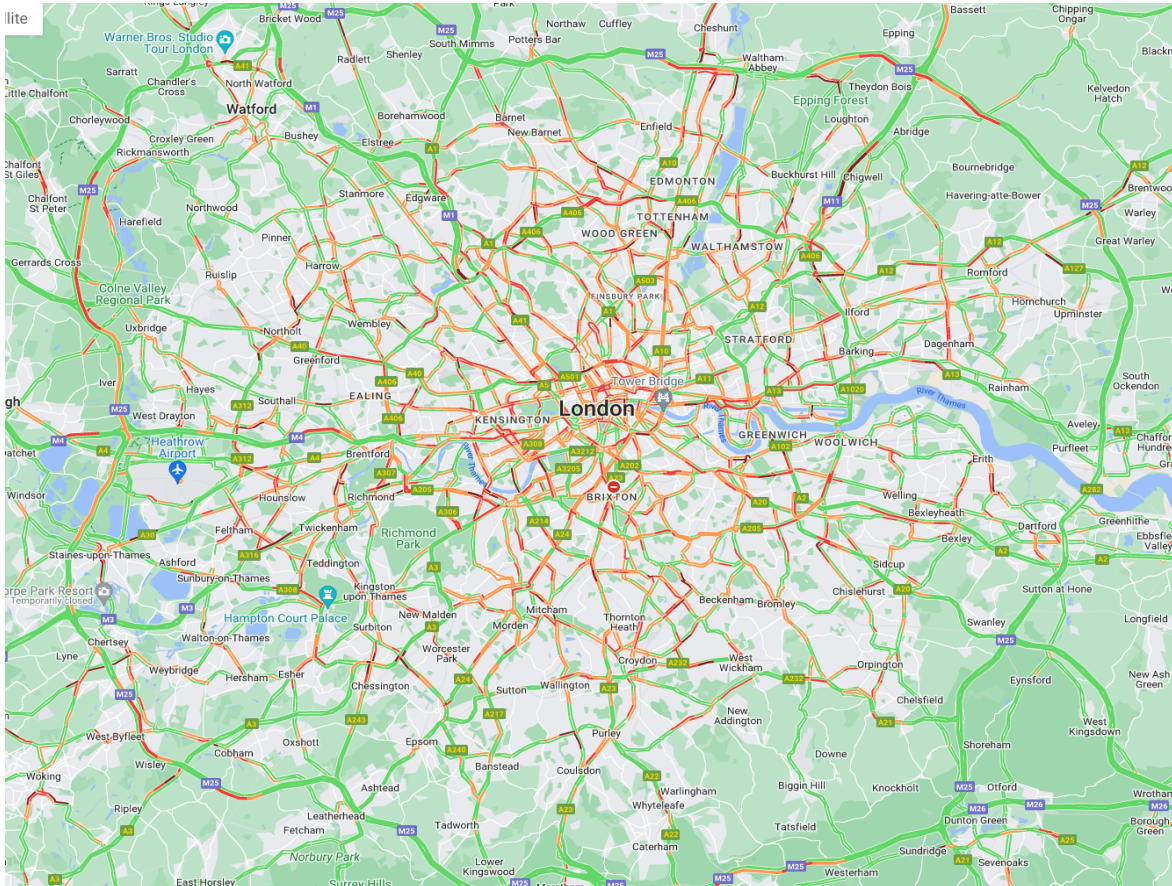


Figure 3.5 Google Maps Snapshot of Traffic Congestion in Greater London – 9:00 a.m., February 22, 2024.

Traffic congestion data were collected at hourly intervals to capture temporal variability, and subsequently converted into a georeferenced raster grid aligned with the predefined spatial grid. For each grid cell, the hourly congestion level was calculated by average congestion level. The detailed traffic congestion data processing workflow is provided in Appendix B.5.

The processed traffic congestion metrics are integrated into the air quality modeling framework as explanatory variables that represent traffic-related emission influence.

These traffic features serve as input features alongside other data like meteorology and land use. Each training sample (for a given location and time) includes not only meteorological variables (temperature, wind, etc.) and geographic descriptors, but also one or more traffic congestion indicators. This allows the model to learn associations between traffic intensity and pollutant concentrations. For example, higher congestion levels typically correlate with higher NO₂ or ultrafine particle concentrations near roads, due to increased vehicle emissions during stop-and-go traffic. By providing the model with a numeric measure of congestion, I enable it to capture these traffic-related pollution spikes.

Meteorological Data

Weather conditions strongly influence air quality, affecting both the dispersion and chemical transformation of pollutants (EPA, 2025). For example, calm air allows pollutants to accumulate near their source, whereas strong winds dilute and transport pollutants away, leading to lower local concentrations. Because of these effects, incorporating meteorological data is essential for accurate air quality modeling. Meteorological inputs help identify pollution source regions and forecast pollution episodes (such as inversion events or high-pollution days), and they form a critical component of air quality simulation systems. In summary, without accounting for meteorology, any predictive air quality model would fail to represent the real-world behavior of pollutants.

Meteorological Variables and Their Impact on Air Quality

- **Wind Speed and Direction:** Wind governs the transport and dispersion of airborne pollutants. Low wind speeds lead to poor dispersion (allowing pollutant concentrations to build up), while strong, turbulent winds mix and carry pollutants away, reducing concentrations over the source area. Additionally, wind direction determines the pathway of pollutant plumes, which helps in identifying the origin of emissions and affected downwind areas.

- **Temperature:** Temperature affects atmospheric chemistry and the vertical mixing of air. Higher temperatures (often coupled with intense sunlight) accelerate photochemical reactions that produce ozone and smog, but may decrease NO_2 . Temperature profiles (e.g. inversions, where temperature increases with altitude) can also trap pollutants near the ground, underscoring the need to include temperature data for accurate pollution prediction.
- **Humidity:** Atmospheric humidity (water vapor content) influences both chemical reaction pathways and particulate matter behavior. Moist conditions can enhance the formation of secondary pollutants – for example, water vapor facilitates the conversion of sulfur dioxide gas into aerosol sulfate (acid droplets) – and can cause aerosol particles to grow larger by absorbing moisture. In contrast, very dry conditions might limit these hygroscopic growth processes and certain chemical reactions, affecting pollutant lifetimes and visibility. Moreover, dry weather can increase the resuspension of dust and soil particles due to reduced surface adhesion, contributing to elevated levels of particulate matter.
- **Rainfall:** Rainfall has a cleansing or “scavenging” effect on the atmosphere. Rain washes out particulate matter and dissolves water-soluble gaseous pollutants, effectively removing them from the air. As a result, heavy or frequent rain events typically lead to lower pollutant concentrations and improved air quality immediately afterward.
- **Pressure:** High-pressure is characterized by stable atmospheric conditions that can lead to air stagnation. This stagnation inhibits the dispersion of pollutants, causing them to accumulate near the surface and potentially leading to pollution episodes. Conversely, low-pressure systems are associated with turbulent air movement, which enhances the dispersion of pollutants and typically results in improved air quality.

Data Acquisition and Preprocessing from OpenWeatherMap

Meteorological data for the model were obtained from the OpenWeatherMap (OWM) service, which provides global weather information through an API. OWM aggregates data from multiple sources – including global forecast models, local weather stations, satellite observations, and radar – to provide comprehensive weather metrics for any location. For the study area, I retrieved gridded weather data at approximately 1 km resolution (spacing between data points ~ 1 km) covering the entire modeling domain. The dataset included all required variables (temperature, humidity, wind speed, wind direction, and precipitation rates) at regular time intervals (hourly snapshots in this case), providing the necessary temporal coverage for air quality simulations.

The raw OWM data underwent preprocessing to ensure it aligned with the air quality model's requirements in both space and time. First, I standardized units and formats from OWM to match my model conventions (e.g., converting temperatures to degrees Celsius if provided in Kelvin, wind speeds to m/s, and accumulated rainfall to consistent rate units). Next, I synchronized the meteorological data timeline with the air quality modeling timeline. Each weather timestamp from OWM was matched to the corresponding model time step (since both were hourly, direct matching was used; if there had been any mismatch in timing, interpolation in time would be applied to fill any gaps).

To map the coarse-resolution meteorological data into the unified 100m standard grid that ensures a consistent input representation required by the ViT-based model, I applied a linear interpolation technique. Essentially, for each meteorological variable and each time step, the values from OWM's ~ 1 km grid were interpolated to estimate values at every 100 m grid cell within the domain. The detailed meteorological data collection and processing workflow is provided in Appendix B.6.

After interpolation, each 100 m grid cell has an associated meteorological profile (temperature, wind, humidity, etc.) at each hourly time step, aligned with the spatial and temporal framework of the air quality model. This integrated meteorological

dataset is then used to drive the air quality model. Table 3.3 shows the Meteorological variables used in this study.

Table 3.3 Meteorological variables used in my study

Meteorological variables	Unit	Description in OWM
Temperature	°C	Temperature in degrees Celsius (°C)
Humidity	%	A percentage of relative humidity, ranging from 0% (completely dry) to 100% (fully saturated)
Rainfall	mm	Rain volume for the last 1 hour
Wind speed	m/s	Wind speed for the last 1 hour
Wind direction	degrees	Denotes the compass direction from which the wind originates, where 0° represents north, 90° east, 180° south, and 270° west
Pressure	hPa	Atmospheric pressure on the sea level
Clouds	%	Cloudiness, the fraction of the sky covered by clouds, expressed as a percentage, with 0% representing a clear sky and 100% indicating complete cloud cover

Data Alignment and Integration

Following the acquisition of datasets described previously (including Breathe London (BL) sensor data, London Air Quality Network (LAQN) reference data, land use features, traffic variables, and meteorological data), it was necessary to align and integrate these datasets spatially and temporally before modeling air quality. This alignment process ensured consistency and comparability across different sources.

- **Spatial Alignment:** A standardized spatial grid covering Greater London (latitude: 51.28°N–51.70°N; longitude: -0.52°W–0.34°E) with a 100 m × 100 m resolution was defined. Initially, the study area was geographically registered using the WGS84 coordinate system, and the Greater London boundary shapefile was extracted. Subsequently, all BL sensors, LAQN reference stations, and auxiliary covariate data (land use, traffic, and meteorological features) were assigned to their corresponding grid cells, ensuring spatial consistency. For grid cells containing multiple measurement stations, pollutant concentrations were represented by averaging all station values within the cell.
- **Temporal Alignment:** Data from all sources were synchronized at an hourly frequency throughout the study period. Timestamps were standardized to Coordinated Universal Time (UTC+0) to maintain consistent temporal referencing across datasets.

The spatial mapping method is detailed comprehensively in Appendix A. After completing spatial mapping, each grid cell within the standardized grid contained aligned values from all relevant data sources. Finally, the spatially and temporally aligned datasets were integrated into a unified dataset at hourly intervals, structured explicitly to facilitate the subsequent development of air quality models.

Prior to model training, continuous variables were standardized using normalization as detailed below:

- **Normalization of Continuous Variables:** Continuous covariates, including meteorological parameters such as temperature and humidity, were scaled to a standard $[0, 1]$ range through min-max normalization. This scaling approach ensures that variables with inherently larger numerical ranges do not disproportionately influence the model training process. Normalization also promotes faster convergence for machine learning algorithms sensitive to feature scales. By standardizing these continuous features, the model can

effectively and appropriately weigh their contributions relative to other input variables.

These procedures resulted in a refined and consistent set of input variables. In summary, following these steps, the finalized dataset was ready to be employed for training and evaluating the proposed air quality prediction method, as described in subsequent sections.

3.2.3 Model Optimization and Validation Method

I randomly selected 20% of the PM_{2.5} and NO₂ monitoring stations in London Air Quality Network as an independent test dataset. The remaining 80% of stations were used for training the model. To evaluate model performance, I conducted a 5-fold cross-validation, partitioning the remaining LAQN stations randomly into five subsets. For each iteration, four subsets were used for training, and the fifth subset was reserved for validation. This 5-fold cross-validation yielded a robust and unbiased estimate of model generalizability. I implement the model by PyTorch 2.6 (pytorch, 2024) using a RTX 3090 GPU.

Model Optimization

To train the proposed ViT-based air quality model, I minimize the Mean Squared Error (MSE) between the model's predictions $\hat{y}_i$ and the ground-truth air quality values y_i . The MSE loss is given by:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3.2)$$

Where N is the number of training samples. This loss function penalizes large errors more than smaller ones and is well-suited for regression tasks, driving the model to predict pollutant concentrations as close as possible to observed values.

Optimizer – AdamW: I employ the AdamW optimizer (Loshchilov & Hutter, 2019) with decoupled weight decay, which offers several advantages over the standard Adam optimizer (Kingma & Ba, 2015) for transformer-based models. AdamW modifies Adam by decoupling the L_2 weight decay regularization term from the gradient-based update rule. This decoupling ensures that weight decay acts directly on parameter values (shrinking them each update) rather than conflating with the gradient of the loss. In practice, AdamW’s decoupled weight decay leads to more stable convergence and better generalization performance, allowing adaptive gradient methods to match or exceed the generalization of SGD with momentum. These benefits are especially pronounced in training transformers, which contain a large number of parameters and can easily overfit. By using AdamW, I can tune the learning rate and weight decay independently, helping ViT model avoid overfitting while still converging rapidly on the training data.

Learning Rate Schedule: Training uses a two-phase learning rate schedule combining an initial warm-up and subsequent linear decay, a strategy often adopted for transformer training (Devlin et al., 2019; Vaswani et al., 2017). During the first 5 epochs, I gradually increase the learning rate from a near-zero value to the peak rate of 5×10^{-5} . Formally, if $\eta(e)$ denotes the learning rate at epoch e , then for $e \leq 5$ (the warm-up phase) $\eta(e)$ is incremented linearly such that $\eta(5) = 5 \times 10^{-5}$. Beyond epoch 5, a linear decay schedule is applied: the learning rate is decreased each epoch so that it reaches 1×10^{-5} by epoch 50. In other words, for epoch e in $6 \leq e \leq 50$, I set

$$\eta(e) = 5 \times 10^{-5} - \frac{e - 5}{45} (5 \times 10^{-5} - 1 \times 10^{-5}) \quad (3.3)$$

This schedule starts with a cautious, low learning rate (to prevent unstable updates in early training) and then gradually reduces the learning rate as training progresses, which aids in fine-tuning the model parameters as convergence is approached. The warm-up phase is critical for transformer models – it avoids shocking the network

with large gradient steps at the start, thus stabilizing training (Popel & Bojar, 2018), while the gradual decay helps in navigating toward a local minimum when updates need to be smaller.

Regularization and Early Stopping: To further guard against overfitting, I integrated an early stopping criterion based on validation performance. Training is halted if the validation loss does not improve for **10 consecutive epochs** (patience = 10). Early stopping is a well-established regularization technique that prevents the model from *over*-optimizing on the training data to the detriment of generalization (Prechelt, 1998). By monitoring a held-out validation set and stopping when the error plateaus or worsens, I ensure that ViT model does not continue training long after it has learned the general data patterns. A patience of 10 epochs allows the model some leeway to plateau briefly and possibly recover, which is important because validation loss in deep learning can fluctuate slightly even when the overall trend is improving (Goodfellow et al., 2016). In my experiments, this early stopping strategy typically stopped training well before the maximum of 50 epochs, once the model’s capacity to generalize started to diminish. This not only mitigated overfitting but also reduced unnecessary computation.

Throughout training, I tracked metrics such as training loss, validation loss, and (if relevant) accuracy or mean absolute error on the validation set at each epoch. I saved model checkpoints at the epoch with the lowest validation loss. This checkpoint was later used for final evaluation on the test set.

Ablation Study

I conducted an ablation study to assess how incorporating low-cost sensor data and sensor calibration affects the performance of the ViT deep learning model for air quality prediction. In this experiment, ViT was trained to predict urban **PM_{2.5}** and **NO₂** concentrations (at hourly 100m spatial resolution) under three different data configurations, while keeping the model architecture and training procedure consistent. The three configurations were:

1. **Reference-only (Ref-only):** training and evaluating the model using only data from LAQN reference monitors, without incorporating any low-cost Breathe London measurements.
2. **Raw Breathe London + Reference (BL_raw+Ref):** using both raw (uncalibrated) Breathe London low-cost sensor measurements and LAQN reference data.
3. **Calibrated Breathe London + Reference (BL_calib+Ref):** using calibrated Breathe London sensor measurements together with LAQN reference data.

All other factors (auxiliary covariates, model structure, hyperparameters, and training epochs) were identical across these runs, isolating the effect of the dataset differences. Model performance was evaluated with the Pearson correlation coefficient (r) to measure how well predicted concentration patterns match observed variations, and the root mean square error (RMSE) to quantify absolute prediction error (in pollutant concentration units, e.g. $\mu\text{g}/\text{m}^3$).

3.2.4 Comparison Models: Land Use Regression, Random Forest, and Convolutional Neural Networks

This study benchmarks the estimation performance of my proposed ViT model against traditional machine learning methods such as RF, LUR and CNN.

Land Use Regression (LUR)

The LUR model was implemented as a multiple linear regression using the auxiliary covariates (same as that of ViT model, quantify surrounding environment) derived for each site. In my implementation, I generate a 100m-by-100m grid cell centered on each measurement site and compute features (as described in the previous dataset and preprocessing section) within that cell. Previous LUR studies often employ multiple buffer sizes around monitoring locations (for example, summarizing road lengths within radius of 100 m, 300 m, 500 m, etc.) to determine which scale of

influence best correlates with pollution levels. However, my study does not utilize multiple buffer sizes, as my aim is not to test numerous buffer configurations in search of an optimal LUR model. Instead, I fixed a 100m gird buffer for all sites and focused on comparing different modeling approaches under consistent conditions.

Random Forest (RF)

The random forest model was trained on the exact same set of predictor variables used in the LUR model, allowing comparison between the linear regression and a non-linear ensemble approach. I used a standard random forest regression algorithm, which constructs an ensemble of decision trees on bootstrapped samples of the training data. Key hyperparameters (such as the number of trees and the maximum tree depth) were tuned on the training set to optimize performance while avoiding overfitting. By using identical inputs and a consistent tuning process, any performance differences between RF and LUR can be attributed to the modeling technique rather than differences in data representation.

Convolutional Neural Network (CNN)

I also implemented CNN as a deep learning baseline to compare against the ViT model. The CNN model was trained on the exact same image-like grid data used in the ViT model, which introduced in 3.2.2. The CNN architecture consisted of multiple convolutional layers (to automatically extract local features from the input grid), interleaved with pooling layers, and followed by fully connected layers to produce the final pollution estimate. I trained the CNN using the same data and procedures as the ViT model (e.g. using the AdamW optimizer for a fixed number of epochs), thereby ensuring a consistent training regime for both deep learning models.

3.3 Experiments and Results

The results, summarized in Table 3.4, show a clear pattern: integrating raw low-cost sensor data without calibration can degrade air quality model performance, whereas integrating calibrated low-cost sensor data improves performance significantly.

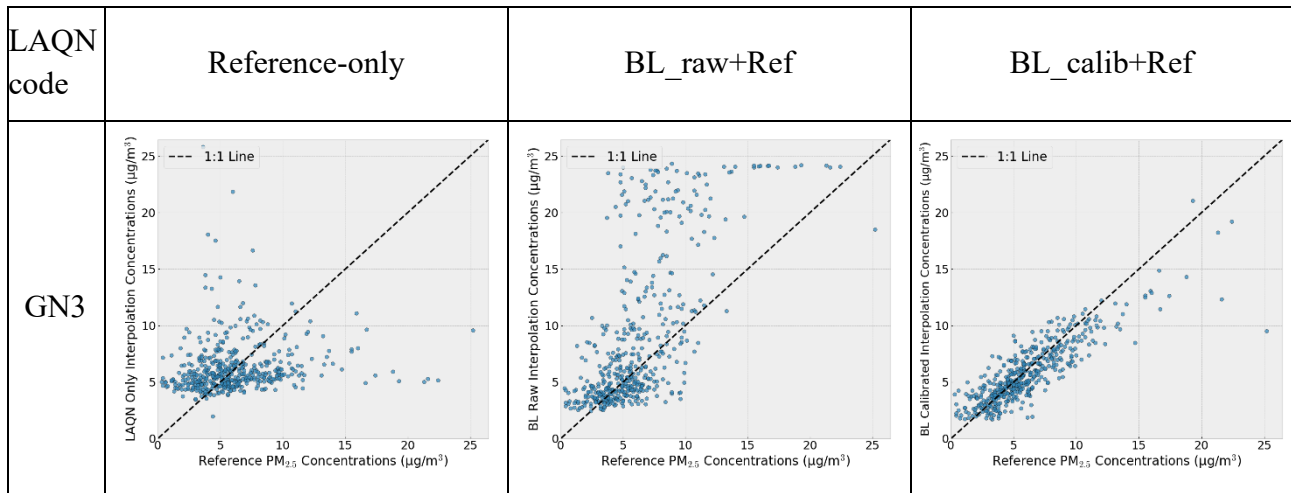
Table 3.4 The NO₂ and PM_{2.5} performance of proposed ViT-based model for each configuration (Reference-only, Raw Breathe London + Reference, Calibrated Breathe London + Reference)

Pollution	Model	Data	Pearson Correlation	RMSE
PM _{2.5}	ViT	Ref	0.4319	4.7479
	ViT	BL_raw+Ref	0.6780	6.4457
	ViT	BL_calib+Ref	0.7992	3.3132
NO ₂	ViT	Ref	0.3182	12.8437
	ViT	BL_raw+Ref	0.4777	13.2247
	ViT	BL_calib+Ref	0.7001	10.8870

For the baseline Ref-only configuration, the model achieved a Pearson correlation coefficient (r) of 0.43 for PM_{2.5} and 0.32 for NO₂, with corresponding RMSE values of 4.75 and 12.84, respectively. When raw low-cost sensor data were added (BL_raw+Ref), while the model captured pollutant variability better (as shown by higher correlations), the predictive accuracy declined, as evidenced by higher RMSE. In fact, including uncalibrated low-cost sensor data raised the PM_{2.5} correlation to $r = 0.68$ (approximately 57% increase) and the NO₂ correlation to $r = 0.48$ (around 50% increase), compared to the Ref-only baseline. However, this improvement came at the cost of higher RMSE values: PM_{2.5} RMSE increased to 6.45 (~36% higher than Ref-only), and NO₂ RMSE marginally increased to 13.22 (~ 3% higher than Ref-only). In other words, the raw sensor data introduced additional noise or bias that degraded the model's accuracy – the predictions deviated more from the true values even though they tracked the fluctuations better. This result suggests that without calibration, the low-cost sensor measurements contain errors large enough to outweigh their benefits, leading to worse overall performance than using only high-quality reference data.

By contrast, integrating calibrated low-cost sensor data with reference data (BL_calib+Ref) resulted in substantial improvements in both accuracy and correlation for PM_{2.5} and NO₂, attaining a PM_{2.5} correlation of $r = 0.80$ and an NO₂ correlation of $r = 0.70$ – improvements of approximately 80% (PM_{2.5}) and 100% (NO₂) relative to the Reference-only baseline. Concurrently, the prediction error significantly decreased: PM_{2.5} RMSE was reduced to 3.31 (~ 29% lower than Ref-only baseline), and NO₂ RMSE fell to 10.89 (~16% lower than Ref-only baseline). Notably, BL_calib+Ref not only outperformed the Ref-only baseline, but also dramatically outperformed BL_raw+Ref, reducing the PM_{2.5} RMSE from 6.45 to 3.31 (a nearly 48% reduction compared to using raw low-cost sensor data), while also substantially improving the correlation. This highlights that calibration effectively unlocks the value of the low-cost sensor data: once biases and systematic errors are corrected, the low-cost sensor’s measurements can be reliably combined with reference measurements, leading to more accurate and consistent predictions.

Figures 3.6 and 3.7 present scatter plots of predicted vs. observed PM_{2.5} and NO₂ at four randomly selected LAQN stations under three configurations: (1) Reference-only, (2) Raw BL + Ref, and (3) Calibrated BL + Ref.



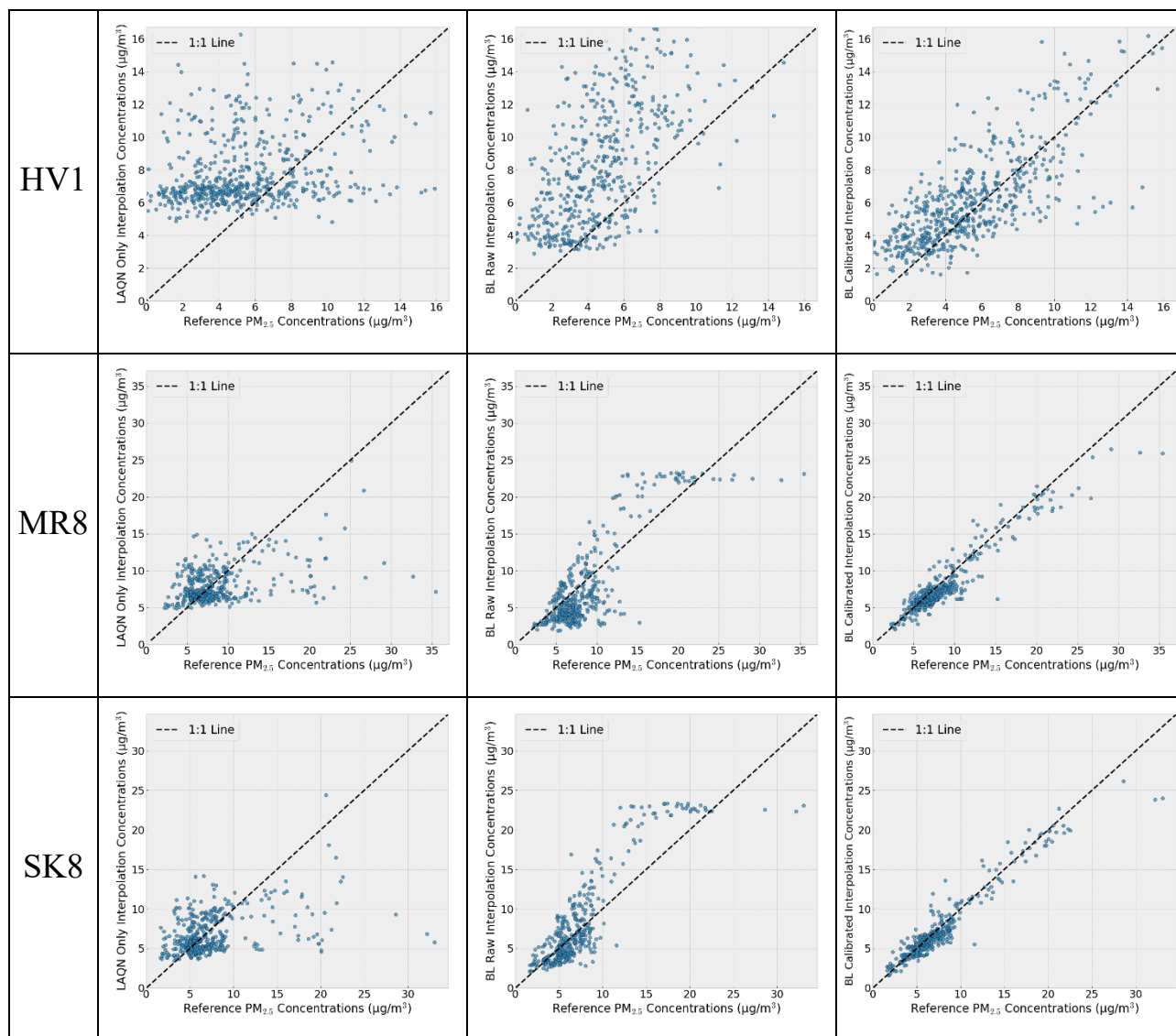


Figure 3.6 Scatter plots of ViT predicted vs. observed $PM_{2.5}$ concentrations at four randomly selected London Air Quality Network stations: GN3, HV1, MR8, and SK8. The three columns represent different data configurations: (1) Reference-only (left), (2) Raw Breathe London + Reference (middle), (3) Calibrated Breathe London + Reference (right).

LAQN Code	Reference-only	BL_raw+ Reference	BL_calib+Ref
-----------	----------------	-------------------	--------------

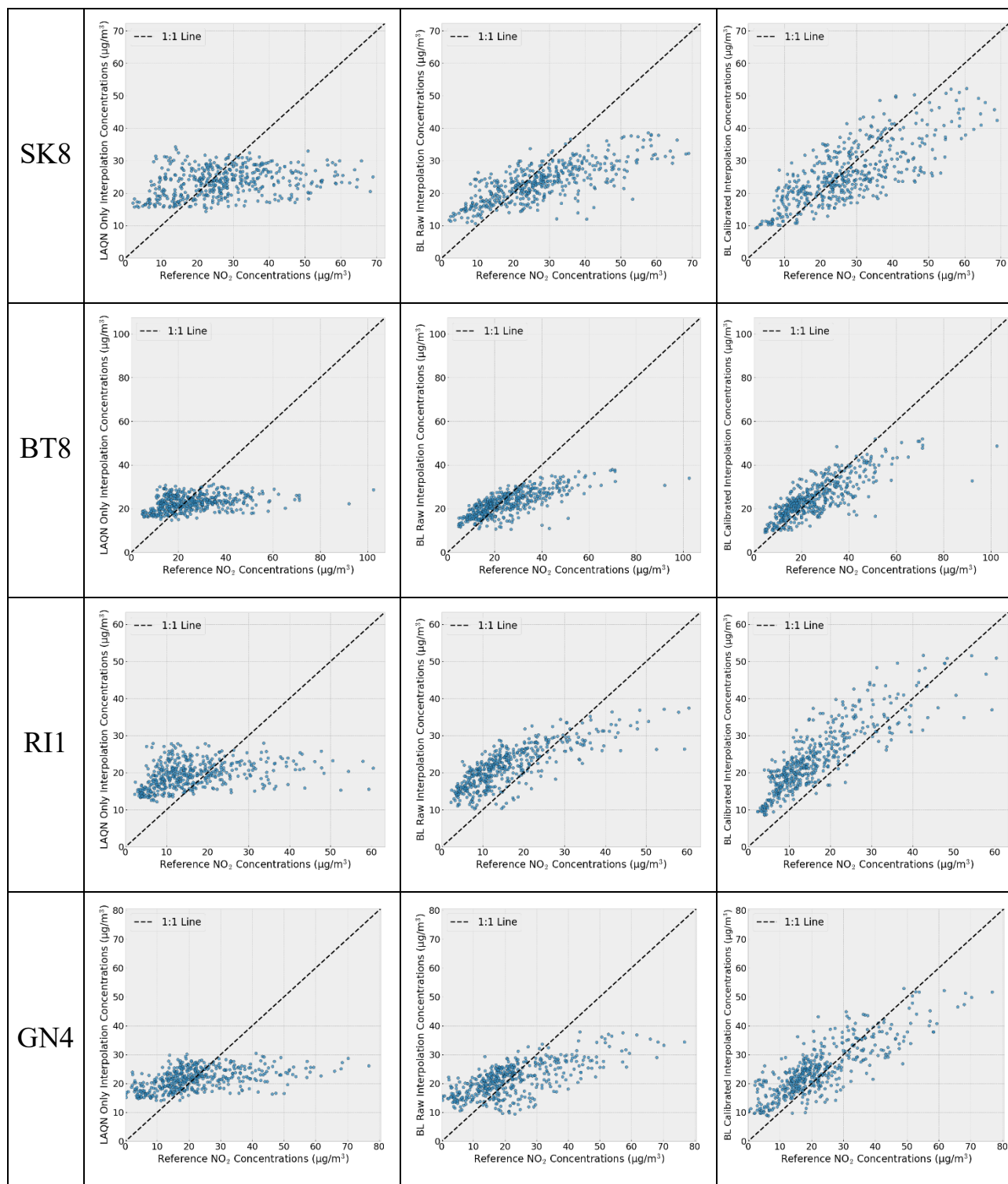


Figure 3.7 Scatter plots of ViT predicted vs. observed NO₂ concentrations at four randomly selected London Air Quality Network stations: SK8, BT8, RI1, and GN4. The three columns represent different data configurations: (1) Reference-only (left), (2) Raw Breathe London + Reference (middle), (3) Calibrated Breathe London + Reference (right).

These quantitative findings are illustrated by the 45-degree scatter plots of predicted vs. reference concentrations for each configuration (with the 45° line representing perfect agreement). In the Ref-only scatter plots, the data points show considerable spread around the 45° line, indicating large deviations between predicted and reference values. Many points fall off the ideal line, consistent with the relatively low correlation and higher RMSE – the model captures some trend but with significant error.

When adding raw sensor data (BL_raw+Ref), the scatter plot points begin to align more along a linear trend, reflecting the improved correlation (the model's predictions vary more closely with the reference values). However, these points do not tightly cluster on the 45° line – instead, they may appear systematically offset or more widely dispersed around the line. This misalignment visualizes the model's bias and inaccuracy when using uncalibrated data, in line with the increased RMSE.

Finally, under the BL_calib+Ref configuration, the scatter plot points lie much closer to the 45° diagonal with markedly less scatter. The predictions match the reference observations across the concentration range, forming a tight cloud along the line of perfect agreement. This close alignment implies that most predicted values are very near their corresponding observed values, corroborating the high Pearson r and low RMSE achieved with calibrated sensor inputs.

In summary, the progression of these scatter plots (from Ref-only to raw Breathe London data to calibrated Breathe London data) vividly demonstrates the ViT model's performance gains: the initially diffuse and biased cloud of points (Ref-only) becomes somewhat more trend-aligned but still error-prone with raw sensor data, and ultimately converges toward the ideal 45° line once the sensor data are calibrated.

Comparison of different models

Table 3.5 presents the performance of different models on both PM_{2.5} and NO₂. The proposed ViT model outperforms both Random Forest and Land use regression, achieving higher r and lower RMSE.

Table 3.5 Air quality model performance for PM_{2.5} and NO₂

Pollution	Model type	Pearson Correlation	RMSE
PM _{2.5}	LUR	0.7491	3.8171
	RF	0.7444	3.6519
	CNN	0.7766	3.6116
	ViT	0.7992	3.3132
NO ₂	LUR	0.6342	12.0773
	RF	0.6682	11.9267
	CNN	0.6762	11.5768
	ViT	0.7001	10.8870

Spatiotemporal Patterns

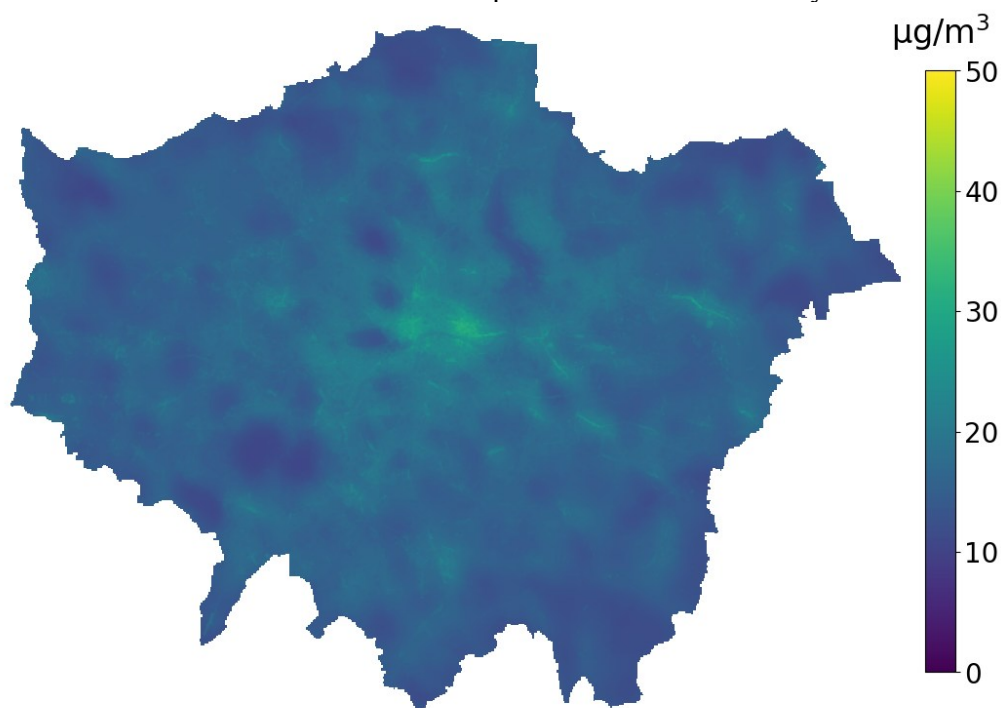
The proposed ViT model provides fine-scale spatiotemporal variation concentration fields. I illustrate the ViT model's spatiotemporal air quality predictions for a representative winter weekday (February 14, 2024).

NO₂ maps estimation

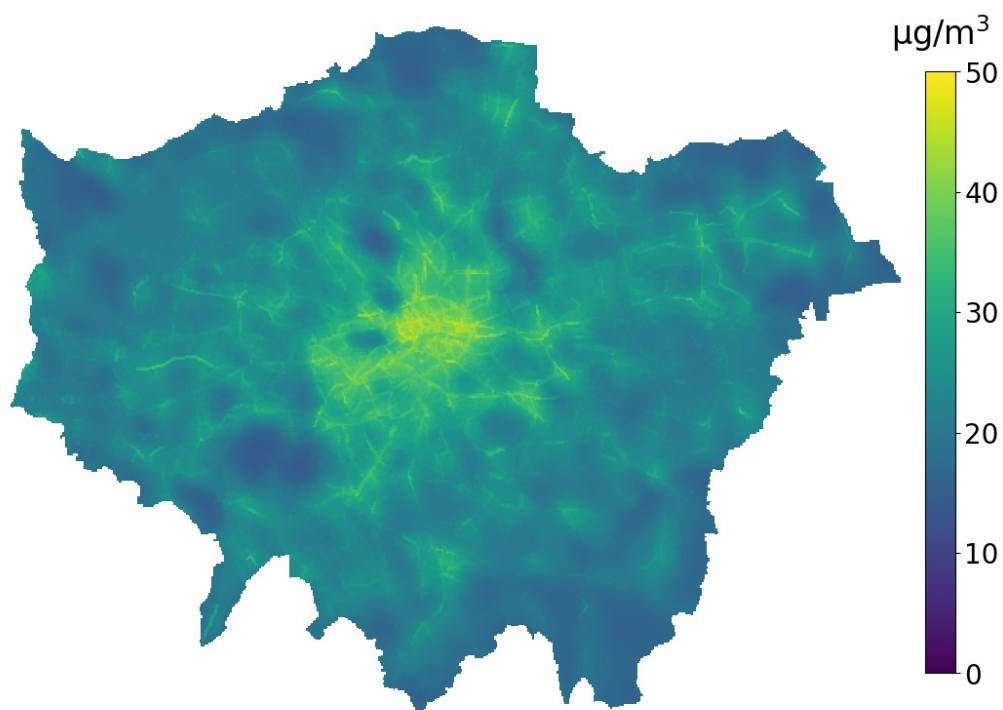
Figure 3.8 shows the spatial distribution of NO₂ at 6 AM, 8 AM, 12 PM, 6 PM, 8 PM, and 10 PM on 14 February 2024. Higher concentrations (indicated by green) are concentrated along major roadways during the rush hour (from 8 AM to 6 PM), while lower concentrations (indicated by blue) prevail in outlying residential and rural areas. The map illustrates sharp spatial gradients in NO₂: traffic-heavy zones exhibit significantly elevated levels relative to the background. After 6 PM, NO₂ concentrations are predicted to decline due to the combined effects of reduced traffic emissions. After 10 PM, much of the urban area experiences lower NO₂ levels,

approaching near-background concentrations except in the dense urban center. This diurnal pattern underscores the strong temporal dynamics of NO₂ driven by traffic patterns. Overall, the ViT model's NO₂ predictions demonstrate how this pollutant's concentration patterns closely follow vehicular emission activity, leading to sharp temporal fluctuations and highly localized spatial hotspots in the city. Detailed hourly results for NO₂ are provided in Appendix D.

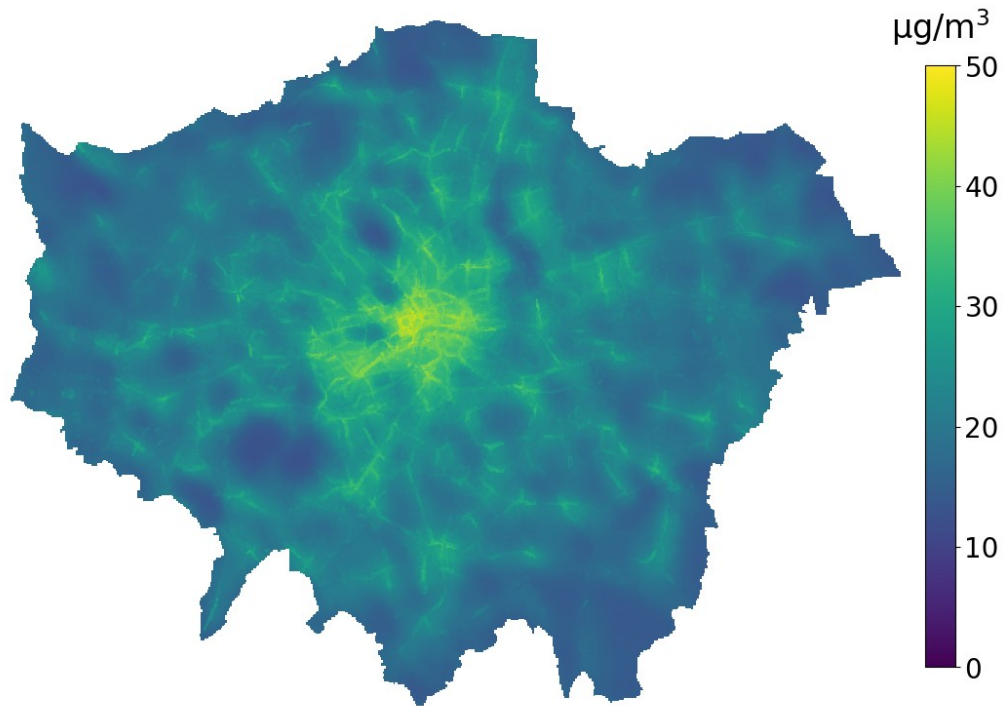
ViT-Based Estimated NO₂ Map at 6 AM on 14 February 2024



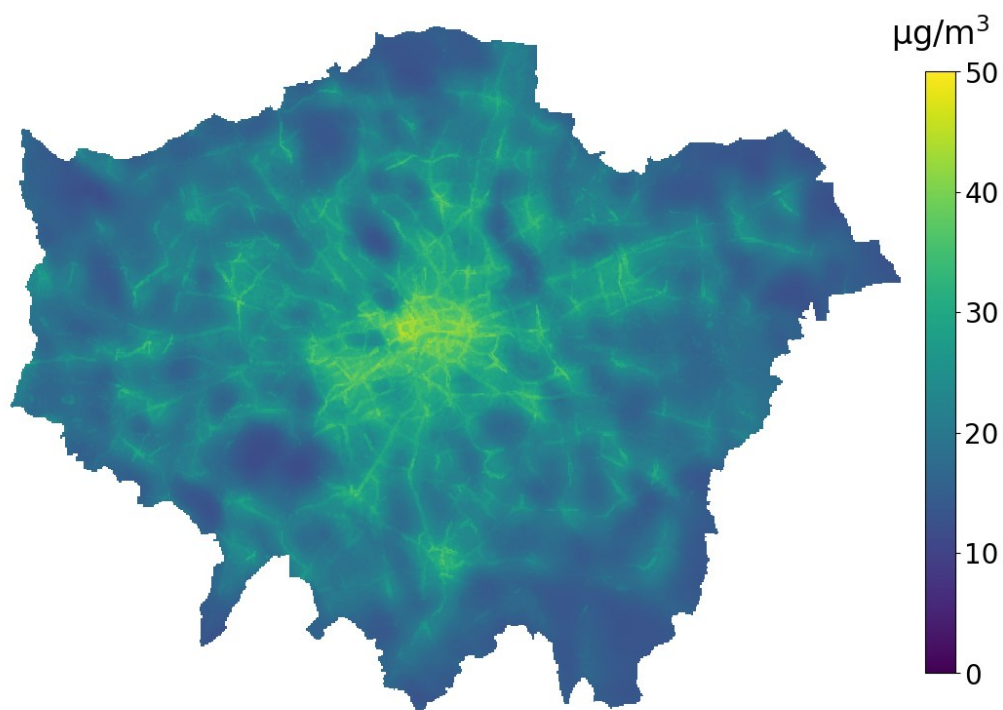
ViT-Based Estimated NO₂ Map at 8 AM on 14 February 2024



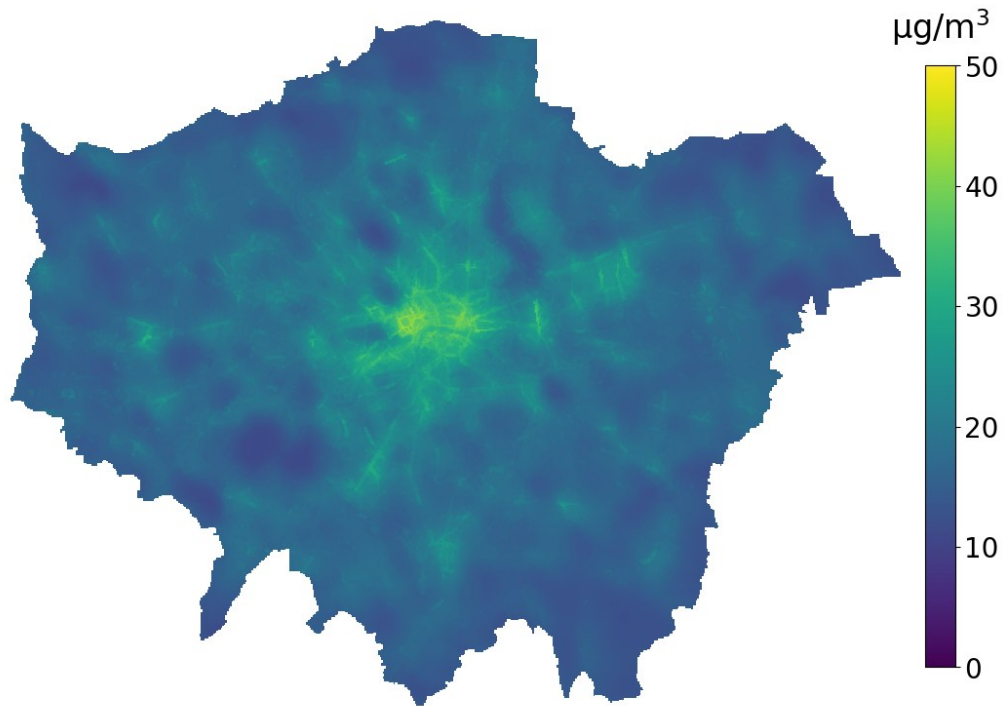
ViT-Based Estimated NO₂ Map at 12 PM on 14 February 2024



ViT-Based Estimated NO₂ Map at 6 PM on 14 February 2024



ViT-Based Estimated NO2 Map at 8 PM on 14 February 2024



ViT-Based Estimated NO2 Map at 11 PM on 14 February 2024

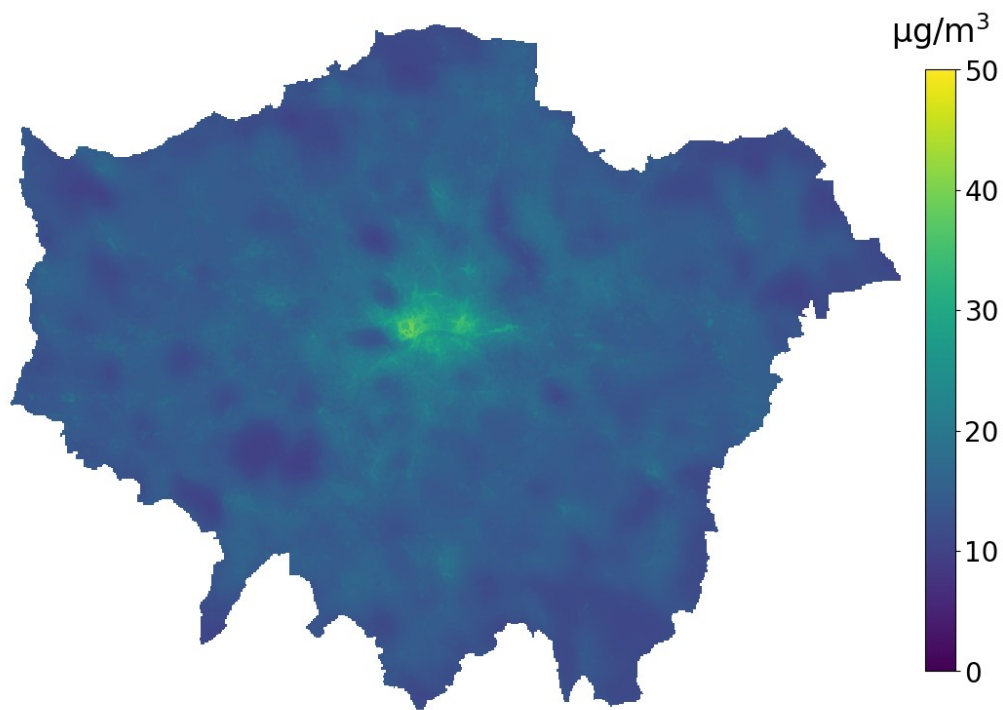


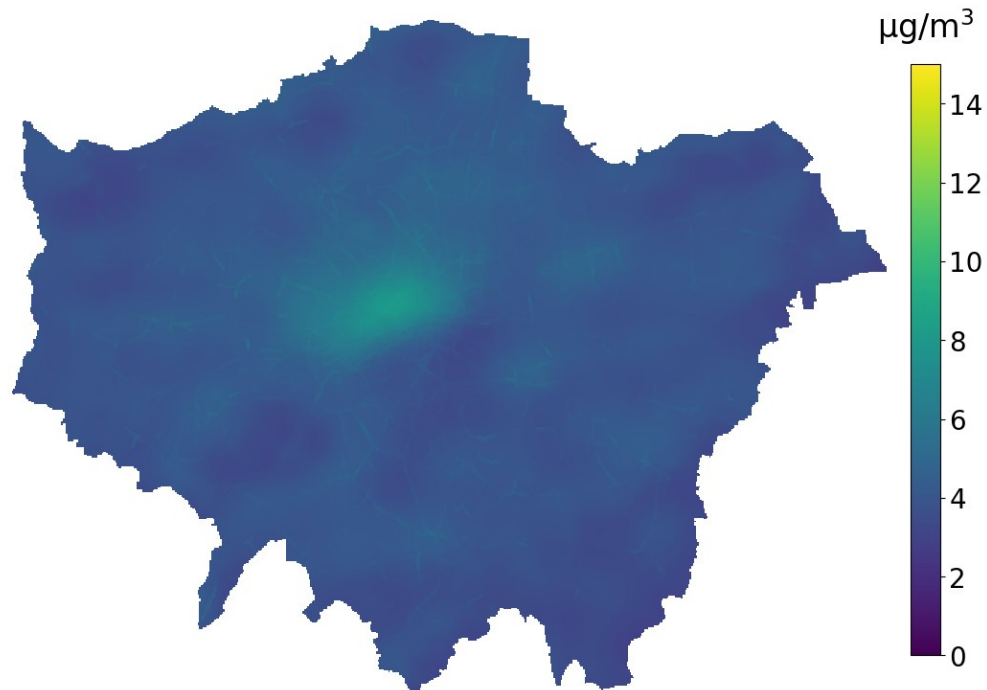
Figure 3.8 Estimated NO₂ maps from the ViT model at 6 AM, 8 AM, 12 PM, 6 PM, 8 PM, and 11 PM on 14 February 2024.

PM_{2.5} maps estimation

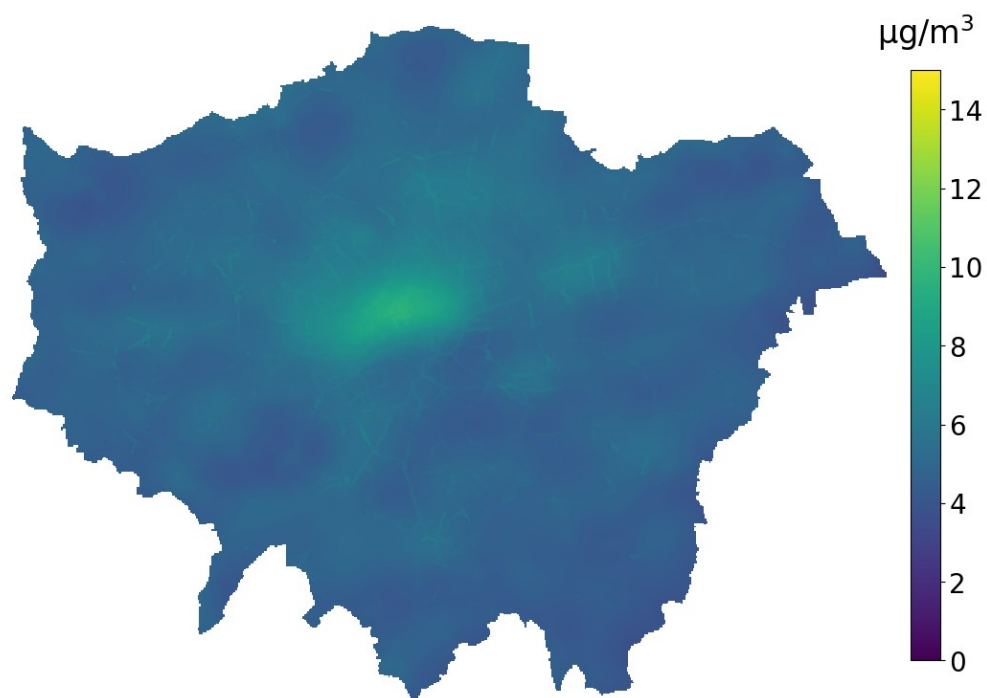
ViT-Based Estimated PM_{2.5} Map at 8 AM on 14 February 2024



ViT-Based Estimated PM2.5 Map at 12 PM on 14 February 2024



ViT-Based Estimated PM2.5 Map at 7 PM on 14 February 2024



ViT-Based Estimated PM2.5 Map at 11 PM on 14 February 2024



Figure 3.9 Estimated PM_{2.5} maps from the ViT model at 8 AM, 12 PM, 7 PM, 11 PM on 14 February 2024.

Figure 3.9 illustrates PM_{2.5} distribution across London at 8 AM, 12 PM, 7 PM, 11 PM on February 14, 2024. Compared to NO₂, the diurnal cycle of PM_{2.5} is flatter, whereby peak-to-off-peak differences in concentration are smaller, and its spatial distribution is more uniform across the city. This contrasting behavior reflects the distinct emission sources and atmospheric lifetimes of the two pollutants. PM_{2.5} originates from multiple source types and can remain suspended or be transported over longer timescales, whereas NO₂ is closely tied to immediate vehicular emissions near roadways. The ViT model’s outputs clearly capture these differences: in the model predictions, PM_{2.5} forms a more regionally dispersed haze, whereas NO₂ appears as sharply defined plumes around traffic-intensive locations. In summary, while traffic influences both pollutants’ levels, NO₂ pollution is shown to be highly temporally peaked and spatially concentrated, whereas PM_{2.5} pollution is more spatially pervasive with more gradual temporal variations.

3.4 Discussion

In this study, I propose a high-resolution urban air quality mapping approach based on Vision Transformer architecture. Leveraging dense observations from over 400 calibrated Breathe London sensor nodes distributed across the city, along with auxiliary covariates such as land use, traffic congestion, and meteorological variables, the proposed ViT-based model produces citywide air quality estimates at a fine spatiotemporal resolution of $100\text{ m} \times 100\text{ m} \times 1\text{ hour}$. This resolution reflects a balance between computational feasibility, data availability, and practical management needs. From a computational perspective, producing hourly predictions across all of Greater London at resolutions finer would substantially increase GPU memory requirements and training time, particularly for vision transformer-based architectures whose self-attention mechanism scales quadratically with the number of spatial tokens. From a data perspective, many covariates used in the model (e.g., meteorological data) inherently represent aggregated conditions rather than micro-

scale variation, meaning that modelling at substantially finer resolutions would not necessarily yield additional meaningful information. A 100 m grid also provides sufficient spatial granularity for policy and management applications. This resolution is capable of capturing major roadside–background contrasts, neighbourhood-scale hotspots, and local exposure differences that are relevant for urban interventions such as traffic management, school-zone protection, and environmental justice assessments. As such, the 100 m resolution offers a practical and scientifically appropriate compromise between model performance, data realism, and policy relevance.

My ViT model achieved high agreement with ground-truth measurements – for example, Pearson correlations around 0.80 for $\text{PM}_{2.5}$ and 0.70 for NO_2 – outperforming conventional modeling approaches. This study demonstrates that combining a calibrated LCS network with an advanced ViT-based model can overcome the limitations of sparse reference monitoring and provide more reliable and granular insights into urban air quality.

Major Contributions

This work makes two significant contributions to the field of urban air pollution mapping:

- **Integrating Low-Cost Sensors:** I demonstrate the significant advantage of incorporating dense networks of low-cost sensors to enhance fine-scale air quality estimations. Specifically, after integrating data from the Breathe London sensor network, the correlation (Pearson's r) with reference-grade monitors improved markedly, increasing from 0.43 to 0.80 for $\text{PM}_{2.5}$ and from 0.31 to 0.70 for NO_2 . Such pronounced improvements clearly validate the utility of dense LCS deployments. Crucially, I also empirically substantiate the necessity of rigorous sensor calibration. Prior to modeling, all $\text{PM}_{2.5}$ and NO_2 sensors underwent systematic calibration against reference-grade monitors, mitigating inherent drift and measurement variability characteristics

of low-cost devices. This calibration step proved essential for ensuring reliable model inputs; uncalibrated low-cost sensor readings, if used naively as shown in my study, would introduce substantial errors even as they increased the apparent correlation. This aligns with DeSouza et al. (2022) that low-cost sensor measurements “need adequate calibration to obtain good quality”. My findings thus underscore low-cost sensor calibration as a prerequisite for accurate, high-fidelity air quality mapping.

- **ViT-Based Attention Framework for Air Quality Mapping:** I developed a novel air quality mapping framework based on Vision Transformer architecture, which is, to my knowledge, the first application of ViT in high-resolution air quality modeling. The ViT framework treats the city as a sequence of spatial “grid” that encode various features (such as sensor readings, land-use data, meteorological information, and traffic conditions) and uses multi-head attention to effectively integrate these heterogeneous inputs by capturing long-range spatial dependencies and learning complex relationships among them. This approach moves beyond the localized receptive fields of convolutional neural networks and the simplistic linear assumptions of traditional regression models. The ViT-based framework adeptly accommodates complex urban heterogeneity and can be transferred to urban environments beyond London.

Performance Evaluation and Comparison with Traditional Models

The proposed ViT model achieved strong agreement with reference measurements. For $\text{PM}_{2.5}$, my model attained a Pearson correlation coefficient of 0.80 and a root mean square error (RMSE) of $3.31 \mu\text{g}/\text{m}^3$ between the predicted and reference concentrations. For NO_2 , the model yielded a Pearson correlation of 0.70 with an RMSE of $10.89 \mu\text{g}/\text{m}^3$. The slightly lower correlation for NO_2 likely reflects the greater spatial heterogeneity of NO_2 sources (e.g., highly localized traffic emissions) and the challenges of measuring NO_2 with low-cost sensors (Apte et al., 2017a), yet the model still captures a majority of the variance in NO_2 levels.

Crucially, my ViT-based approach outperformed traditional models, including land-use regression (LUR), Random Forest (RF), and a convolutional neural network (CNN) baseline. The LUR model I developed was well established and achieved performance metrics comparable to those reported in previous studies. For example, my LUR model yielded an RMSE of $3.8 \mu\text{g}/\text{m}^3$ for $\text{PM}_{2.5}$ and $12.1 \mu\text{g}/\text{m}^3$ for NO_2 , which closely aligns with the values reported by Wang et al. (2022) in Greater London (RMSE of $5.0 \mu\text{g}/\text{m}^3$ for $\text{PM}_{2.5}$ and $12.7 \mu\text{g}/\text{m}^3$ for NO_2). These similarities indicate that my LUR implementation is representative of the typical performance of such models in urban environments. The drawback of LUR models, which typically rely on linear regression, have inherent limitations in fully exploiting nonlinear relationships among pollutants and predictor variables. Random Forest model, which can capture non-linear effects, performed slightly better than LUR, consistent with findings that incorporating machine learning can improve predictive over linear models (Jain et al., 2021b). Nonetheless, the RF lacked a mechanism to account for spatial structure explicitly – it treated spatial coordinates or land-use features independently and fall short when confronted with persistent issues such as spatial heterogeneity, multi-source data integration, and the need to account for spatial dependencies. The CNN baseline, though adept at recognizing local spatial patterns, was limited by its fixed receptive field (Luo et al., 2016). My ViT-based approach addresses these shortcomings by leveraging multi-head attention mechanism that enables the model to learn intricate spatial relationships across the entire urban grid. This allows it to effectively integrate heterogeneous data sources (such as land use, traffic, meteorology, and sensor readings) while simultaneously capturing both local and long-range spatial dependencies to generate more accurate estimates of air quality.

Temporal and Spatial Correlation in ViT-based Air Quality Modelling

As shown in Figure 3.8 and 3.9, $\text{PM}_{2.5}$ and NO_2 exhibit temporal autocorrelation driven by recurring emission patterns (e.g., diurnal traffic cycles), meteorological persistence, and chemical lifetime, as well as pronounced spatial correlation arising from emission source proximity, urban morphology, and atmospheric transport.

Temporal correlation in $\text{PM}_{2.5}$ and NO_2 is primarily governed by short-term emission regularities and meteorological inertia, resulting in smooth hour-to-hour evolution at a given location. In contrast, spatial correlation reflects the heterogeneous structure of the urban environment. NO_2 , in particular, exhibits sharp spatial gradients due to localized traffic emissions and street-canyon effects, whereas $\text{PM}_{2.5}$ tends to display smoother spatial fields influenced by regional background and secondary formation. Capturing such spatial dependencies requires modelling multi-scale interactions across the urban domain (ranging from local to regional transport), rather than relying solely on local neighbourhood information. Traditional convolutional architectures are constrained by fixed receptive fields and therefore struggle to represent these non-local spatial relationships. The ViT framework addresses this limitation by treating the urban grid as a sequence of spatial tokens and applying global self-attention. This allows the model to dynamically weigh the importance of both immediate neighbours and distant spatial drivers, effectively integrating heterogeneous inputs such as land use, road density, and meteorological fields to capture the complex spatial hierarchies of urban pollution.

Sensitivity Analysis of Input Variables

To identify which input variables are most influential in explaining model behaviour, I conducted a leave-one-feature-out sensitivity analysis in which the ViT-based model was retrained repeatedly, each time excluding a single input variable while keeping all other settings unchanged. Changes in RMSE and correlation relative to the full-feature baseline were used as indicators of each variable’s marginal contribution to predictive performance.

Table 3.6 Sensitivity of ViT-based NO_2 and $\text{PM}_{2.5}$ predictions to input variables. Only variables with $|\Delta\text{RMSE}| \geq 0.1$ are reported.

Pollution	Missing Variable	$\Delta\text{Correlation}$	ΔRMSE
$\text{PM}_{2.5}$	Calibrated Breathe London	−0.36	1.43
	Wind direction	−0.06	0.6
	Cloud cover	−0.06	0.47

	Humidity	−0.05	0.39
	Building density	−0.04	0.31
	Temperature	−0.03	0.26
	Rainfall	−0.03	0.22
	Green	−0.03	0.2
	Traffic	−0.03	0.18
	Wind speed	−0.01	0.15
NO ₂	Calibrated Breathe London	−0.38	1.95
	Traffic	−0.03	0.20

* For reference, the ViT model achieved a correlation of 0.7992 (RMSE = 3.3132) for PM_{2.5} and 0.7001 (RMSE = 10.8870) for NO₂ (see Table 3.5)

Table 3.6 presents the results of a systematic leave-one-feature-out sensitivity analysis conducted to identify which input variables are dominant in explaining the behaviour of the ViT-based air quality model. Variables are ranked by the magnitude of performance degradation relative to the full-feature model, and only features with $|\Delta\text{RMSE}| \geq 0.1$ are reported. By retraining the model while excluding one variable at a time and evaluating changes in RMSE and correlation relative to the full-feature baseline, this analysis provides a robust assessment of each variable's marginal contribution to predictive performance.

Across both pollutants, the most pronounced performance degradation occurs when calibrated low-cost sensor observations are removed. Beyond sensor observations, the sensitivity patterns differ markedly between pollutants, reflecting their distinct physical and chemical drivers. For NO₂, traffic-related variable emerges as the most influential auxiliary input after sensor data ($\Delta\text{RMSE} = 0.20$), consistent with the dominant role of road traffic emissions in shaping NO₂ concentrations at fine spatial scales. In contrast, PM_{2.5} predictions exhibit pronounced sensitivity to meteorological variables, particularly wind direction ($\Delta\text{RMSE} = 0.60$), cloud cover ($\Delta\text{RMSE} = 0.47$), and relative humidity ($\Delta\text{RMSE} = 0.39$). This pattern reflects the strong influence of regional transport, atmospheric stability, and nonlinear secondary aerosol formation processes on PM_{2.5} variability. The comparatively smaller impact of traffic variables on PM_{2.5} further highlights the mixed local–

regional nature of particulate matter, in contrast to the strongly localised emission structure of NO₂.

Overall, the sensitivity analysis confirms that the predictive behaviour of the proposed framework is driven by a small number of dominant variables, with calibrated sensor observations playing a central role and auxiliary inputs exerting pollutant-specific secondary influences. This clarity in variable importance strengthens confidence in both the interpretability and transferability of the modelling framework.

Limitations and Practical Applicability of the Proposed Model

Despite the strong predictive performance demonstrated in this chapter, several limitations and potential failure modes must be acknowledged when considering real-world deployment of the proposed ViT-based framework. First, the model's performance is fundamentally contingent on the availability of dense, well-calibrated low-cost sensor observations. As demonstrated by the sensitivity analysis, removing calibrated Breathe London observations lead to the largest degradation in predictive accuracy for both PM_{2.5} and NO₂, indicating that the framework cannot fully compensate for the absence of direct observational constraints using auxiliary variables alone. In cities with sparse low-cost sensor coverage or limited opportunities for reference co-location, model performance is therefore expected to degrade, particularly at fine spatial scales.

Second, the framework implicitly assumes that key auxiliary inputs (particularly traffic and meteorological variables) are available at sufficient spatial and temporal resolution. Sensitivity experiments indicate that NO₂ predictions are particularly reliant on traffic-related features, while PM_{2.5} predictions are more sensitive to meteorological variables such as wind direction, humidity, and cloud cover. In settings where these data streams are unavailable, unreliable, or highly aggregated, predictive accuracy may decline, and uncertainty is likely to increase. This

highlights the importance of aligning model complexity with realistic data availability in prospective deployment contexts.

Third, although the proposed model produces high-resolution concentration fields, it does not explicitly resolve microscale physical processes, such as street-canyon flow dynamics. While some street-canyon-related effects may be partially captured via proxy variables, such as road network density, traffic-related indicators, land-use characteristics, these inputs are unlikely to fully represent the complex geometry and flow dynamics of street canyons. Meteorological variables, particularly wind speed and wind direction, are also included and may help account for variations in ventilation, pollutant accumulation, and downwind transport that influence street-level concentrations. Nevertheless, the absence of explicit geometric representation limits the model's ability to fully capture microscale dispersion processes. Future work could address this limitation by incorporating dedicated morphological descriptors of the built environment, such as street width-to-building height ratios, to more directly characterise street-canyon effects.

Implications for Urban Air Quality Management and Public Health

The implications of this work extend beyond model performance metrics, offering actionable insights for urban planning, environmental policy, and public health. First, a key contribution lies in the ability to generate accurate, high-resolution pollution surfaces at **100 m spatial and hourly temporal resolution**, which is an order of magnitude finer than most conventional models that typically operate at 500–1000 m grids. For instance, De la Cruz (2024) applied ensemble model (RF/XGB) to map daily PM_{2.5} and NO₂ at 1km resolution over Great Britain. Even at 500 m resolution, high-resolution pollution maps have shown value in identifying intra-urban hotspots and supporting air quality interventions (Z. Li et al., 2021); by increasing the resolution an order of magnitude, my approach provides an even more refined evidence base for urban air quality management. City authorities can use the 100 m resolution pollution maps to pinpoint when and where pollutant levels are highest – down to specific neighborhoods, road segments or even street intersections. This

enables more targeted and location-specific interventions than those possible with coarse data. For example, the map reveals a persistent NO₂ hotspot along a particular traffic corridor during morning rush hour, traffic management measures (such as rerouting vehicles, adjusting signal timings, or restricting heavy-duty trucks during peak times) could be implemented in that micro-area to reduce emissions. Similarly, urban planners could identify street canyons or high-emission zones and install localized solutions such as green barriers or enhanced ventilation. The high temporal resolution (hourly updates) means the impact of such interventions can be monitored in near-real-time. This level of granularity moves beyond traditional citywide averages and helps ensure that efforts are focused on where they will have the greatest effect.

Second, my approach demonstrates a practical pathway to cost-effective environmental monitoring. By using a network of low-cost sensors (made reliable through calibration) in combination with a powerful AI model, cities can achieve detailed pollution mapping. High-resolution pollution maps have direct benefits for public health by enabling more accurate exposure assessment. Rather than assuming that air quality is uniform across a city or within large zones, I can now estimate what people are breathing at the scale of their home, school, or workplace. Epidemiological studies have long been limited by exposure misclassification due to reliance on sparsely located monitors that do not represent population-level variability. My 100 m × 100 m × 1 h maps help bridge that gap. Recent research has noted that leveraging dense sensor networks to produce time-resolved concentration surfaces can significantly improve estimates of personal exposure (Jain et al., 2021b). With my model, health agencies could identify micro-environments with extreme pollution levels (for example, a housing block next to a busy highway may experience far higher NO₂ than the city average) and correlate these with health outcome data. This opens the door to more precise studies of pollution-related health effects and better-targeted public health advisories.

In summary, this study introduces an innovative ViT-based framework for urban air quality mapping that advances both methodological capabilities and practical

applicability of pollution monitoring. I have shown that careful sensor calibration coupled with a cutting-edge attention-based model yields superior mapping performance than those of traditional data-driven methods. The novel combination of dense sensor networks and vision transformers holds great promise for smarter cities: it empowers stakeholders with actionable, fine-scale information to combat urban air pollution and protect public health. For example, such maps can inform dynamic traffic management strategies during peak pollution hours, or support targeted health advisories for vulnerable populations. As urban environments continue to grow and change, such AI-driven, high-resolution mapping tools will be increasingly invaluable for creating sustainable and healthy cities.

The ViT-based modelling framework introduced in this chapter demonstrates strong capability in capturing complex spatial patterns and temporal dynamics in urban air pollution. While these results highlight the predictive strengths of the model, they also underscore the need to quantify the uncertainty associated with both the calibrated low-cost sensor inputs and the model-generated pollution fields. Understanding and communicating this uncertainty is essential for ensuring transparency, reliability, and practical usability of the resulting air quality maps. In the next chapter, I extend the modelling framework by introducing comprehensive uncertainty quantification methods for both types of uncertainty, and by developing a Bayesian Maximum Entropy (BME) fusion system that integrates multi-source uncertainty information to produce probabilistic, uncertainty-aware air quality estimates.

Chapter 4: Uncertainty Analysis in Air Quality Estimation

Although numerous previous studies have leveraged deep learning for air quality modeling, the uncertainty in their predictions, arising from both noise in measurements (especially from low-cost sensors) and inherent uncertainty of the model, has often been overlooked. Quantifying such uncertainties is crucial for assessing the trustworthiness of estimation, which in turn improves interpretability and supports critical decision-making processes. Moreover, uncertainty estimation can reveal structural weaknesses in monitoring networks such as under-instrumented areas, where additional data collection could further reduce uncertainties. However, despite the growing importance of uncertainty assessment, most existing machine learning models for air quality estimation prioritize predictive accuracy alone, often neglecting to quantify the confidence associated with their outputs.

A comprehensive uncertainty quantification framework should address: 1. Input uncertainty, such as uncertainty of low-cost sensor measurements and inherent uncertainty of the air quality model. 2. Output uncertainty, by systematically integrating data with varying uncertainty levels to provide an overall quantification of uncertainty.

Given the increasing emphasis on uncertainty analysis, my study focuses on:

- Quantifying uncertainty from both low-cost sensor measurements and ViT-based air quality model predictions.
- Introducing a Bayesian Maximum Entropy (BME) approach to comprehensively quantify uncertainty by integrating data with varying levels of uncertainty from low-cost sensors, reference monitoring stations, and air quality model predictions.

By moving beyond point predictions and incorporating confidence assessments, this study takes a critical step toward developing trustworthy and uncertainty-aware urban air quality models.

4.1 Literature Review

Air quality models are increasingly used not only to predict pollutant concentrations but also to quantify uncertainty associated with those predictions (Carnevale et al., 2016). This shift toward uncertainty-aware modeling has become essential for supporting reliable air quality management and evidence-based policy decisions. For example, areas predicted to be high-risk but with large uncertainty may require additional monitoring before any intervention is implemented, whereas high-confidence estimates can guide targeted mitigation strategies.

Uncertainty arises from various sources – from measurement errors in low-cost sensors to incomplete knowledge in complex prediction models. Researchers broadly classify these uncertainties into two types: aleatoric and epistemic (Lee & Lee, 2024). Aleatoric uncertainty refers to the inherent noise in observed data (e.g. sensor measurement error), while epistemic uncertainty refers to uncertainty in the model itself (Murad et al., 2021a). In practical modeling scenarios, both forms contribute to the total predictive uncertainty.

This section first reviews each type of uncertainty within the context of air quality monitoring and modeling. It then introduces data fusion approaches, namely the Bayesian Maximum Entropy (BME) framework, as a principled probabilistic method for integrating multiple data sources with varying uncertainty characteristics. Finally, I identify the research gaps and contributions of the present work.

Aleatoric Uncertainty in Air Quality Measurements

Aleatoric uncertainty refers to the inherent randomness in measurements. In air quality applications, this primarily encompasses measurement errors and natural

fluctuations in pollutant concentrations. For example, ambient pollutant readings from monitoring devices can vary due to instrument noise. Low-cost sensors (LCS), which are increasingly deployed for high-resolution air quality monitoring, tend to exhibit particularly high aleatoric uncertainty. LCS measurements often deviate from reference-grade instruments in ways that cannot be fully corrected by calibration. Calibration procedures are designed to remove error components that can be explained by the chosen input variables and calibration model structure, thereby substantially reducing systematic discrepancies between low-cost sensors and reference instruments. However, measurement variability that cannot be accounted for by the available inputs or the adopted calibration regime—such as sensor-intrinsic noise or unobserved influencing factors—remains present. These unexplained components constitute a key source of residual uncertainty in calibrated LCS measurements. A field study with 100 low-cost NO₂ sensors in London found an average measurement error of approximately 35% (root-mean-square error) even after calibration (Peters et al., 2021). Such large residual errors underscore that sensor noise remains significant source of uncertainty in LCS-based air quality data.

For regulatory-grade instruments, measurement uncertainty might be relatively low and well-characterized by instrument specifications. In contrast, for LCS devices, one must empirically estimate their uncertainty – for instance, by co-locating sensors with reference monitors to quantify error distributions (Peters et al., 2021). Effectively accounting for aleatoric uncertainty is crucial; if ignored, the model may place undue confidence in noisy sensor readings, leading to spurious conclusions about air quality levels (Murad et al., 2021a). Aleatoric uncertainty, as an inherent property of each low-cost sensor measurement, could be carried forward into data fusion framework.

A common approach to quantify such uncertainty is calculating the variance (or standard deviation) of repeated measurements conducted simultaneously under identical environmental conditions. For instance, Hassani et al. (2023) conducted an uncertainty analysis by mounting eleven calibrated "Snifferbike" sensors on a single bike and simultaneously measuring pollutant levels across different environments.

Peters et al. (2021) employed sensor co-location with reference-grade monitors to empirically characterize error distributions. While empirical estimation methods (such as co-location studies or repeated measurements) are valuable, they face practical limitations:

- Deploying all low-cost sensors simultaneously at a single location is often labor-intensive and costly, especially for large sensor networks.
- Uncertainty characterized at one specific location typically may not generalize well to other locations due to varying environmental conditions. In other words, uncertainty (measurement variance) may fluctuate across spatially and temporally diverse conditions.

Although calibrated low-cost sensor networks generate abundant observational data, current research lacks comprehensive methods for systematically quantifying the uncertainty associated with individual calibrated sensors across extensive urban deployments, particularly at specific locations and time steps.

Research Gap and Contribution: For individual low-cost sensors, typically only a single measurement per sensor per time is available, precluding direct calculation of variance through repeated simultaneous measurements. To overcome this limitation, I propose a novel approach to estimate sensor-specific uncertainty by leveraging co-located calibrated LCS measurements and reference monitor data. Specifically, I design a statistical model that predicts the variance of calibrated LCS readings estimated via co-location with reference monitors under diverse conditions, effectively providing an uncertainty estimate for any calibrated sensor measurement (detailed in Section 4.2.1).

Epistemic Uncertainty in Air Quality Models

Epistemic uncertainty reflects a model's lack of knowledge – for example, uncertainty in model parameters or structure due to limited training data. Epistemic uncertainty can be reduced in principle if more data or knowledge becomes available

(Lee & Lee, 2024). In the context of air quality modeling, epistemic uncertainty is associated with the predictive model used to estimate pollutant concentrations (such as a statistical regression, chemical transport model, or machine learning model). A complex model like a deep neural network may fit the training data well yet still be uncertain when making predictions at new locations or times that differ from the training distribution. This uncertainty stems from the model's incomplete learning of the true mapping from inputs (e.g. meteorological factors, emissions, sensor readings) to outputs (pollutant concentrations). When the model encounters conditions unlike those seen in training, its predictions become extrapolations and thus inherently uncertain (Vignesh et al., 2023).

Modern deep learning models for air quality typically do not provide uncertainty estimates by default; they yield a single point estimate of concentration. However, several techniques have been developed to approximate epistemic uncertainty in such models. These methods provide probabilistic predictions (e.g., predictive distributions) instead of point estimates, thereby addressing the limitations of deterministic approaches (Caldeira & Nord, 2020).

Bayesian Deep Learning is an emerging approach that enables deep neural networks to generate predictions along with uncertainty estimates, and it has gained increasing attention in recent years (Abdar et al., 2021). Recent advances in Bayesian Deep Learning have been successfully applied in fields ranging from image segmentation to autonomous driving (Mehta et al., 2021; Kendall et al., 2018; Xiao et al., 2019; Meyer et al., 2020). However, advanced neural networks have many parameters, and using Bayesian inference to derive explicit posterior distributions is computationally difficult. Bayesian neural networks (BNNs), for example, offer reliable uncertainty estimates but are difficult to implement and scale to very deep or large networks (Murad et al., 2021b), necessitating approximation techniques.

Monte Carlo dropout (MCDO) is a simple yet effective technique for estimating uncertainty without full Bayesian inference, which offers a more efficient means of approximating Bayesian inferences by enhancing both computational speed and

cost-effectiveness (Hsu et al., 2022a,b; Kumar et al., 2022a,b). MCDO provides a practical approximation to Bayesian inference for deep neural networks by leveraging dropout at inference time (Afshari et al., 2023). The Dropout mechanism in MCDO, originally introduced as a regularization technique during training to prevent overfitting, randomly deactivates a subset of neurons with a given probability p (Srivastava et al., 2014). Gal and Ghahramani (2016b) demonstrated that applying dropout during inference is mathematically equivalent to performing variational Bayesian inference in a deep Gaussian process. In other words, a neural network with dropout can be interpreted as a Bayesian model where the dropout mask instantiates a random subset of the model's weights, effectively sampling from a learned approximate posterior distribution over weights. By integrating over many such random dropout realizations, MCDO approximates the model's predictive distribution and thus quantifies epistemic uncertainty (uncertainty in the model). This approach captures the model's confidence in its predictions without requiring an explicit Bayesian neural network formulation. Specifically, the MCDO method employs deep learning model (e.g. neural network) trained using dropout. During inference, instead of turning off dropout (as would typically be done in standard neural networks), MCDO keeps dropout active, performing multiple inference (e.g., 50 or 100 times). This process produces a sample of predictions from an implicit ensemble, where each pass corresponds to a perturbed model. The variability in predictions provides an estimate of epistemic uncertainty (Gal Yarin, 2016).

In air quality applications, MCDO has been successfully used to capture uncertainty in deep learning models (Mokhtari et al., 2021). Murad et al. (2021) applied MCDO to estimate predictive uncertainty in forecasts of multiple pollutants (e.g., $\text{PM}_{2.5}$, NO_2) over urban monitoring networks. By flagging time periods with wide predictive intervals (low-confidence forecasts), the system enabled proactive review and, where necessary, supplementary measurements or manual inspection before any automated actions, thereby reducing the likelihood of false alarms or missed pollution events.

In summary, epistemic uncertainty captures the model's lack of knowledge, particularly under unfamiliar input conditions. Addressing this uncertainty—through techniques like MCDO—is a key step toward improving model transparency and reliability in air quality modeling.

Data Fusion Framework

The two forms of uncertainty discussed above—aleatoric uncertainty inherent in measurement data and epistemic uncertainty arising from model predictions—manifest at different stages of an air quality modeling pipeline. A comprehensive, uncertainty-aware modeling framework ideally accounts for both uncertainty types simultaneously; however, existing studies typically address these uncertainties separately (Hanigan et al., 2017; Hassani et al., 2023). To fully integrate both data and model uncertainties, a rigorous data fusion approach is beneficial because it explicitly combines observational data (along with their inherent aleatoric uncertainties) with predictive model outputs, which include epistemic uncertainties. Such integration allows uncertainties from different sources to be simultaneously propagated and quantified within a unified analytical framework, enhancing overall confidence in air quality estimation.

Data fusion methods systematically merge multiple datasets (e.g., observational measurements and model outputs) to produce a combined estimate that surpasses the quality of individual sources (Carnevale et al., 2020). In particular, data fusion techniques, which are closely related to data assimilation, are built on mathematically robust frameworks such as the Best Linear Unbiased Estimator (BLUE), which provides optimal, uncertainty-aware integration of different information streams (Lahoz & Schneider, 2014). When designed appropriately, data fusion frameworks can propagate both aleatoric and epistemic uncertainties into the fused result.

Past studies have shown the benefit of such fusion: for instance, combining chemical transport model outputs with monitoring data via Bayesian Maximum Entropy led

to improved ozone estimates in North Carolina, with higher accuracy and precision than using interpolation alone (de Nazelle et al., 2010). Bayesian Maximum Entropy (BME) is a rigorous data fusion method uniquely designed to integrate diverse data sources (each characterized by distinct types and degrees of uncertainty) within a unified probabilistic framework (Akita et al., 2012). Unlike conventional data assimilation approaches such as Optimal Interpolation or Kalman Filtering, which typically assume Gaussian uncertainty and require explicit covariance structures, BME flexibly incorporates uncertainties represented by arbitrary probability distributions and allows for incomplete or soft information. This flexibility makes BME particularly suitable for environmental modeling scenarios involving heterogeneous data sets, such as air quality observations from diverse sensors, expert judgments, and model outputs, each accompanied by differing uncertainty characterizations (Christakos & Serre, 2000). Originally developed by Christakos et al. (1990), BME provides a mathematically robust method to incorporate diverse sources of information (or "knowledge bases") into spatial estimations. Unlike traditional interpolation methods, which typically use only exact measurements, BME distinguishes between *hard data* (precise measurements with negligible uncertainty) and *soft data* (information with associated uncertainty described by a probability distribution) (Akita et al., 2014). This categorization enables BME to integrate not only observed pollutant concentrations from reference stations but also auxiliary information such as model outputs or low-cost sensor readings that come with uncertainty.

In the BME framework, knowledge integration occurs in two stages. First, a prior stage employs maximum entropy principles on general knowledge (e.g., known trends, physical laws, or broad model predictions) to construct a prior probability density function (PDF) of the spatial field of interest. Next, a posterior stage applies Bayesian updating (Bayesian conditionalization) to refine this prior PDF by incorporating site-specific data, both hard and soft. The result is a BME posterior distribution for the quantity of interest (e.g., pollutant concentration at an unsampled location) that reflects all available information and its uncertainties. The expected

value of this posterior distribution is used as the best estimate, while the posterior standard deviation provides a quantitative measure of estimation uncertainty. Notably, if only hard data are used, BME reduces to the classic kriging estimator, meaning BME is a consistent generalization of kriging for cases involving uncertain data. This property underscores BME's importance as a unifying approach: it retains the well-known advantages of kriging when high-quality data are available, and extends those capabilities to handle less certain information sources (Serre & Christakos, 1999). BME's utility has also been demonstrated in real-world air pollution applications; for example, Yu et al. (2009) used BME to estimate residential PM₁₀ and ozone exposures over multiple time scales, achieving spatiotemporal predictions with quantified uncertainties. Akita et al. (2014) applied BME to integrate results from an existing land-use regression (LUR) model and a chemical transport model (CTM) with observational data, estimating annual average NO₂ concentrations over Catalunya, Spain. Similarly, Hanigan et al. (2017) employed BME to combine LUR predictions with NO₂ measurements, incorporating uncertainty through informative priors.

However, previous BME studies have primarily focused on blending measurements from reference monitoring stations, which are typically treated as hard data without uncertainty, while neglecting uncertainty quantification when incorporating low-cost sensor data.

Research Gap and Contribution

Research Gap: Based on the above review, low-cost sensor networks produce a wealth of data but with significant measurement uncertainties; yet few studies quantify the aleatoric uncertainty from low-cost sensors (Peters et al., 2021). Meanwhile, the epistemic uncertainty inherent in advanced air quality models (e.g. deep learning model) has often been overlooked. Moreover, no existing research has systematically integrated **aleatoric uncertainty**—arising from air quality measurements, particularly those collected by low-cost sensor (LCS) networks—with **epistemic uncertainty** inherent in advanced air quality models such as deep

learning, within a unified data fusion framework. The potential of such integrated frameworks to produce reliable, uncertainty-informed air quality maps remains largely unexplored. This gap is especially critical given the growing dependence on LCS networks, whose data is increasingly widespread but often highly uncertain, and the expanding use of deep learning models in air quality research.

Contributions: To address these gaps, this study proposes a novel uncertainty-aware air quality modeling framework. The key contributions of this work are:

- **Uncertainty Estimation (Aleatoric and Epistemic):** I explicitly quantify aleatoric uncertainty associated with low-cost sensor (LCS) measurements and epistemic uncertainty arising from the deep learning model as input data into data fusion framework.
- **Joint Uncertainty Integration:** I advance the application of Bayesian Maximum Entropy for air quality data fusion by incorporating the above uncertainties as input information. The BME method fuses these disparate sources – uncertain LCS data, high-confidence reference monitor data, and uncertain deep learning model outputs – to produce a spatial map of pollutant concentration along with a quantified uncertainty.
- **Improved Uncertainty-Aware Predictions:** Through case studies in London (detailed in Section 4.2), I show that my integrated approach yields informative results. Specifically, the framework provides predictions with quantified uncertainty, which offer insight into where the prediction is confident and where it is less certain. This allows for more nuanced risk assessment; for example, identifying locations where predicted pollution is high and uncertainty is low (high priority for action) versus locations where predicted pollution is high but uncertainty is also high (priority for further measurement). This illustrates the practical utility of uncertainty-aware modeling for environmental risk management.

The following sections will detail the development of this framework and evaluate its performance, illustrating the benefits of an uncertainty-aware perspective in air quality science.

4.2 Methodology and Research Design

This study proposes a Bayesian Maximum Entropy (BME)-based data fusion framework designed to produce high-resolution air quality concentration maps along with explicit quantification of associated uncertainties. This approach addresses a critical research gap concerning the integrated management of aleatoric (measurement-related) and epistemic (model-derived) uncertainties, particularly within the context of low-cost sensor (LCS) networks.

The methodology encompasses three core steps:

- First, I quantify the aleatoric uncertainty associated with calibrated LCS measurements, as detailed in Section 4.2.1.
- Second, I estimate the epistemic uncertainty inherent in proposed air quality model predictions in Chapter 3 through the MCDO method, elaborated in Section 4.2.2.
- Third, Section 4.2.3 describes the integration of multiple data sources within a carefully designed BME-based fusion workflow, specifically combining:
 - Deep learning model outputs, incorporating epistemic uncertainty;
 - LCS observations, including aleatoric uncertainty;
 - High-precision reference measurements from the London Air Quality Network (LAQN)

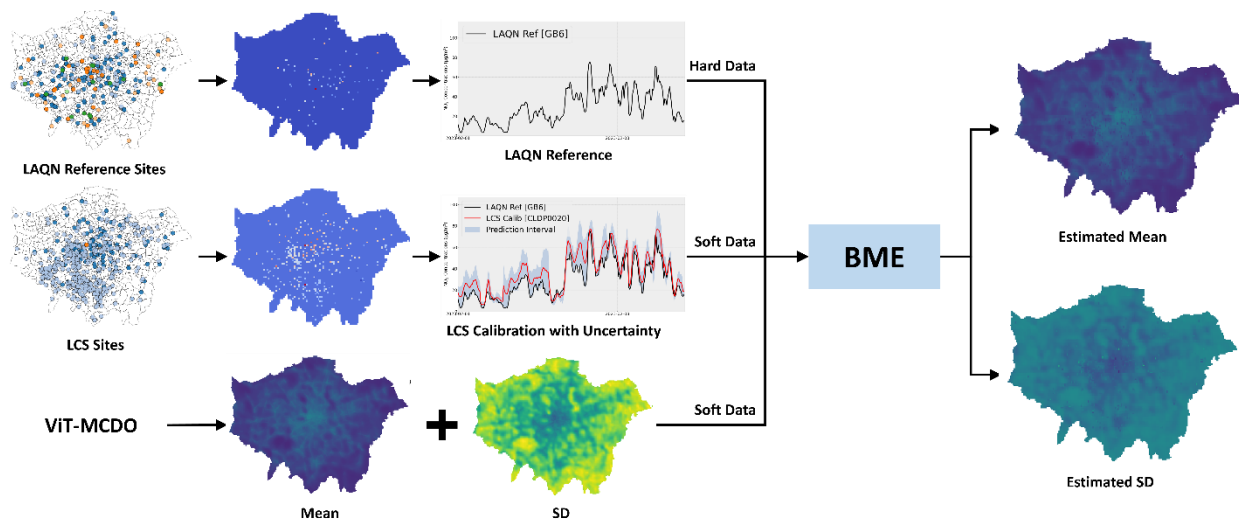


Figure 4.1 The proposed BME framework for uncertainty estimation

Figure 4.1 shows the proposed workflow optimally fuses heterogeneous data characterized by distinct uncertainty levels. Consequently, the framework generates hourly concentration estimates and corresponding uncertainty maps for $\text{PM}_{2.5}$ and NO_2 at a fine spatial resolution ($100 \text{ m} \times 100 \text{ m}$) across Greater London.

4.2.1 Quantifying Low-Cost Sensor Uncertainty as Inputs to the Data Fusion Framework

After calibration, the distribution of measurements from the low-cost sensors closely aligns with that of reference-grade monitors (as demonstrated in Figures 2.18–2.19). This empirical consistency supports the assumption that the calibrated low-cost sensor measurements are essentially unbiased estimates of the true pollutant concentration. In other words, calibration removes systematic errors (bias), allowing us to treat the sensor's output as an accurate reflection of ambient pollutant levels on average. It is important to note that this consistency does not imply perfect agreement for individual sensors or at all time periods. Residual discrepancies may persist due to sensor-specific characteristics, temporal variations, and unmodelled environmental influences. Nevertheless, from a statistical perspective, the

calibration results support the working assumption that calibrated low-cost sensor measurements can be treated as approximately unbiased estimators of ambient pollutant concentrations when averaged across sensors and over time. Given this approximate unbiasedness assumption, the remaining challenge is to quantify the random uncertainty in these calibrated readings. I address this by leveraging co-located reference monitor data to characterize the variability of sensor errors.

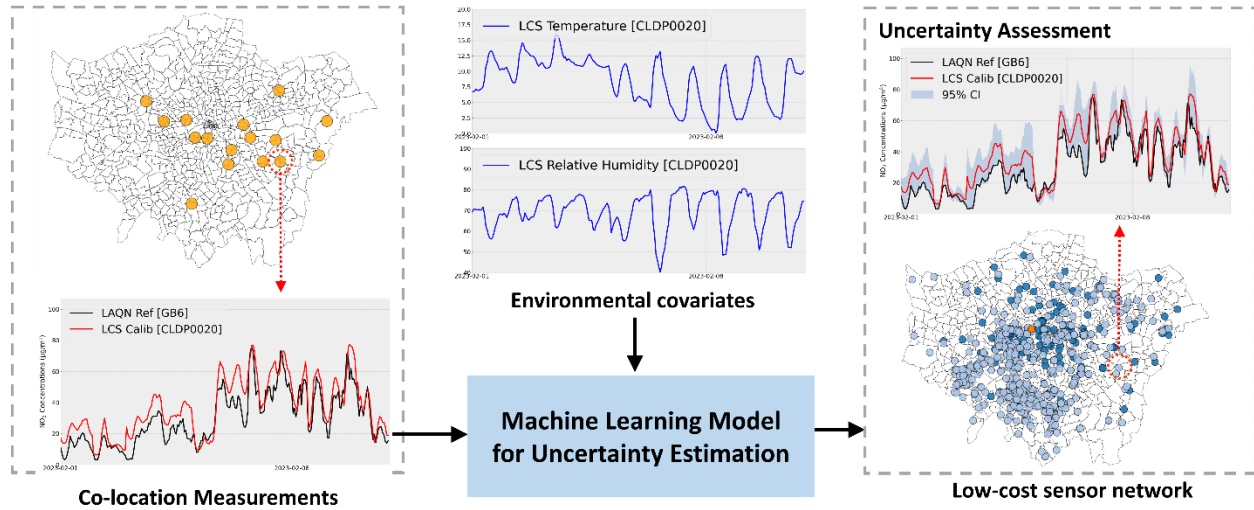


Figure 4.2 The proposed framework for LCS uncertainty estimation

Although the calibrated sensor estimates are unbiased on average, the variance of their errors is not necessarily constant. In practice, the residuals between each pair of colocation BL sensor reading and the reference measurement (i.e. the differences $r_t = y_t^{\text{lcs}} - y_t^{\text{ref}}, r_t \sim \mathcal{N}(0, \sigma^2)$) can exhibit heteroskedasticity; their dispersion depends on factors like environmental conditions and the magnitude of the measurement itself. For example, changes in ambient temperature or humidity might degrade or improve sensor precision, and the sensor may be inherently less reliable at higher pollutant concentrations. It is important to model this non-constant error variance rather than assume homoskedastic (constant-variance) errors; doing so enables more accurate uncertainty quantification and prevents mischaracterization of the sensor's confidence under varying conditions.

To capture the heteroskedastic error behavior, I employ a random forest regression model that predicts the squared residual r_t^2 as a function of relevant predictors. The model is trained on all available co-located sensor-reference data, using features that are expected to influence measurement error:

T_t : ambient temperature at time t .

H_t : ambient relative humidity at time t .

y_t^{lcs} : the calibrated low-cost sensor reading (PM_{2.5} or NO₂) at time t .

Formally, I model the error variance as:

$$r_t^2 = g(T_t, H_t, y_t^{\text{lcs}}) + \eta_t \quad (4.1)$$

where $g(\cdot)$ represents the function learned by the random forest and η_t is a random noise term (the unexplained portion of r_t^2). In essence, $g(T_t, H_t, y_t^{\text{lcs}})$ estimates the expected squared error (or variance of the measurement error) under the conditions at time t . I chose a random forest for this task because it can flexibly model nonlinear interactions between the predictors and does not require specifying a fixed functional form for the error variance. By training on a comprehensive set of paired BL and reference observations, the model learns how the calibrated BL sensor's uncertainty (error variance) varies with temperature, humidity, and concentration level.

In this study, I employed the random forest (RF) method exclusively, as it is a well-established and widely used machine learning algorithm that provides a favorable balance between predictive performance and computational efficiency. Compared to linear models, RF can more effectively capture complex nonlinear relationships in the data; and in contrast to deep learning approaches, it is less demanding in terms of training data volume, hyperparameter tuning, and computational resources. My primary objective was to illustrate a feasible approach for quantifying LCS

uncertainty rather than exhaustively evaluating multiple models to identify an optimal technique.

4.2.2 Quantifying ViT-based Air Quality Model Uncertainty as Inputs to the Data Fusion Framework

Although advanced AI models (e.g. Deep learning models) emphasize accuracy, they frequently overlook the reliability of their predictions. In this study, I delve deeper into the integration of the ViT-based air quality model and Monte Carlo Dropout method to characterize uncertainty in the ViT-based model estimates.

Mathematically, let $f(x; W)$ be a deep model with learned weights W . Introducing dropout means that at each forward pass I sample a binary mask m for the neurons (or equivalently for a subset of weights) from a Bernoulli distribution (with dropout rate p). I denote the masked weight set as $W \odot m$ (where $\odot$ is elementwise product, setting a subset of weights to zero). A single stochastic forward pass with dropout yields a prediction $\widehat{y}^{(t)} = f(x; W \odot m^{(t)})$, where t indexes one Monte Carlo sample. Repeating this T times produces a set of predictions $\widehat{y}^{(1)}, \widehat{y}^{(2)}, \dots, \widehat{y}^{(T)}$ drawn from the model's approximate posterior predictive distribution. The **predictive mean** is then estimated as the average of these samples:

$$E[y|x, D] \approx \frac{1}{T} \sum_{t=1}^T \widehat{y}^{(t)} \quad (4.2)$$

and the **predictive variance** (epistemic uncertainty) can be estimated as:

$$Var[y|x, D] \approx \frac{1}{T} \sum_{t=1}^T \left(\widehat{y}^{(t)} \right)^2 - \left(\frac{1}{T} \sum_{t=1}^T \widehat{y}^{(t)} \right)^2 \quad (4.3)$$

This variance corresponds to the sample variance of the T stochastic forward passes. Notably, Gal and Ghahramani (2016b)’s derivation shows that as $T \rightarrow \infty$, this Monte Carlo estimation converges to the predictive distribution of a Bayesian Neural Network under a particular prior and variational approximation. MCDO thus provides a probabilistic Bayesian interpretation of dropout: each random dropout mask approximately samples from the posterior over network weights, and averaging predictions corresponds to marginalizing over this posterior. Importantly, this approach does not require changing the model beyond using dropout—the same trained model can produce uncertainty estimates via multiple stochastic forward passes. This combination of simplicity and theoretical grounding makes MCDO an attractive method for uncertainty quantification in deep learning models.

Applying MCDO in the proposed ViT-Based Air Quality Model

To implement Monte Carlo Dropout for uncertainty assessment in ViT-based air quality model, I follow these steps:

1. Model Architecture with Dropout: In the ViT model, I apply dropout with probability p in each vision transformer block (on the attention weights and hidden units) and in the final regression layer (last layer of CNN). The dropout rate p is common values range from 0.1 to 0.5, I experimented with dropout rates of 0.1 and 0.25.

2. Enabling Dropout at Inference: For prediction (inference) time, I enable the dropout layers during multiple stochastic forward passes. Perform T forward passes of the model for each input x . In each pass $t = 1, \dots, T$, the model will randomly drop different units, yielding a prediction $\hat{y}^{(t)}$. Typical values of T in literature range from 20 to 200; a larger T yields a more stable uncertainty estimate at the cost of more computation. In my case, I choose $T = 100$ forward passes as a reasonable trade-off. It is worth noting that these T forward passes are independent samples from the model’s approximate posterior; if computational resources allow, they can be run in parallel batches to speed up the process.

3. Aggregating Predictions for Mean and Uncertainty: After obtaining $\widehat{y}^{(1)}, \widehat{y}^{(2)}, \dots, \widehat{y}^{(T)}$ from the T stochastic passes, aggregate them to quantify uncertainty. Compute the **Monte Carlo mean prediction** $\bar{y} = \frac{1}{T} \sum_{t=1}^T \widehat{y}^{(t)}$, which serves as the final point estimate for the model's output. Next, compute the **model uncertainty**. A straightforward measure is the sample variance: $s^2 = \frac{1}{T} \sum_{t=1}^T \left(\widehat{y}^{(t)} - \bar{y} \right)^2$, which reflects the model's *epistemic uncertainty* for that prediction. A high variance s^2 indicates the model's predictions were highly sensitive to weight perturbations (dropout), implying low confidence in the prediction for that input; conversely, a low variance suggests that all sampled subnetworks produced similar outputs, indicating high confidence. The outcome of this step is a probabilistic prediction: a mean estimate with an associated uncertainty. Since the ViT model outputs a spatial map of pollutant concentrations, I would compute the mean and variance at each grid across the T runs, yielding an uncertainty map.

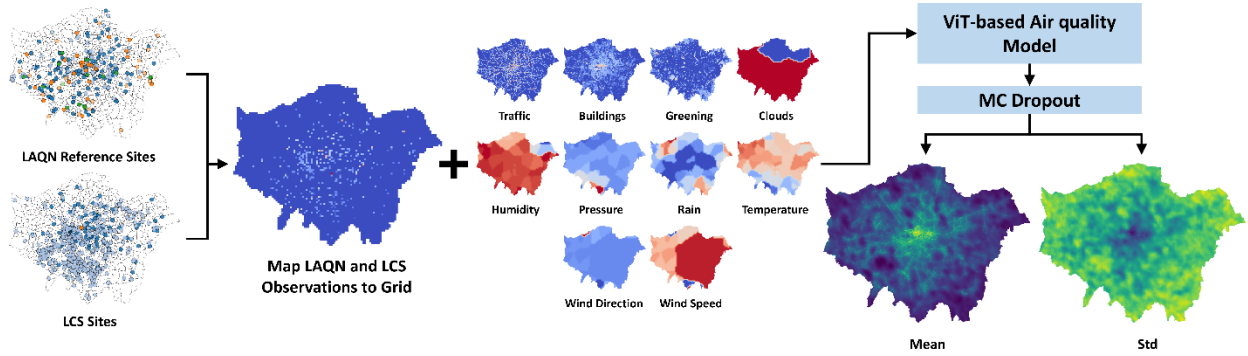


Figure 4.3 The proposed framework for uncertainty estimation of proposed ViT-Based Air Quality Model by Monte Carlo Dropout

Whereas Chapter 3 introduced the ViT-based air quality modeling approach, the focus in this chapter is on quantifying uncertainty in ViT-based air quality model predictions using MCDO, thereby enhancing their reliability. Since the datasets and model employed here are identical to those described in Chapter 3, detailed descriptions are omitted to avoid redundancy.

For each hourly prediction, I generated 100 air pollution maps using Monte Carlo Dropout ($T = 100$) to quantify uncertainty in the predictions of NO_2 and $\text{PM}_{2.5}$ concentrations. These 100 maps were subsequently concatenated, forming the final output of the ViT-MCDO model. The final predicted pollutant concentration at each grid cell was computed as the average of these samples, and the uncertainty was quantified using their standard deviation.

Uncertainty Evaluation Metrics: To rigorously assess the quality of the probabilistic predictions, I employed several metrics commonly used in uncertainty quantification. These include:

Prediction Interval Coverage Probability (PICP): PICP measures the fraction of true values that fall within a predicted interval. For instance, if I form the prediction interval from the mean and standard deviation, PICP is the percentage of times the actual observed pollutant concentration lies inside this interval.

Mean Prediction Interval Width (MPIW): This metric represents the average width of the prediction intervals. While PICP gauges reliability (coverage), MPIW gauges sharpness or specificity of the prediction. A narrower interval (low MPIW) is more informative but if it's too narrow, it may not cover the true values (low PICP). I seek a balance where intervals are as tight as possible while still achieving the desired coverage. MPIW is usually reported in the same units as the target (e.g., $\mu\text{g}/\text{m}^3$ for pollutant concentration) and I calculated it for the 95% intervals in results.

Continuous Ranked Probability Score (CRPS): Continuous Ranked Probability Score (CRPS) is a proper scoring rule that quantifies the distance between the cumulative distribution function (CDF) of the predicted distribution and the empirical CDF corresponding to the observation (represented as a step function). As a generalization of the mean absolute error (MAE) to the probabilistic setting, CRPS evaluates both the accuracy and the sharpness of a probabilistic prediction. Unlike point-based metrics, CRPS captures the full shape of the predictive distribution. A lower CRPS indicates a probabilistic prediction that is both accurate and confident.

In contrast, probabilistic predictions that are either overconfident but inaccurate, or accurate but overly uncertain, tend to yield higher CRPS values. I employed CRPS to evaluate the overall quality of the probabilistic predictions produced by the ViT-MCDO model, as it provides a more comprehensive assessment of whether the predicted distribution is both reasonable and reliable. CRPS is defined as:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - I(x \geq y))^2 dx \quad (4.4)$$

where $F(x)$ is the CDF of the predictive distribution. $I(x \geq y)$ is the indicator function, which equals 1 if $x \geq y$, and 0 otherwise.

If the predictive distribution is Gaussian (i.e., normally distributed) with mean μ and standard deviation σ , and the true observation is y , then the CRPS has the following closed-form solution:

$$\text{CRPS}(\mu, \sigma, y) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\phi(z) - z(2\Phi(z) - 1) \right] \quad (4.5)$$

Where:

- μ : the predicted mean (expected value of the distribution).
- σ : the predicted standard deviation (uncertainty of the prediction).
- $z = \frac{y-\mu}{\sigma}$: the standardized error (i.e., how many standard deviations the true value y is from the mean).
- $\phi(z)$: the standard normal probability density function (PDF), defined as:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (4.6)$$

- $\Phi(z)$: the standard normal cumulative distribution function (CDF), defined as:

$$\Phi(z) = \int_{-\infty}^z \phi(t) dt \quad (4.7)$$

4.2.3 Bayesian Maximum Entropy Approach for Integrating Data with Varied Uncertainties

Bayesian Maximum Entropy (BME) Framework

The Bayesian Maximum Entropy (BME) is a robust two-stage probabilistic data fusion method capable of systematically integrating datasets with varying degrees of uncertainty (Christakos et al., 2002). At the prior stage, maximum entropy theory is used to process the general knowledge base describing global characteristics of the spatial random field, such as its mean trend and the covariance function, which produces a prior PDF depicting the spatial process. Then at the posterior stage, a Bayesian conditionalization rule is used to update this prior PDF with respect to the site specific hard and soft data available, which produces a BME posterior PDF describing the value of the spatial process at any estimation point of interest.

1. Prior Stage: Maximum Entropy Prior Distribution

At the prior stage, BME employs the principle of maximum entropy to derive a prior probability density function (PDF) that encodes general knowledge G about the underlying spatial random field $Z(s)$, such as pollutant concentrations (Christakos, 1990). This general knowledge typically includes statistical constraints, such as

known mean trends, covariance structures, physical laws, or any other informative contextual assumptions about the field.

Formally, let $Z(s)$ denote the random field of interest defined over spatial locations s . The general knowledge G is represented by constraints on expectations of specified feature functions $g_\alpha(Z)$:

$$E[g_\alpha(Z)] = G_\alpha \quad (4.8)$$

where these feature functions $g_\alpha(Z)$ can include, for example, the mean $m_Z(s) = E[Z(s)]$ or the covariance function $C_Z(s, s') = \text{Cov}[Z(s), Z(s')]$.

Under these constraints, the maximum entropy principle selects the least biased prior joint PDF $f_G(z)$ that avoids introducing unnecessary assumptions beyond the given knowledge G . This selection is formulated as a variational optimization problem:

$$\begin{aligned} \max_f H[f] = - \int f(z) \ln f(z) dz, \text{ subject to } \int f(z) g_\alpha(z) dz \\ = G_\alpha \text{ for all } \alpha, \end{aligned} \quad (4.9)$$

where $H[f]$ is the entropy of the joint PDF $f(z)$ of the random field (Christakos, 1990). Solving this constrained optimization using Lagrange multipliers yields the maximum entropy prior distribution f_G .

Practically, if the available general knowledge includes only the mean and covariance, the resulting maximum-entropy prior is a multivariate Gaussian distribution:

$$f_G(Z) \sim N(m_Z, C_Z) \quad (4.10)$$

which precisely reproduces these known moments (Akita et al., 2012). Thus, this Gaussian prior provides the least biased representation of the spatial random field consistent with the given knowledge.

2. Posterior Stage: Bayesian Update with Hard and Soft Data

In the posterior stage, BME performs a Bayesian conditionalization to update the prior PDF with site-specific data S , which consists of both hard and soft data (Christakos, 1990; Akita et al., 2012). Hard data provide exact values z_h at certain locations (e.g. reference monitor readings), while soft data provide probabilistic information f_S about the variable (e.g. a predictive distribution from a model or an observation with measurement uncertainty). Let $K = G \cup S$ denote the complete knowledge base. BME computes the posterior PDF f_K (also denoted $f_{G \cup S}$) for an unsampled location s_k by conditioning the prior on all site-specific information. For example, if Z_h and Z_s are random variables representing the field at hard-data and soft-data locations respectively, the BME posterior for $Z_k = Z(s_k)$ can be expressed as:

$$f_K(z_k | z_h, f_S) \propto \int f_G(z_k, z_h, z_s) f_S(z_s) dz_s \quad (4.11)$$

where $f_G(z_k, z_h, z_s)$ is the prior joint PDF (from the prior stage) over the estimation point, hard data, and soft data, and $f_S(z_s)$ is the probability density encoding the soft data information. In practice, hard data enter this expression as conditions (point masses at the observed z_h values), and soft data enter as weighting functions that adjust the prior distribution. The result $f_K(z_k | K)$ is the BME posterior distribution of the quantity $Z(s_k)$ given all knowledge K . This posterior fully quantifies the uncertainty of the estimate at s_k after assimilating the data. For instance, the BME posterior mean $E[Z(s_k) | K]$ is taken as the best estimate, and the posterior variance $\text{Var}[Z(s_k) | K]$ provides a measure of predictive uncertainty at that location (Akita et al., 2012). The formal inclusion of soft data enables BME to utilize

information that kriging (which uses only hard data) would ignore, thus often achieving more accurate and robust spatial predictions (Akita et al., 2012).

Integrating Multi-Source Uncertain Data Using the Bayesian Maximum Entropy Framework

In this study, the BME framework is employed to fuse three heterogeneous sources of air quality data: (1) uncertainty-aware air quality model outputs, (2) low-cost sensor measurements with uncertainty, and (3) regulatory-grade reference station observations. I treat the high-quality reference station readings as hard data, whereas the other two sources are incorporated as soft data with associated uncertainty distributions. The air quality model provides a predicted concentration field along with uncertainty, which was introduced in Section 4.2.2. The calibrated low-cost sensors, while providing point measurements, still carry non-negligible calibration uncertainty as introduced in 4.2.1 and are thus modeled as soft data. By integrating these sources through the BME update, the posterior distribution at any location reflects both the reliable anchor of reference monitors and the extensive coverage of model predictions and low-cost sensor measurements. Consequently, the BME approach in this study yields a spatial concentration map with rigorous uncertainty quantification, leveraging all available information in a principled manner. The inclusion of soft data in addition to hard data leads to more informative estimates than using hard data alone, demonstrating BME's advantage in integrating multi-source air quality data (Akita et al., 2012; Christakos, 1990). The framework's ability to reduce to kriging in the absence of soft data ensures consistency with classical methods, while its capacity to incorporate uncertain information provides a powerful generalization for environmental prediction.

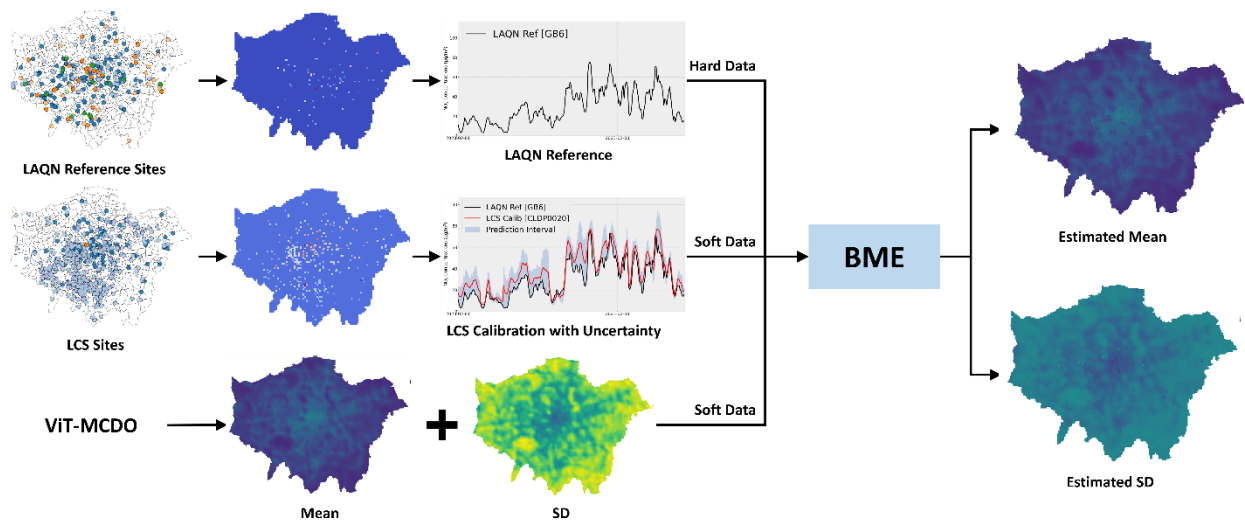


Figure 4.4 The BME framework for integrating data source with varying uncertainty

Within the BME framework, integrating these sources means that at locations of reference monitors, the estimation is tightly constrained by the observed values, whereas at other locations, the air quality model predictions and LCS data inform the estimates in proportion to their reliability. BME inherently gives precedence to high-confidence data while still utilizing lower-confidence information by formally accounting for their uncertainty (Nazelle et al., 2010).

For example, where the air quality model's performance is strong (low prediction error), its influence on the BME posterior will be greater, effectively injecting detailed spatial patterns learned by the model. Conversely, in areas or times where the model is less certain, the BME approach will rely more on actual sensor measurements, thereby preventing the model's uncertainties from degrading the final estimate. Similarly, LCS data, when available, provide local corrections to model estimates – but their influence is tempered by their calibration uncertainty. The BME data fusion thus acts as a dynamic weighting scheme driven by uncertainty: information from the deep model and LCS is incorporated as a function of its trustworthiness, and the combined result is a spatial pollutant field that is both informed by rich data sources and cognizant of their limitations.

Detrending Data Inputs Using a Global offset

The global offset refers to a smoothed spatial trend surface representing the large-scale variation in pollutant concentrations across the study area. In geostatistical analyses, the global offset is generally removed from the data prior to the analysis in order to produce residuals that are as homogeneous as possible. In this study, I estimated the global offset by applying a moving average smoother to the ViT-based model predictions, using a 300×300 m grid resolution. I chose this resolution so that the smooth trend should produce low variance in the residuals while maintaining adequate residual autocorrelation. As Xu et al. (2016b) point out when a trend surface is smoother and less variable across space, the transformed residuals retain more variability. Conversely as the surface becomes less smooth the residuals are less autocorrelated.

Prior Probability Density Functions of BME

Parametrization for Prior Probability Density Functions of Soft Data: I construct the PDFs by combining soft data including the LCS measurements and ViT-based air quality model. At the soft data points (100m grid centroids and BL sensor locations), the PDF was estimated from the uncertainties associated with the expected values at each point, calculated in Section 4.2.1 and 4.2.2, respectively.

Determination of the Spatial Covariance Model

The sample covariogram was calculated from the residuals of all hard and soft data after removing global offset, and was then used to fit a positive definite covariogram model using an automated weighted least-squares procedure. In this study, exponential, Gaussian, and spherical models were tested, and the exponential model which produced the smallest mean square error was used in the analysis.

Validation of the BME Posterior PDFs

I computed the BME posterior PDFs at the nodes of a two-dimensional network of estimation points on the regular 100m map grid. The expected value of the BME posterior PDF was used as an estimate of the residual concentration at the estimation point, and the corresponding BME posterior standard deviation provides a useful characterization of the associated estimation uncertainty. The deterministic global offset interpolated on the estimation location was finally added back onto the residual estimate to obtain the NO_2 and $\text{PM}_{2.5}$ estimate.

I randomly selected 20% of the London Air Quality Network $\text{PM}_{2.5}$ and NO_2 monitoring stations to serve as an independent test dataset. Bayesian Maximum Entropy (BME) requires the inversion of an $N \times N$ covariance matrix, where N represents the total number of hard and soft data points. The computational complexity for matrix inversion is $O(N^3)$ in time and $O(N^2)$ in memory, typically accomplished through standard Cholesky decomposition. Given the extensive number of LCS observations and fine-scale air quality model outputs, the resulting covariance matrix becomes very large, leading to significant computational demand. Due to limited computational resources (an NVIDIA RTX 3090 GPU), each hourly BME data fusion calculation took over 10 hours. Consequently, I opted not to perform a cross-validation because of the prohibitively high computational burden. Additionally, I selected data from a single representative day (14th February 2024) for model fitting and validation. The BME framework was implemented using MATLAB R2024a (MathWorks Inc., Natick, MA, U.S.A.) in combination with the BMElib2.0b toolbox developed by Serre et al. at the University of North Carolina (BMElib, 2024).

To assess the effectiveness of the proposed BME framework, I compared hourly concentrations recorded at test reference monitoring stations with the probabilistic predictions from the BME and ViT-MCDO models. Due to the computational intensity associated with the BME method, the analysis was restricted to a representative day, February 14, 2024.

In the ViT-MCDO method, probabilistic predictions were derived using the mean and standard deviation from ViT-MCDO outputs (with parameters: $T=100$, dropout rate $=0.1$). The BME method produced probabilistic predictions through data fusion of ViT-MCDO predictions, calibrated Breathe London (BL) sensor data, and reference station observations within the Bayesian Maximum Entropy framework.

Evaluation Metrics: To quantify the accuracy of each method, 2 error statistics were calculated: root-mean-square error (RMSE), Pearson’s correlation coefficient(r). To rigorously assess the quality of the probabilistic predictions, I adopted four metrics previously described in Section 4.2.2: PICP, MPIW and CRPS.

4.3 Experiments and Results

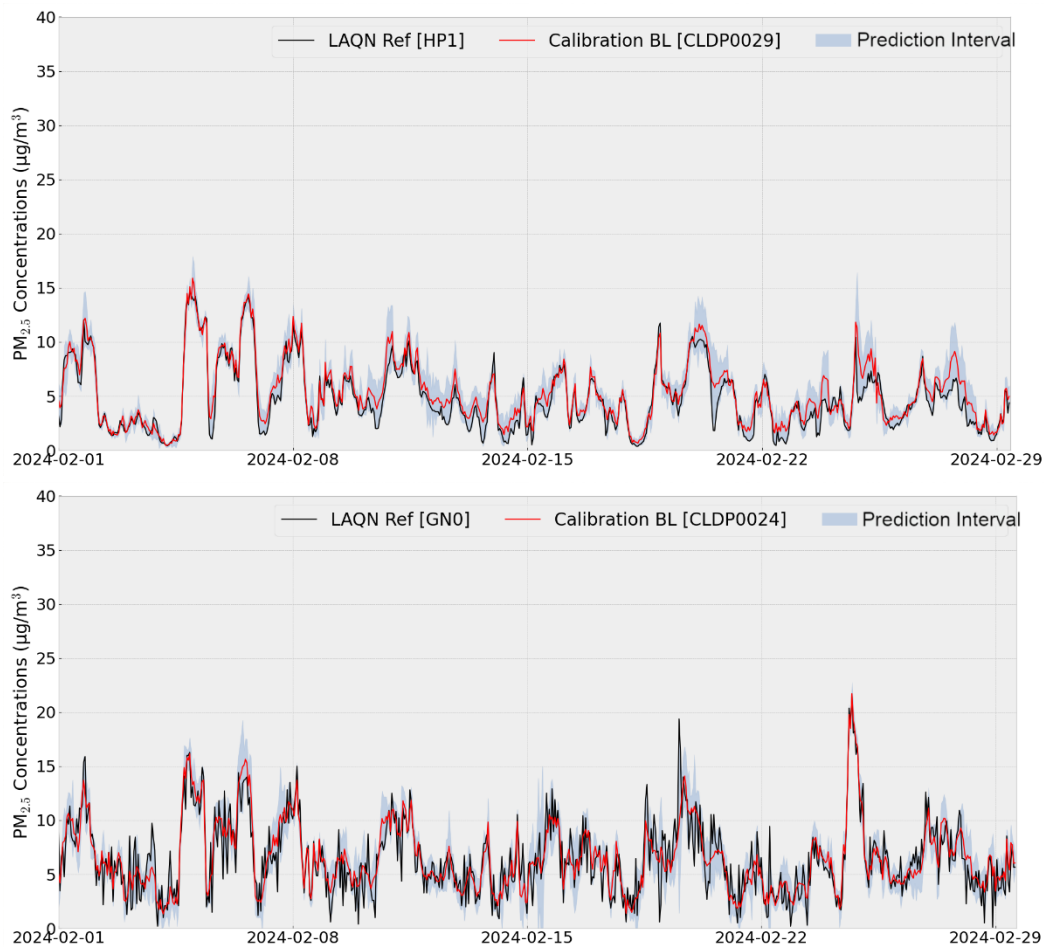
4.3.1 Aleatoric Uncertainty in Low-cost Sensor Measurements

The training dataset consists of calibrated measurements from the 19 co-located BL nodes described in Chapter 2, alongside their corresponding reference measurements from co-located LAQN stations. The environmental covariates (temperature and humidity) were also sourced from the BL nodes, as detailed in Chapter 2. The independent test dataset includes additional calibrated measurements from BL nodes with nearby LAQN stations, as illustrated in Section 2.3.2.

I trained a random forest regression model to quantify the uncertainty associated with calibrated low-cost sensor measurements. The details of random forest implementation are provided in Section 2.2.2 and are not repeated here. Specifically, uncertainty is represented by the predicted standard deviation σ , computed as the square root of the estimated variance r_t^2 .

To intuitively visualize uncertainty, I derive a Prediction Interval (PI) for each calibrated sensor measurement. Assuming approximately normally distributed errors, this interval is calculated as $\hat{y} \pm \sigma$, where $\hat{y}$ denotes the calibrated sensor measurement produced by the Transformer-based calibration model described in

Section 2.2.3. This means the prediction interval of calibrated sensor measurement lies within the shaded band around $\hat{y}$ defined by $\pm\sigma$. Figures 4.5 and 4.6 show example time series of the calibrated BL measurements against reference measurements, with the corresponding prediction interval bands. They illustrate that the vast majority of reference measurements fall inside the prediction intervals, demonstrating good empirical coverage.



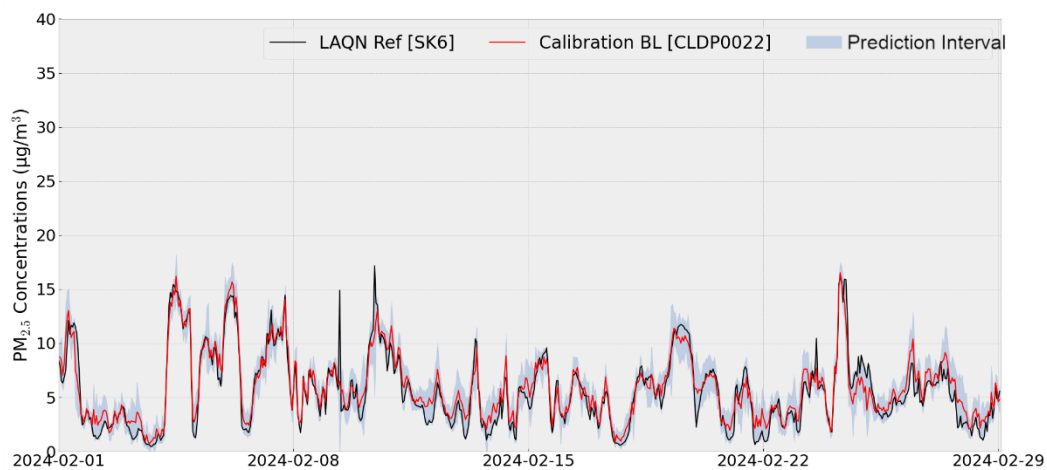
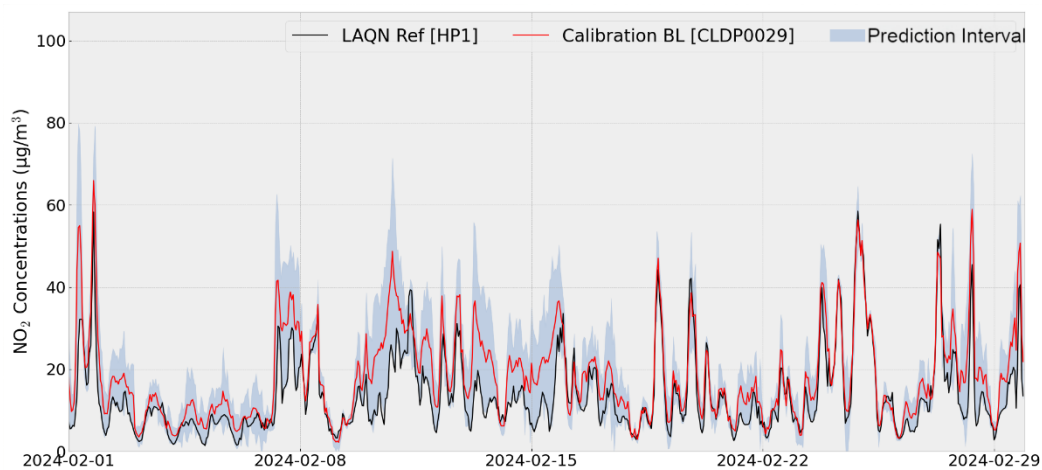


Figure 4.5 Calibrated Breathe London PM_{2.5} measurements with corresponding prediction interval at 3 random selected stations HP1, GN0, and SK6 over February 2024.



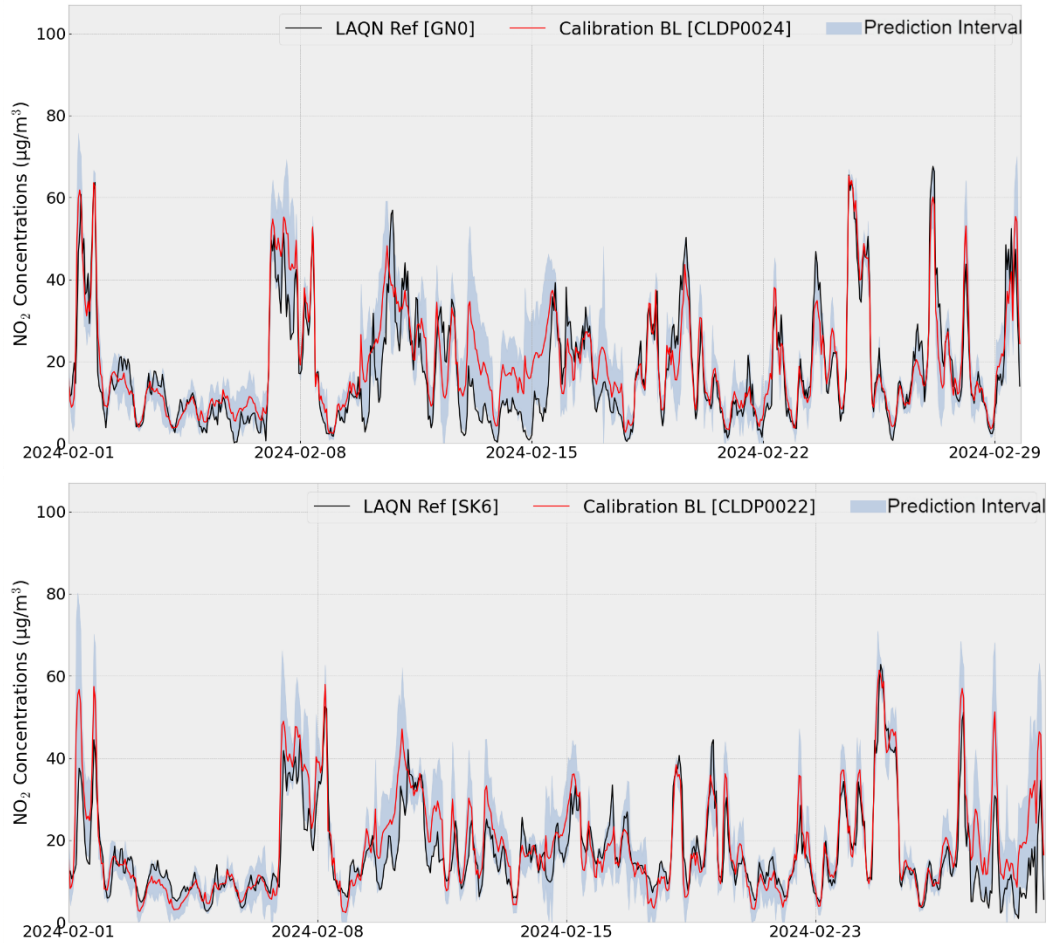


Figure 4.6 Calibrated Breathe London NO₂ measurements with corresponding prediction interval at 3 random selected stations HP1, GN0, and SK6, February 2024.

Finally, I incorporate these uncertainty estimates into my broader framework for uncertainty-aware air quality mapping. In the downstream Bayesian Maximum Entropy (BME) data fusion process, each calibrated sensor measurement is treated as an observation with an associated uncertainty - the predicted standard deviation - which is used as an input to BME, a geostatistical data fusion method that can weigh data points by their uncertainty. By providing BME with a standard deviation (SD) for each measurement (by modeling the sensor reading as a Gaussian-distributed “soft” observation with y_t^{lcs} and SD), the measurements with higher uncertainty

contribute less to the fused air quality estimate, while more certain measurements are given greater influence.

4.3.2 Epistemic Uncertainty in Air Quality Model Estimation

I discuss the ViT-MCDO model's ability to provide reliable predictions and account for uncertainty through standard deviation estimates. Table 4.1 shows the uncertainty estimation for NO₂ and PM_{2.5} using Monte Carlo Dropout in the proposed ViT-based air quality model.

Table 4.1 Uncertainty estimation for NO₂ and PM_{2.5} using Monte Carlo Dropout in the proposed ViT-Based air quality model

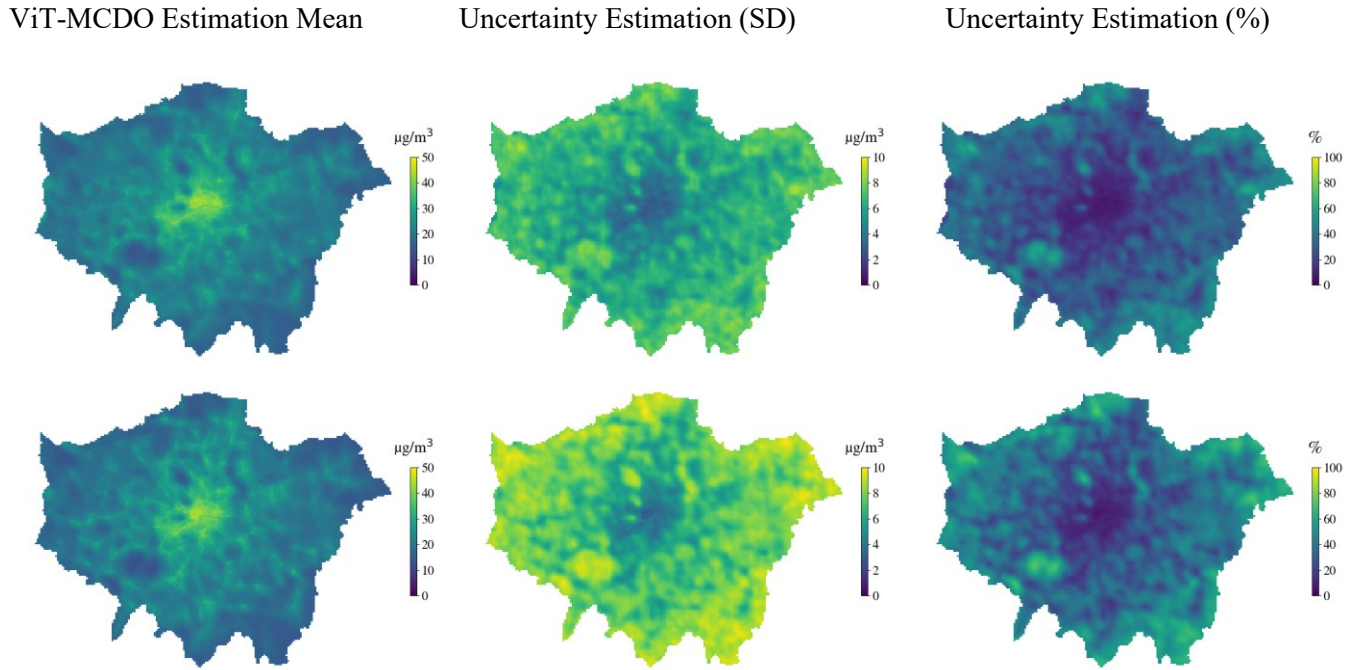
Pollution	Model	T	Dropout	RMSE↓	r↑	PICP↑	MPIW↓	CRPS↓
NO ₂	ViT	-	-	10.61	0.65	-	-	-
	ViT-	100	0.1	10.77	0.65	0.53	14.18	6.17
	MCDO	100	0.25	11.49	0.63	0.72	23.69	6.59
PM _{2.5}	ViT	-	-	3.09	0.79	-	-	-
	ViT-	100	0.1	3.18	0.79	0.61	4.49	1.71
	MCDO	100	0.25	3.84	0.75	0.71	7.72	2.13

* ↓: the lower the better, ↑: the larger the better. T: multiple inference times

When applying Monte Carlo Dropout (MCDO) for uncertainty estimation within a ViT-based model for NO₂ and PM_{2.5} predictions, the dropout rate significantly affects both prediction accuracy and uncertainty quantification quality. As shown in Table 4.1, increasing the dropout rate from 0.1 to 0.25 results in a marked improvement in the Prediction Interval Coverage Probability (PICP) for both pollutants, indicating that a larger proportion of true observations are captured within the prediction intervals. Specifically, the PICP improves from 0.53 to 0.72 for NO₂, and from 0.61 to 0.71 for PM_{2.5}. However, this coverage improvement comes at the expense of increased Mean Prediction Interval Width (MPIW) from 14.18 to 23.69 for NO₂, and from 4.49 to 7.72 for PM_{2.5}, suggesting that the model exhibits greater uncertainty at higher dropout rates.

ViT-MCDO (dropout=0.1) yields lower Continuous Ranked Probability Scores (CRPS; 6.17 for NO₂, 1.71 for PM_{2.5}) compared to dropout=0.25 (6.59 for NO₂, 2.13 for PM_{2.5}), demonstrating sharper probabilistic predictions at the lower dropout rate. Regarding accuracy, ViT-MCDO (dropout=0.1) not only performs comparably to the baseline ViT model but also surpasses ViT-MCDO (dropout=0.25). RMSE and Pearson correlation coefficient consistently indicate that higher dropout slightly reduces prediction accuracy.

Figure 4.7 presents estimated NO₂ concentration maps and corresponding uncertainty maps (SD) generated by ViT-MCDO (dropout=0.1, T=100) at selected hours. Distinct spatial uncertainty patterns illustrate the model's adaptive uncertainty quantification capability. Additional detailed results are available in Appendix E.



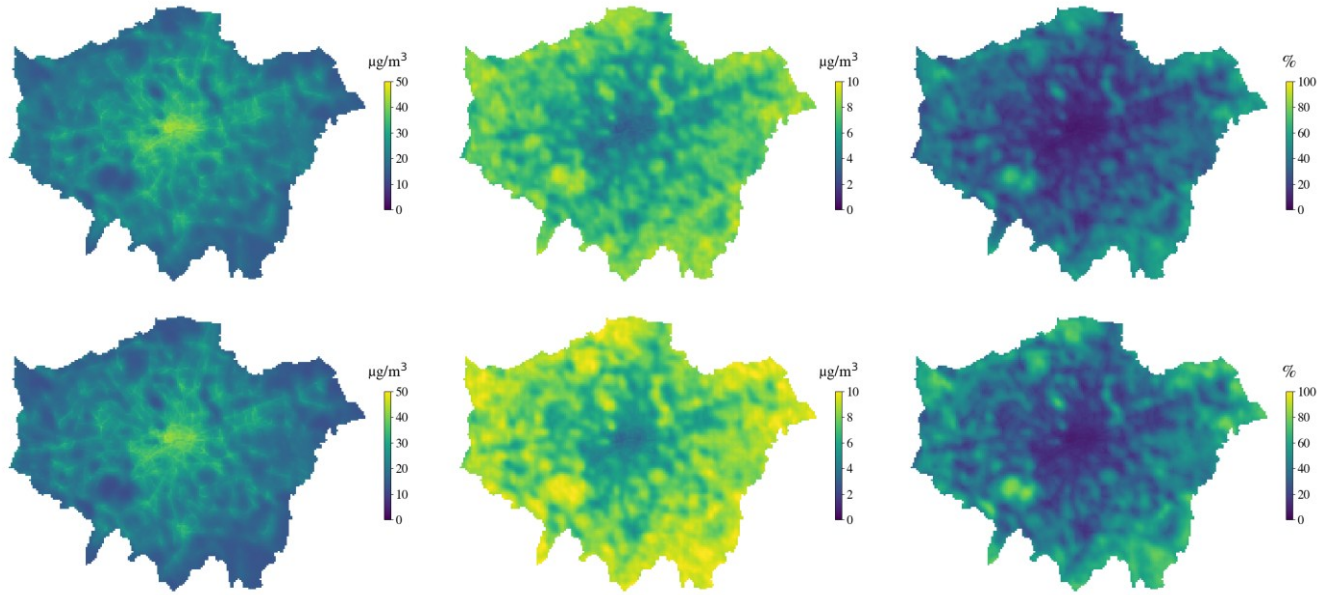


Figure 4.7 Estimated NO₂ Concentration and Uncertainty Maps Using Monte Carlo Dropout in the Proposed ViT-Based Air Quality Model at 9 AM, 12 PM, 3 PM, and 6 PM on 14 February 2024 (from top to bottom).

Finally, I incorporate these uncertainty estimates of ViT-MCDO into my broader framework for uncertainty-aware air quality mapping in the downstream BME data fusion process.

4.3.3 The Performance of Bayesian Maximum Entropy Framework for Integrating Data with Varied Uncertainties

Table 4.2 lists the statistics of the validation analysis for PM_{2.5} and NO₂ obtained by the BME and ViT-MCDO (T=100, dropout rate= 0.1) at test reference stations.

Table 4.2 Evaluation metrics for ViT-MCDO and Bayesian Maximum Entropy based on validation reference stations.

Pollution	Model	RMSE↓	r↑	PICP↑	MPIW↓	CRPS↓
PM _{2.5}	ViT-MCDO	3.11	0.63	0.46	4.75	1.85
	BME	2.29	0.70	0.33	2.27	1.41

NO ₂	ViT-MCDO	9.44	0.59	0.55	14.91	5.61
	BME	5.62	0.85	0.20	2.08	3.47

* ↓: the lower the better, ↑: the larger the better.

The BME model demonstrates improved performance over the standalone ViT-MCDO model across all accuracy metrics. Specifically, for PM_{2.5}, BME achieves a lower RMSE (2.29 vs. 3.11) and a higher correlation coefficient (0.70 vs. 0.63) compared to ViT-MCDO. This indicates that BME's predictions align more closely with reference values, as expected from multi-source data fusion. In terms of uncertainty quantification, BME also demonstrates superior performance. It produces substantially narrower prediction intervals, achieving a significantly lower Mean Prediction Interval Width (MPIW; 2.27 vs. 4.75 for ViT-MCDO). Additionally, BME yields a lower Continuous Ranked Probability Score (CRPS; 1.41 vs. 1.85), suggesting more accurate and sharper probabilistic predictions. However, this improved sharpness (narrower and more confident predictive distributions) comes at the cost of reduced coverage, with a lower PICP (0.33 vs. 0.46).

The NO₂ results exhibit similar trade-offs, with even more pronounced effects. BME achieves significantly lower RMSE (5.62 vs. 9.44) and higher correlation coefficient (0.85 vs. 0.59), suggesting substantial improvements in accuracy. This indicates that BME more effectively captures spatial patterns by integrating low-cost sensor data with reference observations. Similarly, BME yields a notably lower CRPS (3.47 vs. 5.60), suggesting improved probabilistic alignment with reference values. BME also produces substantially narrower uncertainty intervals (MPIW: 2.08 vs. 14.91). However, this increased sharpness reduces coverage (PICP: 0.20 vs. 0.55). Despite this trade-off, the substantially lower RMSE and CRPS demonstrate markedly more accurate and confident fused-data predictions from BME.

In summary, results in Table 4.2 strongly support the effectiveness of the BME framework in combining multiple data modalities for improved probabilistic air pollution prediction. While ViT-MCDO provides a solid deep learning foundation,

integrating its predictions with BME leads to substantial gains in accuracy and reliability, particularly for pollutants such as NO_2 with strong local variability.

Figure 4.8 presents spatial comparison of NO_2 concentration estimates and associated uncertainties (standard deviation, SD) from ViT-MCDO (dropout rate = 0.1, $T = 100$) and the BME approach at 17:00 on February 14, 2024. The results illustrate that the BME-derived uncertainties are notably smaller than those estimated by ViT-MCDO, reflecting the enhanced prediction of BME due to its integration of multiple data sources.

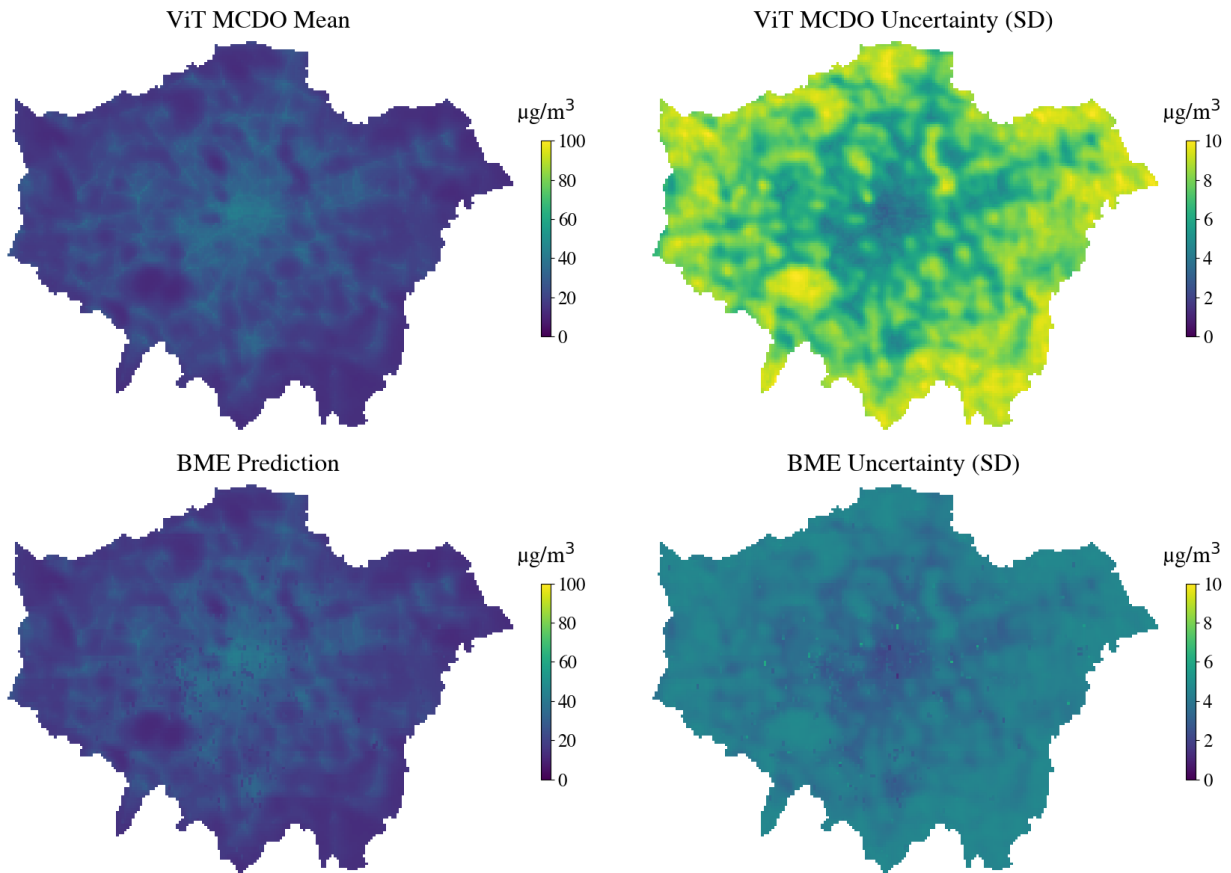


Figure 4.8 NO_2 estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, $T=100$) VS BME and at 5 PM, February 14, 2024

Relationship Between Estimated Concentration and Predictive Uncertainty

Figure 4.9 presents the relationship between the estimated NO₂ concentrations and their associated predictive uncertainties (expressed as standard deviation, SD) derived from the BME approach at 17:00 on February 14, 2024. The results reveal a strong inverse relationship between the estimated NO₂ levels and the uncertainty, as indicated by a Pearson correlation coefficient of $r = -0.93$.

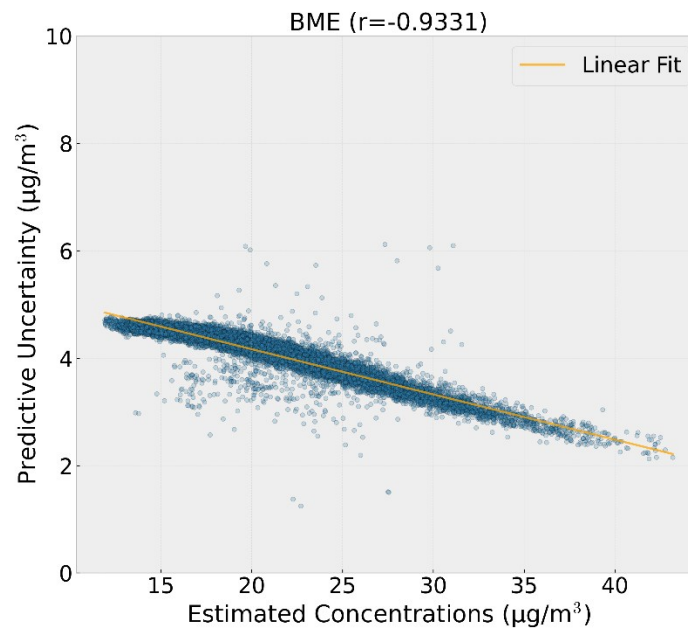


Figure 4.9 Relationship between estimated NO₂ concentrations and predictive uncertainty from BME at 5 PM, February 14, 2024

Relationship Between Sensor Density and Predictive Uncertainty

Figure 4.10 illustrates the relationship between BME-derived NO₂ predictive uncertainty and local sensor density, defined as the number of available sensors within a 3 km × 3 km spatial window, based on 24-hour averaged concentrations for 14 February 2024. A clear and statistically meaningful negative association is observed ($r = -0.6611$), indicating that regions with higher sensor density tend to exhibit substantially lower predictive uncertainty.

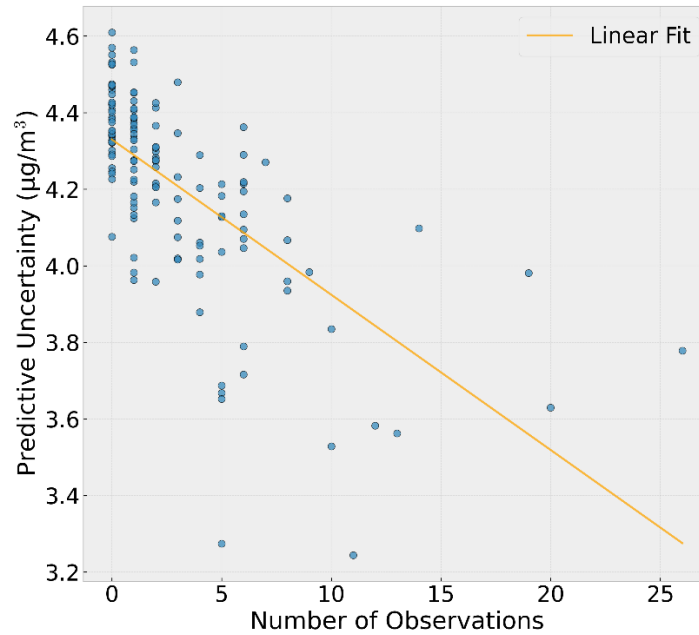


Figure 4.10 Relationship between sensor density and predictive uncertainty from BME (3KM grid, 24-hour average, February 14, 2024, $r=-0.6611$)

The choice of a $3\text{ km} \times 3\text{ km}$ aggregation scale represents a methodological trade-off between spatial representativeness and statistical robustness. Smaller windows would yield insufficient sensors, while larger windows would dilute local variability and reduce the number of independent samples available for analysis.

Through this integration, my approach produces high-resolution air quality maps with quantified uncertainty at every grid cell. This is a improvement over traditional methods that might ignore model uncertainty or sensor error. In my case, by estimating and fusing the aleatoric and epistemic uncertainties from sensor measurements and air quality model with BME framework, I get access to an uncertainty-aware pollution surface where areas of higher predictive uncertainty are explicitly indicated (through larger BME posterior standard deviations), guiding users to interpret the results with appropriate caution.

4.4 Discussion

This study introduces a novel, uncertainty-aware air quality mapping framework that simultaneously quantifies and integrates aleatoric uncertainty from low-cost sensor (LCS) measurements and epistemic uncertainty from a ViT-based air quality model through a Bayesian Maximum Entropy data fusion approach. By systematically integrating heterogeneous data sources, including measurements from low-cost sensors, reference monitors, and predictions from the ViT-based model, with their associated uncertainties, the BME framework generates pollutant concentration maps accompanied by quantified uncertainty at a spatial and temporal resolution of $100\text{ m} \times 100\text{ m} \times 1\text{ hour}$. Compared with the ViT-based air quality model presented in Chapter 3, the proposed BME data fusion framework has not only improved prediction accuracy but has also provided a corresponding uncertainty map, offering deeper insights into the reliability and confidence of air quality estimates.

Major Contributions

This work makes three significant contributions to the uncertainty analysis in air quality estimation:

1. Quantifying Aleatoric Uncertainty for All Low-cost Sensor Network

In conventional uncertainty estimation, it is common to assume a fixed error margin around sensor readings based on aggregate performance statistics (Peters et al., 2021). However, such static bounds fail to reflect sensor-specific behavior. In contrast, my approach estimates dynamic, instance-specific uncertainty that adapts to contextual factors such as ambient temperature, relative humidity and pollutant concentration levels.

Specifically, I model the variance of calibrated sensor errors using a random forest regression trained on co-located sensor-reference data. The model learns how the magnitude of the residual error varies as a function of environmental conditions, thereby capturing heteroskedasticity in sensor behavior. The predictors used include

the ambient temperature, relative humidity, and the calibrated LCS reading itself. As a result, the estimated error variance is no longer constant, but tailored to each measurement instance based on its environmental context.

Importantly, although this model is trained using co-located data, the learned relationships generalize across the full sensor network. This enables the estimation of adaptive uncertainty even for sensors that are not co-located with reference monitors. By providing uncertainty that dynamically adjust to local environmental conditions, my framework offers significantly improved interpretability over conventional fixed-margin methods, allowing users to identify when and where sensor measurements are more or less reliable, rather than relying on one-size-fits-all thresholds.

2. Qualifying Epistemic Uncertainty of ViT-based Air Quality Model

I incorporated uncertainty quantification techniques by employing Monte Carlo Dropout to estimate epistemic uncertainty in my proposed ViT-based air quality model. To evaluate the robustness of the uncertainty estimates, I experimented with different dropout rates (0.1 and 0.25) and performed 100 stochastic forward passes during inference. The results show that increasing the dropout rate leads to higher predictive variance, consistent with the findings of Mokhtari et al. (2021), suggesting that the model expresses greater uncertainty under higher dropout regimes.

Using a configuration of dropout rate = 0.1 and 100 inference samples, I generated high-resolution ($100\text{ m} \times 100\text{ m} \times 1\text{ hour}$) epistemic uncertainty maps for Greater London. To the best of my knowledge, this is the first fine-scale uncertainty map of air quality estimation for Greater London. The spatial distribution of uncertainty reveals lower uncertainty levels in central London, whereas higher uncertainty is observed in peripheral and background regions.

In addition to this spatial pattern, Figure 4.7 reveals an inverse relationship between predicted pollutant concentration and epistemic uncertainty, with lower uncertainty associated with higher concentration levels and higher uncertainty observed in low-

concentration regions. This relationship is closely linked to the deployment characteristics of the low-cost sensor network. More than 70% of the Breathe London sensors are located at roadside or traffic-influenced sites, where pollutant concentrations (particularly NO₂) are systematically higher. As a result, the model is trained on a dense set of observations in these high-concentration regimes, enabling it to learn more stable and confident representations of pollution dynamics in road-adjacent environments.

3. Conversely, lower-concentration areas, such as suburban and urban background regions, are relatively under-represented in the training data. In these regions, the model has fewer observations to constrain its predictions, leading to elevated epistemic uncertainty. The observed inverse correlation between concentration and uncertainty therefore reflects not only spatial variation in sensor density, but also the uneven coverage of concentration regimes within the training dataset. This pattern is consistent with the interpretation of epistemic uncertainty as a measure of model knowledge gaps arising from limited or imbalanced training data.
- Integrating Data with Varied Uncertainties**

This work introduces a novel uncertainty-aware framework for urban air quality modeling that quantifies aleatoric (low-cost sensor measurement) and epistemic (ViT-based model output) uncertainties, and then combines them using the Bayesian Maximum Entropy (BME) method. The BME framework assigns optimal weights to each data source—including low-cost sensors, reference monitors, and ViT-based model outputs—based on their associated uncertainties. This results in more accurate pollutant concentration estimates and better uncertainty quantification, consistent with findings in prior studies (Akita et al., 2014; Hanigan et al., 2017).

My approach bridges modern deep learning with geostatistical data fusion, integrating multiple uncertainty sources within a unified probabilistic framework. This hybrid system captures complex spatial patterns while simultaneously accounting for measurement variability and model uncertainty, yielding a

operationally meaningful solution for urban pollution mapping and evidence-based environmental decision-making.

The results highlight the value of jointly modeling multiple uncertainty sources—particularly the often-overlooked measurement uncertainty from LCS networks. To my knowledge, few studies have integrated low-cost sensor data at this scale due to data quality concerns (Gressent et al., 2020). By treating low-cost sensor measurements as probabilistic inputs with empirically derived error distributions, my framework improves both the reliability and interpretability of air quality assessments.

To further validate this reliability, I conducted a formal test of the relationship between sensor density and predictive uncertainty from BME. A negative correlation ($r = -0.66$) between the number of observations and the associated uncertainty is illustrated in Figure 4.10. Regions with higher sensor density exhibit substantially lower predictive uncertainty, as the BME posterior distribution becomes more tightly constrained by nearby observations. In contrast, areas with sparse monitoring coverage rely more heavily on prior assumptions and auxiliary inputs, resulting in wider uncertainty intervals.

Despite its advantages, the BME framework is computationally intensive. In my case, each hourly data fusion task required over 10 hours of runtime, limiting its real-time applicability. While my aim was to demonstrate methodological feasibility, future work should explore more efficient algorithms or approximate inference techniques to improve scalability.

Implications of Uncertainty Estimates for Inequities

Providing uncertainty estimates is essential for reliable air quality management and informed policy-making. For example, uncertainty information can be particularly valuable in applications concerned with environmental equity. To illustrate this, I compared the spatial distribution of NO₂ predictive uncertainty with the Index of

Multiple Deprivation (IMD, a widely used indicator of socioeconomic disadvantage) across Greater London (shown in Figure 4.9).

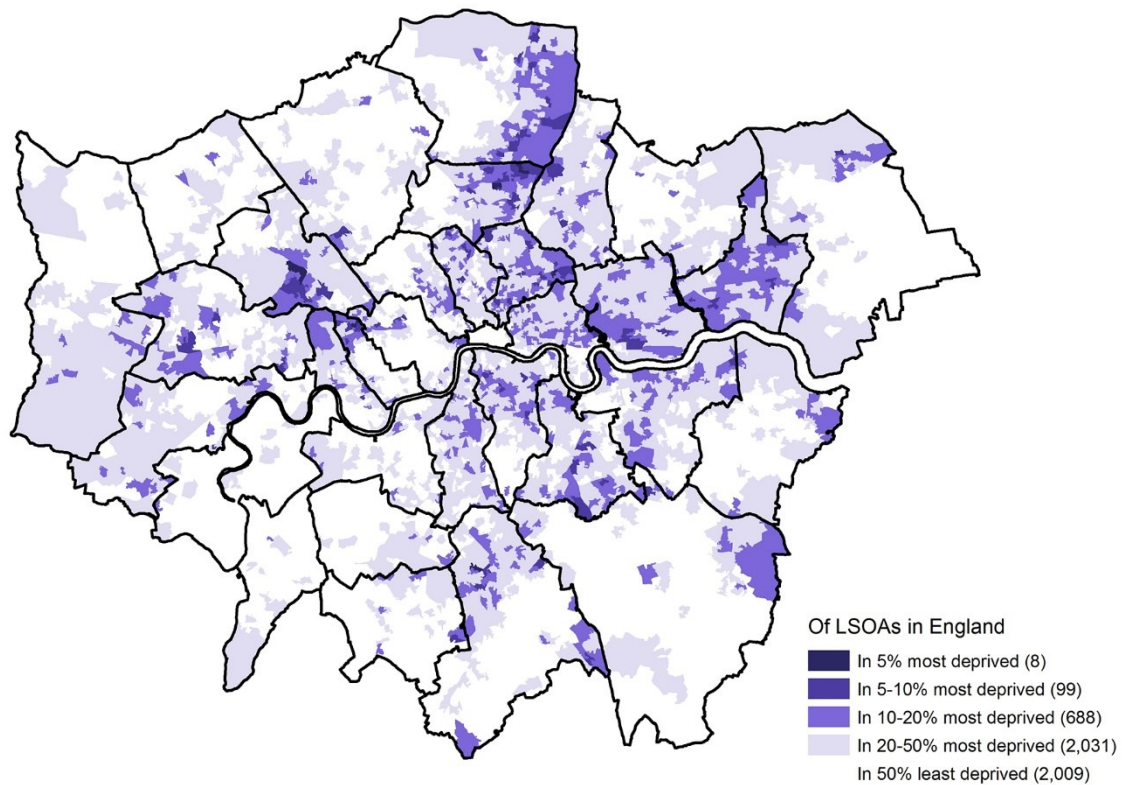


Figure 4.11 Map of Greater London depicting the distribution of relative deprivation by neighborhood (LSOA), based on IMD 2019.

I created a bivariate map that overlays the Index of Multiple Deprivation (IMD) with both predicted air pollution concentrations and their associated predictive uncertainties. Here, I used an NO₂ pollution map and its corresponding uncertainty map generated by BME at 5 PM on February 14, 2024. Figure 4.10 illustrates the joint spatial distribution of IMD with air pollution concentrations (A) and with uncertainty (B). I adopted a classic 3×3 bivariate color scheme, in which pollution or uncertainty (yellow scale) and IMD (blue scale) are each categorized into low,

medium, and high levels. Dark red areas indicate locations where both pollution (or uncertainty) and IMD levels are high.

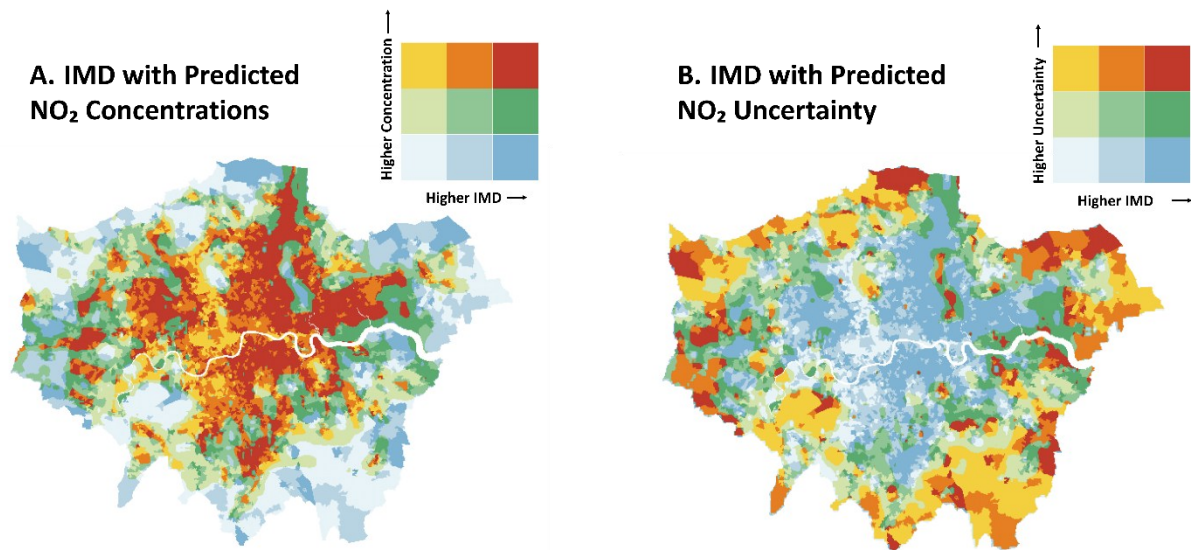


Figure 4.12 Bivariate maps illustrating the overlay of the Index of Multiple Deprivation with NO₂ pollution concentration (A) and predictive uncertainty (B) in Greater London at 5 PM on February 14, 2024.

When overlaid with IMD, several key patterns emerge:

- Many central London neighborhoods and areas along major roadways exhibit both elevated NO₂ concentrations and high levels of deprivation (as indicated by the dark red zones in Panel A). These locations represent clusters of a “double burden,” where residents are simultaneously exposed to severe air pollution and socioeconomic disadvantage.
- When IMD is overlaid with the uncertainty map (Panel B), no clear relationship is observed. In particular, highly deprived neighbourhoods do not necessarily exhibit higher predictive uncertainty.

I performed a correlation analysis between the BME-derived uncertainty and the IMD scores across London. IMD is used here as a proxy for SES, where higher scores represent higher levels of deprivation.

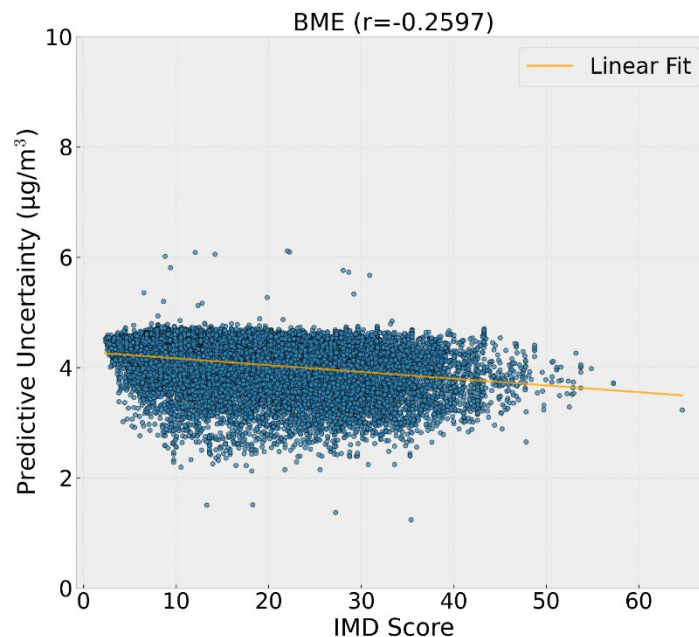


Figure 4.13 Relationship between IMD and predictive uncertainty from BME at 5 PM, February 14, 2024

As illustrated in Figure 4.13, the scatter plot shows that there is no strong or definitive relationship between a neighborhood's socioeconomic status and the model's predictive confidence. These findings invite a more nuanced interpretation of equity in uncertainty-aware air quality modelling. In contrast to a simple assumption that socioeconomically deprived communities are systematically associated with higher predictive uncertainty, the London case study does not provide clear evidence of such a direct inequity. Instead, the joint analysis of NO₂ concentrations, uncertainty, and the IMD reveals a more complex spatial structure.

This highlights an important methodological implication: uncertainty maps should not be interpreted as straightforward indicators of environmental injustice without careful consideration of the mechanisms that generate uncertainty. In the present

framework, uncertainty primarily reflects gaps in the model's knowledge across concentration regimes rather than the intentional or systematic under-monitoring of deprived communities. For example, affluent areas with consistently low concentrations (such as Richmond) may exhibit higher uncertainty despite relatively good infrastructure, illustrating that uncertainty-driven deployment strategies could, if applied uncritically, prioritise sensing in low-exposure areas rather than in communities facing the greatest pollution burden.

Overall, this example underscores that equity-aware evaluation in AI-based environmental modelling requires careful interpretation of uncertainty, grounded in both data-generating mechanisms and deployment context. Rather than assuming that higher uncertainty necessarily signals social inequity, the London case illustrates how uncertainty can reveal non-intuitive interactions between monitoring design, pollution regimes, and socio-spatial structure—reinforcing the need for cautious, evidence-based use of uncertainty metrics in policy applications.

Chapter 5: Conclusions and Future Research

5.1. Summary of Key Findings

In this dissertation, I set out to develop a **fine-scale spatiotemporal air quality modeling framework by combining big but noisy data**. Chapter 1 outlined the overarching problem: ambient air pollution varies on fine spatial and temporal scales, yet traditional monitoring networks (with a few high-quality stations) cannot capture this variability. I identified three specific objectives in Chapter 1 to address this gap: (1) improve the data quality of dense low-cost sensor networks through calibration, (2) develop an advanced deep learning model to map air quality at high resolution using large heterogeneous datasets, and (3) quantify uncertainty from LCS measurements and deep learning model predictions, and then employ an uncertainty-aware data fusion framework to systematically integrate data sources with varying levels of uncertainty. Chapters 2, 3, and 4 each tackled one of these objectives. This section summarizes the key findings from those chapters and shows how each contributes to the field of air pollution mapping.

Chapter 2 – Sensor Calibration: Chapter 2 addressed the critical challenge of handling noisy and unstable data from low-cost sensors. Although these sensors offer substantial spatiotemporal coverage, their raw measurements commonly suffer from significant inaccuracies, including measurement drift and sensitivity to environmental factors such as temperature and humidity. To enhance data quality at scale, this chapter introduced a scalable calibration framework explicitly designed for extensive urban sensor networks. A key methodological advancement presented in this chapter was the development of a novel transformer-based calibration model, which effectively leverages temporal dependencies (e.g., diurnal cycles) present in sensor data. Empirical evaluation in $\text{PM}_{2.5}$ and NO_2 demonstrated that this transformer-based approach significantly improved sensor accuracy, achieving a reduction in RMSE of approximately 10–20% compared to conventional calibration techniques. Consequently, the calibrated outputs from these low-cost sensors closely

approached reference-grade performance, as evidenced by their high correlation with regulatory-grade monitors. Overall, the rigorous calibration methodologies established in Chapter 2 provided an essential foundation for subsequent modeling efforts, ensuring that the input sensor data employed in later chapters was of acceptable quality.

Beyond accuracy improvements, Chapter 2 also yielded important empirical insights into the robustness and practical applicability of large-scale calibration. Sensitivity analyses demonstrated that calibration performance degrades gracefully as the number of available reference co-location sites decreases, rather than failing abruptly. While increasing the number of reference sites improves calibration accuracy, acceptable performance can still be achieved with a substantially reduced reference network, particularly for PM_{2.5}. Additional experiments showed that the absence of auxiliary meteorological inputs (temperature and relative humidity) disproportionately affects NO₂ calibration, reflecting the stronger environmental sensitivity of electrochemical NO₂ sensors. Finally, calibration models trained using roadside-only reference sites were shown to generalize reasonably well to mixed site types, although with modest performance trade-offs. Collectively, these findings demonstrate that the proposed calibration framework is resilient to realistic data and infrastructure constraints, while also clarifying the conditions under which its performance may deteriorate.

Chapter 3 – Deep Learning (ViT-based) Model for Air Quality Mapping:

Building on the calibrated, abundant sensor data, Chapter 3 addressed Objective 2 by developing a deep learning model for fine-scale air quality mapping. I introduced novel model architecture based on a Vision Transformer combined with convolutional neural network components, designed to handle spatially distributed data. The rationale for using a ViT architecture was its strength in capturing complex patterns over space by using attention mechanism, while convolutional layers extracted localized features. The chapter detailed how I integrated various data sources – including calibrated sensor measurements, land-use and traffic information, and meteorological data – as inputs to the model. The ViT-based model was trained

to predict pollutant concentration fields ($\text{PM}_{2.5}$ and NO_2) over the study area at a 100m resolution. The key findings from Chapter 3 were very encouraging: the deep learning model was able to learn intricate spatial relationships and temporal trends in the data, yielding accurate high-resolution pollution maps. The model outperformed baseline approaches such as linear models, random forest or CNN models in test dataset, achieving lower prediction errors and better capturing of local pollution. For instance, the deep model could identify sharp concentration gradients near emission sources (like busy roadways). This demonstrated the feasibility and advantage of applying advanced deep learning (specifically a ViT architecture) to environmental modeling for fine-scale air quality distribution. Chapter 3 also provides a leave-one-feature-out sensitivity analysis shows that calibrated low-cost sensor observations are the most influential information source for both pollutants, while auxiliary variables contribute in pollutant-specific ways: traffic-related features are the most important secondary driver for NO_2 , whereas meteorological variables (wind direction, humidity, cloud cover) play a larger role in $\text{PM}_{2.5}$.

Chapter 4 – Bayesian Maximum Entropy for Uncertainty Analysis: While the deep learning model provided high-resolution pollution maps, an important next step for real-world applications is to characterize the uncertainty associated with these predictions. Given the increasing availability of low-cost sensor networks, there is a growing opportunity to develop integrated frameworks that make intelligent use of large, yet potentially noisy, datasets and innovative air quality model. Chapter 4 addressed Objective 3 by introducing a Bayesian Maximum Entropy (BME) framework to perform uncertainty-aware data fusion. In this chapter, I first quantify data uncertainty for each LCS measurements and uncertainty of ViT-based model air quality estimation and then employ BME framework to systematically integrate data sources with varying levels of uncertainty. The key findings of Chapter 4 highlight the power of this integration: by applying BME, I generated pollution concentration maps that not only incorporate information from the predictive model and sensor data but also provide a probability distribution (with mean and standard deviation) for pollutant levels at every location and time. An important result was

that the BME-fused estimates were more accurate and reliable than using the air quality model alone. For example, the BME approach reduced prediction error compared to the stand-alone deep model by adjusting biases where the model deviated from LCS and reference measurements. While previous studies have focused on improving accuracy, few have considered quantify uncertainty, and none have attempted to jointly integrate aleatoric uncertainty from low-cost sensor measurements and epistemic uncertainty from model predictions within a unified data fusion framework. This study is among the first to operationalize such integration through the BME approach, enabling the generation of spatiotemporally continuous pollutant concentration maps alongside corresponding uncertainty estimates. Chapter 4 fulfills the third research objective by contributing a novel uncertainty-aware methodology, which accounts for both large-scale but noisy sensor networks and deep learning models uncertainties. This approach enhances not only predictive accuracy but also the interpretability and reliability of urban air quality assessments.

Bringing these chapters together, the dissertation's key findings can be synthesized as follows: I successfully created an end-to-end framework to map urban air quality at fine spatial and temporal scales by harnessing various data sources and advanced algorithms. Each component of this framework addressed a specific challenge and contributed to the overall solution. Low-cost sensor networks, once properly calibrated, can serve as a dense information source capturing neighborhood-level pollution variations. Deep learning models, like the developed ViT-based architecture, can transform these calibrated data (along with ancillary inputs) into high-resolution pollution maps. And Bayesian Maximum Entropy data fusion can effectively integrate various uncertainty inputs, yielding optimally unbiased estimates and explicit uncertainties. Collectively, these findings are not only viable but also highly beneficial: it addresses the fine-scale air quality monitoring problem by combining the strengths of big data and advanced deep learning method. This work, therefore, provides a foundation for more informed environmental health

assessments and policy decisions, as city officials and researchers can obtain granular pollution information along with an understanding of its reliability.

5.2. Contributions to Knowledge

This dissertation makes several original contributions to the fields of air quality monitoring, environmental data science, and geospatial modeling. The innovations span methodological advances in sensor calibration and machine learning, as well as practical advances in how I can use air pollution sensors and models. The **key contributions to knowledge** are summarized below:

- **Innovative Low-Cost Sensor Calibration Methodology:** I developed a novel, scalable calibration framework for low-cost air quality sensor network. Prior to this work, the variability and drift in low-cost sensor data posed a major hurdle to their use for air quality monitoring (Castell et al., 2017b; Lewis & Edwards, 2016). Previous studies have recognized these limitations and proposed various calibration approaches, but most are either static, one-time calibrations based on short-term co-location with reference monitors (e.g., Castell et al., 2017b; Spinelle et al., 2015), or sensor-specific models that fail to generalize across time or space (Zimmerman et al., 2018b; Jiao et al., 2016). The cost and logistical complexity of maintaining calibration across large-scale networks represents a critical barrier to fully leveraging low-cost sensors (Spinelle et al., 2015; Zimmerman et al., 2018b). My contribution addresses these barriers with the following core innovations: (1) I designed a periodic re-calibration strategy that leverages co-located sensor-reference data to continuously update calibration models, thereby addressing key challenges such as sensor drift and long-term stability degradation. Unlike conventional one-time calibration methods, this approach allows for dynamic adaptation to changing environmental conditions and sensor aging over time. Furthermore, my results demonstrate that the trained calibration model can be successfully generalized across the entire low-cost sensor network, even for

sensors that were not part of the initial co-location set. (2) To further improve accuracy, I introduced a transformer-based model that explicitly captures sequential patterns in sensor response (e.g., diurnal variation). This is a significant advancement over prior methods such as random forests or linear models, which either ignore time entirely or treat time as an external covariate. My transformer model uses attention to weigh historical sensor readings and environmental conditions, dynamically adjusting calibration based on learned temporal context, which reduced RMSE by $\sim 15\text{--}20\%$ compared to RF and MLR baselines. Despite the proliferation of novel low-cost sensing technologies that promise dense spatial coverage at low deployment cost, so far few studies have been able to make good use of it (Castell et al., 2017c; Lewis & Edwards, 2016; Morawska et al., 2018). This contribution is significant because it enables large networks of affordable sensors to function as reliable monitoring system by mitigating their noise and biases through advanced calibration framework. This contribution provides a foundation for any study aiming to utilize increasing big but noisy data: it demonstrates how to unlock the value of noisy data by improving its quality.

- **Advanced Deep Learning Framework for High-Resolution Air Quality Mapping:** This study proposes a Vision Transformer–based deep learning framework that leverages over 400 calibrated Breathe London sensors to generate high-resolution ($100\text{ m} \times 100\text{ m} \times 1\text{ hour}$) urban air quality estimation. Although previous studies have applied convolutional neural networks to pollution data (Liang et al., 2023; Yi et al., 2018), these models are inherently limited in capturing long-range spatial dependencies (Luo et al., 2016). Graph neural networks (GNNs) also face challenges in capturing complex spatial patterns and often suffer from high computational complexity and reduced efficiency when scaled to large sensor networks (C. Wang et al., 2021). My ViT-based model (Chapter 3) is a pioneering application of transformer architecture for air quality modeling. The proposed architecture, a hybrid of convolutional layers and transformer encoder, is an innovation that

allows us to treat heterogeneous data (including low-cost sensor measurements, land use characteristics, traffic congestion data, meteorological data and so on) like sequence of images, capturing both local detail and broader contextual patterns. Methodologically, this is a significant advance: it demonstrates how complex dependencies in space can be learned from data, without relying solely on pre-defined spatial features or assumptions. Practically, the proposed framework demonstrates superior estimation performance compared to baseline approaches, including land use regression (LUR), random forest (RF), and CNNs. Furthermore, integrating calibrated measurements from over 400 low-cost sensors within the Breathe London network led to substantial performance gains. This finding highlights the critical role that dense low-cost sensor networks can play in enhancing the spatial resolution and predictive accuracy of urban air quality assessments. This work not only delivers a novel modeling tool to the research community but also offers methodological insights into how recent advances in deep learning and the proliferation of novel data sources (e.g., low-cost sensors) can be effectively leveraged for environmental applications. It represents a paradigm shift in air quality mapping—transitioning from coarse, regression-based approaches to high-fidelity, high-resolution maps generated through dense low-cost sensors and advanced deep learning architectures.

- **Uncertainty-Aware Data Fusion via Bayesian Maximum Entropy:** This study proposes a novel uncertainty-aware air quality mapping framework that jointly quantifies and integrates aleatoric uncertainty (originating from low-cost sensor measurements) and epistemic uncertainty (arising from a ViT-based deep learning model) within a unified Bayesian Maximum Entropy (BME) data fusion approach (Chapter 4). Past studies in air pollution mapping has typically emphasized improving predictive accuracy but overlook the confidence of those predictions (Ahn et al., 2025). Geostatistical approaches (e.g. kriging and Bayesian models) inherently provide uncertainty estimates alongside predictions, yet they rely mainly on sparse monitoring

data and may not fully exploit “big but noisy data” from low-cost sensors(Akita et al., 2014; Hanigan et al., 2017). The novel aspect here is treating the model-derived estimates and low-cost sensor measurements as “soft data” and fusing them with reference measurements. The result not only provides a best estimate of pollution concentration but also yields a measure of uncertainty (standard deviation) associated with that estimate at every point. Although BME has previously been used to combine land-use regression outputs with monitor observations(Akita et al., 2014), my contribution shows a practical way to combine “big but noisy” data from low-cost sensors with uncertainty estimates from deep learning-based air quality models. The result is a practical framework that produces detailed pollution concentration maps with uncertainty quantification. This added measure of uncertainty is crucial for decision-makers: it helps identify where the model’s estimates are reliable and where they should be interpreted with caution or supplemented by additional data(Batterman et al., 2020).

Each of these contributions advances the scientific community’s ability to monitor and understand air quality. Taken together, the work pushes the boundaries of what can be achieved with low-cost sensors and modern modeling: from methodological innovations like the ViT model and BME data fusion, to practically utilizing dense low-cost sensor networks. This study is especially well positioned to take advantage of novel data sources emerging from smart cities, digital cities and sensor-rich Internet-of-Things (IoT) infrastructures (Hassani et al., 2023; Sivarajah et al., 2017a), thereby enabling more adaptive, data-driven, and health-conscious urban living.

5.3. Limitations of the Study

Notwithstanding its successes, this study has several important limitations. These limitations arise from technical constraints, methodological choices, and the scope of my experiments. A frank discussion of these limitations is essential for interpreting the results and for guiding future work. Below, I outline the key

limitations associated with each major component of the dissertation, as well as overarching limitations and trade-offs. For each identified limitation, I also propose corresponding future research directions in Section 5.4.

- **Sensor Calibration Limitations:** The calibration approach in Chapter 2, while effective, has certain constraints. First, it relies on the availability of high-quality reference monitors for ground-truth data. In my case study, calibration models were derived from sensors co-located with reference stations; this means the calibration’s accuracy is highest in conditions similar to those in the training phase. If a sensor is deployed in an environment very different from the calibration conditions (e.g., different climate, pollutant mix, or extreme concentration ranges), the calibration model may not perform as well. Future work should explore adaptive learning approaches for sensor calibration to address these limitations, as discussed in Section 5.4.
- **Deep Learning Model Limitations:** The deep learning ViT-based model introduced in Chapter 3 also comes with limitations. One major consideration is the data requirement. Training such a model demands a large dataset of sensor observations and related features. In areas or regions without dense sensor coverage or historical data, the model’s training could be data-limited. Moreover, the model’s performance is sensitive to the quality of input data: any remaining biases or errors in the calibrated sensor data, or in the auxiliary inputs (e.g., if meteorological data are coarse or land-use data are outdated), could affect the model’s predictions. My model was developed and tested in London; if applied to a city with different characteristics (different sources of pollution, meteorology, or urban form), the model would likely need retraining or fine-tuning. Additionally, despite my best efforts at validation, the deep learning model remains a complex, non-linear predictor – essentially a “black box.” It is not always clear why it makes a certain prediction, which can be a drawback for scientific interpretation and for users who require explainable results. These challenges underscore the need for improving model interpretability and transferability, which are discussed in Section 5.4.

- **Bayesian Data Fusion and Uncertainty Analysis Limitations:** The BME-based framework in Chapter 4, while powerful, has its own set of limitations. A fundamental limitation is the assumption of statistical models: BME (and kriging/geostatistics in general) relies on specifying covariance functions and assuming the pollutant field behaves in a certain statistically smooth way. In reality, air pollution processes can violate these assumptions (for instance, sharp discontinuities at urban boundaries or sudden changes due to events). If the chosen covariance model is mis-specified, the BME results could be suboptimal – e.g., uncertainty might be under- or over-estimated, or the fused values might be biased. I did carefully fit covariance parameters based on data, but there is always uncertainty in whether the model form fully captures the true spatial structure. Another limitation is computational scalability: BME involves inverting covariance matrices whose size grows with the number of data points. With the high-density data I have (numerous sensor readings plus many model-predicted points), full BME computation can become very heavy. Therefore, extremely high-resolution mapping or very large regions could still pose a challenge. Additionally, integrating the deep learning model output as “soft data” required an estimate of the model’s error variance. In my work, I used MCDO to set this deep learning model’s uncertainty level. This is inherently approximate – if the estimate of the model uncertainty or sensor uncertainty was off, the data fusion might have either trusted the model too much or too little. Thus, the results are somewhat contingent on a correct quantification of input uncertainties. Finally, while BME gives a probabilistic output, it does not inherently solve any causality or mechanistic understanding – it is still an empirical estimator. Therefore, its usefulness is in improved estimation rather than explaining the underlying causes of pollution patterns (which were outside the scope but worth noting as a limitation of what this approach can provide). Future research directions related to these limitations are outlined in Section 5.4, with a focus on improving BME robustness and efficiency.

- **General Limitations and Trade-offs:** Across the entire study, there are broader limitations and trade-offs to acknowledge. One trade-off is between model complexity and practicality. I chose complex models (deep neural network, BME data fusion) to maximize accuracy and information, but this comes at the cost of higher computational requirements and greater system complexity. A highly complex framework might be harder to implement in real-time or in resource-constrained settings (e.g., smaller city agencies without big data infrastructure). Simpler methods might be less accurate but more deployable; my work favored accuracy and detail, which is a conscious choice that may limit immediate practical adoption until computational tools catch up. Another limitation is related to data coverage and representativeness. Even though the whole point of this study was finding ways to make use of big but messy data, a fundamental challenge remains: the data were not uniformly distributed across space or time. There were notable spatial gaps and time periods with limited sensor coverage, in which the framework could not fully compensate for the complete absence of observations across large areas. In such unmonitored regions, pollutant patterns may still be misrepresented or carry high uncertainty. This underscores the continued need for strategic sensor deployment and supplementary data collection in under-instrumented areas. In addition, my case study was limited to a particular urban region with a certain level of infrastructure and pollution characteristics. The lessons learned and methods developed, while generally applicable, might encounter unforeseen difficulties in a vastly different context (such as rural areas or developing cities with different technology constraints). I also acknowledge that my evaluation of results was constrained to available reference data and cross-validation; while I did rigorous testing, there remains uncertainty in evaluation – for instance, true validation of the fine-scale maps is difficult since no “ground truth” exists at that resolution apart from sensor network itself. This means there may be residual errors or overconfidence in some fine-scale features that I could not fully detect. Finally, there are socio-technical limitations: deploying a dense sensor network and an advanced

modeling framework involves considerations beyond my technical analysis, such as maintenance of sensors, public acceptance, and integration into policy – these were outside the scope of this dissertation but are real hurdles on the path from research to impact (Clements et al., 2017; Sivarajah et al., 2017b). In summary, while the dissertation achieved its goals, it is important to view the findings within these limitations. Recognizing them helps clarify the boundaries of the conclusions and provides motivation for the improvements suggested next.

5.4. Recommendations for Future Research

The research presented in this dissertation opens several avenues for further investigation. To ensure clarity and continuity, the following future research directions are presented in direct correspondence to the limitations identified in Section 5.3. For each limitation, I propose specific strategies to address or mitigate the issue in future studies. Key recommendations for future research include:

- **Adaptive and Continuous Sensor Calibration:** To address sensor drift and reduce reliance on fixed reference stations, future work should explore adaptive learning for sensor calibration. One idea is to implement continuous calibration updates: as new reference measurements or inter-sensor comparisons become available, the calibration model could be adjusted in real-time (Concas et al., 2021b). Techniques like online machine learning or recursive calibration algorithms could enable sensors to self-calibrate to an extent, using background concentration trends or overlapping sensor data (Becnel et al., 2019). Another research direction is the development of calibration models that incorporate sensor network redundancy – for instance, calibrating sensors against each other to detect and correct outliers or drift (if one sensor in an area starts deviating significantly from the others, it could be flagged for maintenance or adjustment) (Narayana et al., 2023). Additionally, laboratory calibration vs. field calibration trade-offs could be studied:

determining what minimal laboratory characterization could complement data-driven field calibration to improve stability (Taştan, 2025). Adaptive calibration approaches will become increasingly important as sensor networks age; implementing machine learning systems that detect drift (perhaps by monitoring the residuals of the calibration model over time) and then correct for it automatically is a promising future development (De Vito et al., 2020). These advances would ensure that the “big data” feeding into the modeling framework remains trustworthy over long deployment periods.

- **Enhancing Deep Learning Model Interpretability and User Engagement:** Although the focus of this dissertation was on accuracy and uncertainty, future work should also consider the interpretability of the models and the usability of the outputs. For the deep learning model, research could apply explainable AI techniques to extract insights about what drives the predictions (Z. Cheng et al., 2025). For instance, analyzing the attention weights in Vision Transformer might reveal which locations or features the model considers most important for predicting pollution at a target site. This could yield scientific insights (e.g., the model might implicitly learn that wind patterns or certain land-use types strongly influence the distribution) (Qiang et al., 2025). Making the model more transparent would also increase trust when presenting the results to stakeholders. On the output side, further work on how to visualize and communicate uncertainty is recommended (Preston & Ma, 2022). I have produced uncertainty maps, but tailoring these for decision-making (e.g., showing confidence levels in a way that a non-expert can grasp) is an area where interdisciplinary research with communication experts could be valuable (Korporaal et al., 2020). Additionally, integrating the model outputs into public platforms or policy workflows will require feedback from users – an opportunity for future researchers to study how high-resolution, uncertainty-tagged air quality information is actually used by city officials, or how it influences public behavior when provided through apps or websites. Learning from this feedback could guide modifications to the modeling

framework to better suit end-user needs (for example, providing forecasts, health risk indicators, or scenario analysis capabilities). In summary, making the framework not only powerful but also interpretable and user-friendly is a meaningful direction for extending this work.

- **Transferability to Other Cities and Generalizability:** Future studies should test and adapt the proposed framework in different cities or regions. Applying my calibration methods and deep learning model to new environments will assess how generalizable the approaches are. This could involve transfer learning or domain adaptation techniques – for example, using the model trained in this dissertation as a starting point and fine-tuning it with a small amount of data from a new city (Poelzl et al., 2025; Yadav et al., 2024). By comparing results across multiple cities (with different climates, topographies, and pollution sources), researchers can identify which components of the model are location-specific and which are universally applicable. Demonstrating transferability is crucial for broader adoption: if a model trained on one city can be quickly re-purposed for another, the framework becomes far more useful. Moreover, studying generalizability may reveal the need for global models that incorporate data from many cities, or it may highlight features (such as satellite-based inputs or globally available data) that help the model perform well everywhere (Njaime et al., 2025). Ultimately, research into scaling my approach to national or global levels – while maintaining accuracy – would be an ambitious but impactful direction.
- **Multi-Pollutant Modeling and Extensions of Uncertainty Analysis:** My framework focused on a single pollutant (e.g., $\text{PM}_{2.5}$ or NO_2). In reality, cities face multiple air pollutants, and there are relationships between them. A valuable future research direction is to extend the modeling framework to multiple pollutants simultaneously (R. Li et al., 2024). This could involve training a multi-task deep learning model that predicts, say, $\text{PM}_{2.5}$, PM_{10} , NO_2 , and O_3 all at once, possibly sharing representations since these pollutants may be influenced by common factors (like traffic or meteorology). Multi-

pollutant models could improve efficiency and even accuracy (if information from one pollutant helps predict another). On the uncertainty side, the BME framework can be expanded to handle multivariate data fusion, where correlations between pollutants are modeled. For example, knowing O₃ levels might inform expected ranges of NO₂, and vice versa. Additionally, future research could apply the uncertainty quantification approach to pollutants with different dynamics (such as O₃, which has strong photochemical behavior) to see if the current covariance modeling is sufficient or if new formulations are needed. Another extension is to incorporate additional data sources in the uncertainty analysis: for instance, integrating satellite-derived aerosol measurements or chemical transport model outputs as additional “soft data” in the BME framework (Hanigan et al., 2017). This would test the framework’s ability to assimilate a wider array of information. Overall, broadening the scope to multiple pollutants and data types will move the framework closer to a comprehensive air quality management tool and will challenge its components to be more flexible and robust.

- **Real-Time Deployment Strategies:** Transitioning from an academic study to an operational system requires research into real-time implementation. One recommendation is to develop a real-time version of the modeling pipeline. This entails handling streaming data from sensors, performing on-the-fly calibration, running the deep learning model continuously (or at short intervals), and updating the BME data fusion as new data arrive. Future work should address the engineering challenges of this deployment: for example, optimizing the deep model to run on faster time scales (potentially using model compression or faster hardware), and employing incremental BME techniques that update estimates without having to rerun a full batch geostatistical model from scratch (Zhong et al., 2016). Research could explore using edge computing (processing data on devices near the sensors) to distribute the computational load, or cloud-based services that can dynamically scale with data input (Idrees et al., 2018). Another aspect is

reliability and fail-safe mechanisms: in real-time use, sensors may go offline or data may arrive late. Developing strategies for graceful degradation (so that the system still provides estimates with higher uncertainty when data are missing, and flags issues to users) would be important (Buelvas et al., 2023).

- **Sensor Network Design and Active Learning:** With an uncertainty-aware model in hand, a compelling future direction is to use it to optimize sensor placements (Guestrin et al., 2005). Research can be done on adaptive network design, where one uses the model’s uncertainty outputs to decide where adding new sensors would most reduce uncertainty. This concept turns the modeling framework into a tool for guiding investments in monitoring infrastructure. By simulating the impact of a candidate new sensor (e.g., using BME to predict how much uncertainty in surrounding areas would drop), I could iteratively build a more cost-effective network. Future work might involve developing algorithms that identify “hotspots of uncertainty” and then solve for the optimal location of additional sensors, possibly subject to practical constraints. This merges with the concept of active learning, where the model can suggest where obtaining more data (through new sensors or temporary mobile monitoring campaigns) would be most informative. Over time, such an approach could dramatically improve map accuracy by intelligently deploying resources. Additionally, as sensor technology evolves (for instance, wearable sensors or sensors on vehicles), integrating these mobile or dynamic data sources is another research avenue. Mobile sensors could fill in spatial gaps but add complexity since their location changes (Van den Bossche et al., 2015). Adapting the framework to handle moving sensors or temporally irregular data would expand its applicability. Ultimately, this line of research aims to create a virtuous cycle: model informs measurement, and new measurements improve the model.
- **Evaluating and mitigating potential disparities:** Another future direction could be to assess the fairness and equity of air quality predictions across demographic and socioeconomic groups. This involves evaluating whether

prediction errors systematically vary across regions with differing levels of deprivation, sensor density, or population vulnerability. If such biases are detected, future work could explore fairness-aware learning approaches, such as reweighting, adversarial debiasing, or region-specific loss functions, to mitigate disparities. This would help ensure the modelling framework supports equitable public health interventions and environmental justice.

- **Implications and Potential Applications of the ViT-based Air Quality Modelling Framework:** The proposed ViT-based air quality modelling framework has important implications for real-world air quality management. Because the model operates on continuously updated low-cost sensor streams and auxiliary datasets, it can, in principle, be deployed in a near-real-time setting, enabling the generation of hourly, fine-scale pollution maps with minimal latency. A further promising direction for future development is to extend the current system from nowcasting to short-term forecasting, such as producing 24-hour predictive air pollution maps. From a modelling perspective, the attention mechanism at the core of the ViT architecture offers a key advantage for forecasting: it can flexibly capture long-range temporal dependencies, thereby supporting the development of a near-real-time air quality forecasting system. Such a capability could support real-time exposure management, inform dynamic public-health alerts, and assist transport and environmental agencies in anticipating and responding to pollution episodes before they occur.

Incorporating Additional Datasets into the ViT Framework: Another promising direction for enhancing the ViT-based air quality modelling framework is the integration of additional large-scale and high-resolution datasets, particularly satellite observations and outputs from physics-based chemical transport models. Satellite observations—such as Aerosol Optical Depth from MODIS, TROPOMI-derived NO₂—provide continuous spatial coverage at regional to global scales, which capture large-scale pollution transport patterns that are often missed by ground-based networks. Similarly, physics-based model outputs, such as those from

CMAQ, WRF-Chem, simulate emissions, diffusion, chemical transformation—that cannot be directly inferred from data-driven methods alone. Although these models may exhibit biases due to uncertainties in emissions inventories or meteorological inputs, their coarse-resolution predictions offer physically grounded representations of large-scale pollution dynamics. Incorporating satellite retrievals and physics-based model outputs as additional input channels to the ViT model may therefore improve its ability to generalize across spatial domains and reduce errors in areas with sparse sensor coverage, effectively leveraging the complementary strengths of data-driven learning and atmospheric science knowledge. This thesis focuses on outdoor air quality by integrating big data and advanced modeling techniques, representing a paradigm shift in air quality monitoring and one which I hope will serve as a blueprint for global cities to protect public health and guide urban environmental policy. By improving transferability, adapting to sensor drift, expanding to multiple pollutants, enabling real-time operation, optimizing sensor deployment, and enhancing interpretability, future researchers can build upon and extend the contributions made here. Such efforts will move us closer to a future where cities worldwide deploy intelligent sensor networks and modeling frameworks to continuously monitor air quality with high fidelity, enabling data-driven decisions that protect public health, support targeted interventions, and foster more sustainable urban environments.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., & Acharya, U. R. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.

Abu-Hani, A., Chen, J., Balamurugan, V., Wenzel, A., & Bigi, A. (2024). Transferability of ML-based Global Calibration Models for NO₂ and NO Low-Cost Sensors. *Atmospheric Measurement Techniques Discussions*, 2024, 1–23.

Afshari, A., Vogel, J., & Chockalingam, G. (2023). Statistical downscaling of SEVIRI land surface temperature to WRF near-surface air temperature using a deep learning model. *Remote Sensing*, 15(18), 4447.

Ahn, S., Kim, A., Chung, Y., Kang, C., Kim, S., Kwon, D., Park, J., Oh, J., Park, J., Moon, J., Song, I., Min, J., Lee, H. J., Kim, H., & Lee, W. (2025). Nationwide Machine Learning-Ensemble PM_{2.5} Mapping Prediction and Forecasting Models in South Korea with High Spatiotemporal Resolution and Health Risk Estimation-Based Evaluations. *Environment & Health*.
<https://doi.org/10.1021/envhealth.4c00201>

Akita, Y., Baldasano, J. M., Beelen, R., Cirach, M., De Hoogh, K., Hoek, G., Nieuwenhuijsen, M., Serre, M. L., & De Nazelle, A. (2014). Large Scale Air Pollution Estimation Method Combining Land Use Regression and Chemical Transport Modeling in a Geostatistical Framework. *Environmental Science & Technology*, 48(8), 4452–4459. <https://doi.org/10.1021/es405390e>

Akita, Y., Chen, J.-C., & Serre, M. L. (2012). The moving-window Bayesian maximum entropy framework: Estimation of PM_{2.5} yearly average concentration across the contiguous United States. *Journal of Exposure Science & Environmental Epidemiology*, 22(5), 496–501.

- Ali, S., Alam, F., Potgieter, J., & Arif, K. M. (2024). Leveraging Temporal Information to Improve Machine Learning-Based Calibration Techniques for Low-Cost Air Quality Sensors. *Sensors*, 24(9), 2930.
- Andersen, Z. J., Bønnelykke, K., Hvidberg, M., Jensen, S. S., Ketzel, M., Loft, S., Sørensen, M., Tjønneland, A., Overvad, K., & Raaschou-Nielsen, O. (2012). Long-term exposure to air pollution and asthma hospitalisations in older adults: A cohort study. *Thorax*, 67(1), 6–11.
- Andersen, Z. J., Hvidberg, M., Jensen, S. S., Ketzel, M., Loft, S., Sørensen, M., Tjønneland, A., Overvad, K., & Raaschou-Nielsen, O. (2011). Chronic Obstructive Pulmonary Disease and Long-Term Exposure to Traffic-related Air Pollution: A Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, 183(4), 455–461. <https://doi.org/10.1164/rccm.201006-0937OC>
- Andersen, Z. J., Kristiansen, L. C., Andersen, K. K., Olsen, T. S., Hvidberg, M., Jensen, S. S., Ketzel, M., Loft, S., Sørensen, M., Tjønneland, A., Overvad, K., & Raaschou-Nielsen, O. (2012). Stroke and Long-Term Exposure to Outdoor Air Pollution From Nitrogen Dioxide: A Cohort Study. *Stroke*, 43(2), 320–325. <https://doi.org/10.1161/STROKEAHA.111.629246>
- Appel, K. W., Bash, J. O., Fahey, K. M., Foley, K. M., Gilliam, R. C., Hogrefe, C., Hutzell, W. T., Kang, D., Mathur, R., & Murphy, B. N. (2020). The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3. 1: System updates and evaluation. *Geoscientific Model Development Discussions*, 2020, 1–41.
- Appel, K. W., Napelenok, S. L., Foley, K. M., Pye, H. O., Hogrefe, C., Luecken, D. J., Bash, J. O., Roselle, S. J., Pleim, J. E., & Foroutan, H. (2017). Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geoscientific Model Development*, 10(4), 1703–1732.
- Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., & Hamburg, S. P. (2017a). High-

Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environmental Science & Technology*, 51(12), 6999–7008. <https://doi.org/10.1021/acs.est.7b00891>

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., & Hamburg, S. P. (2017b). High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environmental Science & Technology*, 51(12), 6999–7008. <https://doi.org/10.1021/acs.est.7b00891>

Araki, S., Yamamoto, K., & Kondo, A. (2015). Application of Regression Kriging to Air Pollutant Concentrations in Japan with High Spatial Resolution. *Aerosol and Air Quality Research*, 15(1), 234–241. <https://doi.org/10.4209/aaqr.2014.01.0011>

Bachechi, C., Rollo, F., & Po, L. (2024). HypeAIR: A novel framework for real-time low-cost sensor calibration for air quality monitoring in smart cities. *Ecological Informatics*, 81, 102568.

Bai, L., Huang, L., Wang, Z., Ying, Q., Zheng, J., Shi, X., & Hu, J. (2020). Long-term field Evaluation of Low-cost Particulate Matter Sensors in Nanjing. *Aerosol and Air Quality Research*, 20(2), 242–253. <https://doi.org/10.4209/aaqr.2018.11.0424>

Batterman, S., Berrocal, V. J., Milando, C., Gilani, O., Arunachalam, S., & Zhang, K. M. (2020). Enhancing models and measurements of traffic-related air pollutants for health studies using dispersion modeling and Bayesian data fusion. *Research Reports: Health Effects Institute*, 2020, 202.

Becnel, T., Sayahi, T., Kelly, K., & Gaillardon, P.-E. (2019). A recursive approach to partially blind calibration of a pollution sensor network. *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, 1–8. https://ieeexplore.ieee.org/abstract/document/8782523/?casa_token=klgH0z_yMJ

MAAAAA:zlZb9Q7Ynl1jsit5n0XsZr7kpz0mGD2E0XZJToU5JxOzOf21t7y34ub1
QOsganvVWi9pUdJJCw8

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., & Marcon, A. (2013). Development of NO₂ and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—The ESCAPE project. *Atmospheric Environment*, 72, 10–23.

Bertazzon, S., Johnson, M., Eccles, K., & Kaplan, G. G. (2015). Accounting for spatial effects in land use regression for urban air pollution modeling. *Spatial and Spatio-Temporal Epidemiology*, 14, 9–21.

Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating Low-Cost Sensor Measurements into High-Resolution PM_{2.5} Modeling at a Large Spatial Scale. *Environmental Science & Technology*, 54(4), 2152–2162. <https://doi.org/10.1021/acs.est.9b06046>

Bos, I., De Boever, P., Int Panis, L., & Meeusen, R. (2014). Physical Activity, Air Pollution and the Brain. *Sports Medicine*, 44(11), 1505–1518. <https://doi.org/10.1007/s40279-014-0222-6>

Breathe London. (2024). *Breathe London*. Breathe London. <https://www.breathelondon.org>

Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment*, 151, 1–11.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010). Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement From

the American Heart Association. *Circulation*, 121(21), 2331–2378. <https://doi.org/10.1161/CIR.0b013e3181dbee1>

Buelvas, J., Múnera, D., Tobón V., D. P., Aguirre, J., & Gaviria, N. (2023). Data Quality in IoT-Based Air Quality Monitoring Systems: A Systematic Mapping Study. *Water, Air, & Soil Pollution*, 234(4). <https://doi.org/10.1007/s11270-023-06127-9>

Burnett, R. T., Pope, C. A., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., Anderson, H. R., Smith, K. R., Balme, J. R., Bruce, N. G., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M. C., Gapstur, S. M., ... Cohen, A. (2014). An Integrated Risk Function for Estimating the Global Burden of Disease Attributable to Ambient Fine Particulate Matter Exposure. *Environmental Health Perspectives*, 122(4), 397–403. <https://doi.org/10.1289/ehp.1307049>

Byun, D., & Schere, K. L. (2006a). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59(2), 51–77.

Byun, D., & Schere, K. L. (2006b). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59(2), 51–77.

Caldeira, J., & Nord, B. (2020). Deeply uncertain: Comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology*, 2(1), 015002.

Carnevale, C., Angelis, E. D., Finzi, G., Turrini, E., & Volta, M. (2020). Application of Data Fusion Techniques to Improve Air Quality Forecast: A Case Study in the Northern Italy. *Atmosphere*, 11(3), 244. <https://doi.org/10.3390/atmos11030244>

Carnevale, C., Douros, J., Finzi, G., Graff, A., Guariso, G., Nahorski, Z., Pisoni, E., Ponche, J.-L., Real, E., & Turrini, E. (2016). Uncertainty evaluation in air quality planning decisions: A case study for Northern Italy. *Environmental Science & Policy*, 65, 39–47.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., & Bartonova, A. (2017a). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99, 293–302.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., & Bartonova, A. (2017b). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99, 293–302.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., & Bartonova, A. (2017c). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99, 293–302.

Cheng, Y., Li, X., Li, Z., Jiang, S., Li, Y., Jia, J., & Jiang, X. (2014). AirCloud: A cloud-based air-quality monitoring system for everyone. *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, 251–265. <https://doi.org/10.1145/2668332.2668346>

Cheng, Z., Wu, Y., Li, Y., Cai, L., & Ihnaini, B. (2025). A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. *Sensors*, 25(13), 4166.

Christakos, G. (1990). A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22(7), 763–777. <https://doi.org/10.1007/BF00890661>

Christakos, G., & Serre, M. L. (2000). BME analysis of spatiotemporal particulate matter distributions in North Carolina. *Atmospheric Environment*, 34(20), 3393–3406. [https://doi.org/10.1016/s1352-2310\(00\)00080-7](https://doi.org/10.1016/s1352-2310(00)00080-7)

Clements, A. et al. (2022). The Enhanced Air Sensor Guidebook. U.S. Environmental Protection Agency. https://cfpub.epa.gov/si/si_public_record_Report.cfm?dirEntryId=356426&Lab=C EMM

Clements, A. L., Griswold, W. G., Rs, A., Johnston, J. E., Herting, M. M., Thorson, J., Collier-Oxandale, A., & Hannigan, M. (2017). Low-cost air quality monitoring tools: From research to practice (a workshop summary). *Sensors*, 17(11), 2478.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., & Dandona, R. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082), 1907–1918.

Cohen, A. J., Ross Anderson, H., Ostro, B., Pandey, K. D., Krzyzanowski, M., Künzli, N., Gutschmidt, K., Pope, A., Romieu, I., Samet, J. M., & Smith, K. (2005). The Global Burden of Disease Due to Outdoor Air Pollution. *Journal of Toxicology and Environmental Health, Part A*, 68(13–14), 1301–1307. <https://doi.org/10.1080/15287390590936166>

Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P., & Tarkoma, S. (2021a). Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration: A Survey and Critical Analysis. *ACM Transactions on Sensor Networks*, 17(2), 1–44. <https://doi.org/10.1145/3446005>

Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P., & Tarkoma, S. (2021b). Low-Cost Outdoor Air Quality Monitoring and

Sensor Calibration: A Survey and Critical Analysis. *ACM Transactions on Sensor Networks*, 17(2), 1–44. <https://doi.org/10.1145/3446005>

Cordero, J. M., Borge, R., & Narros, A. (2018). Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensors and Actuators B: Chemical*, 267, 245–254.

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., & Jayne, J. T. (2017). Use of electrochemical sensors for measurement of air pollution: Correcting interference response and validating measurements. *Atmospheric Measurement Techniques*, 10(9), 3575–3588.

De Vito, S., Di Francia, G., Esposito, E., Ferlito, S., Formisano, F., & Massera, E. (2020). Adaptive machine learning strategies for network calibration of IoT smart air quality monitoring devices. *Pattern Recognition Letters*, 136, 264–271.

deSouza, P., Barkjohn, K., Clements, A., Lee, J., Kahn, R., Crawford, B., & Kinney, P. (2023). An analysis of degradation in low-cost particulate matter sensors. *Environmental Science: Atmospheres*, 3(3), 521–536.

DeSouza, P., Kahn, R., Stockman, T., Obermann, W., Crawford, B., Wang, A., Crooks, J., Li, J., & Kinney, P. (2022). Calibrating networks of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 15(21), 6309–6328.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., & Schwartz, J. (2020). Assessing NO₂ Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging. *Environmental Science & Technology*, 54(3), 1372–1384. <https://doi.org/10.1021/acs.est.9b03358>

- Do, T. H., Tsiligianni, E., Qin, X., Hofman, J., La Manna, V. P., Philips, W., & Deligiannis, N. (2020). Graph-deep-learning-based inference of fine-grained air quality from mobile IoT sensors. *IEEE Internet of Things Journal*, 7(9), 8943–8955.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (No. arXiv:2010.11929). arXiv. <https://doi.org/10.48550/arXiv.2010.11929>
- EPA. (2025). *Meteorological Process Overview*. <https://www.epa.gov/cmaq/meteorological-process-overview>
- Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L., & Popoola, O. (2016). Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sensors and Actuators B: Chemical*, 231, 701–713.
- Fukami, K., Maulik, R., Ramachandra, N., Fukagata, K., & Taira, K. (2021). Global field reconstruction from sparse sensors with Voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 3(11), 945–951.
- Gehring, U., Wijga, A. H., Brauer, M., Fischer, P., De Jongste, J. C., Kerkhof, M., Oldenwening, M., Smit, H. A., & Brunekreef, B. (2010). Traffic-related Air Pollution and the Development of Asthma and Allergies during the First 8 Years of Life. *American Journal of Respiratory and Critical Care Medicine*, 181(6), 596–603. <https://doi.org/10.1164/rccm.200906-0858OC>
- Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., & Subramanian, R. (2021a). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, 158, 105833.

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., & Subramanian, R. (2021b). From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. *Journal of Aerosol Science*, 158, 105833.

Goldberg, M. (2008a). A Systematic Review of the Relation Between Long-term Exposure to Ambient Air Pollution and Chronic Diseases. *Reviews on Environmental Health*, 23(4). <https://doi.org/10.1515/REVEH.2008.23.4.243>

Goldberg, M. (2008b). A Systematic Review of the Relation Between Long-term Exposure to Ambient Air Pollution and Chronic Diseases. *Reviews on Environmental Health*, 23(4). <https://doi.org/10.1515/REVEH.2008.23.4.243>

Googletraffic. (2024). *Googletraffic*. <https://dime-worldbank.github.io/googletraffic/index.html>

Guestrin, C., Krause, A., & Singh, A. P. (2005). Near-optimal sensor placements in Gaussian processes. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 265–272. <https://doi.org/10.1145/1102351.1102385>

Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4), 12–18.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2023). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>

Hanigan, I. C., Williamson, G. J., Knibbs, L. D., Horsley, J., Rolfe, M. I., Cope, M., Barnett, A. G., Cowie, C. T., Heyworth, J. S., & Serre, M. L. (2017). Blending multiple nitrogen dioxide data sources for neighborhood estimates of long-term

exposure for health research. *Environmental Science & Technology*, 51(21), 12473–12480.

Hankey, S., & Marshall, J. D. (2015). Land Use Regression Models of On-Road Particulate Air Pollution (Particle Number, Black Carbon, PM_{2.5}, Particle Size) Using Mobile Monitoring. *Environmental Science & Technology*, 49(15), 9194–9202. <https://doi.org/10.1021/acs.est.5b01209>

Hassani, A., Castell, N., Watne, Å. K., & Schneider, P. (2023). Citizen-operated mobile low-cost sensors for urban PM_{2.5} monitoring: Field calibration, uncertainty estimation, and application. *Sustainable Cities and Society*, 95, 104607.

Hatzopoulou, M., Valois, M. F., Levy, I., Mihele, C., Lu, G., Bagg, S., Minet, L., & Brook, J. (2017). Robustness of Land-Use Regression Models Developed from Mobile Air Pollutant Measurements. *Environmental Science & Technology*, 51(7), 3938–3947. <https://doi.org/10.1021/acs.est.7b00366>

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009. https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper

Heal, M. R., Kumar, P., & Harrison, R. M. (2012). Particles, air quality, policy and health. *Chemical Society Reviews*, 41(19), 6606–6630.

Health burden of air pollution in London | London City Hall. (2021, January 25). <https://www.london.gov.uk/programmes-and-strategies/environment-and-climate-change/environment-publications/health-burden-air-pollution-london>

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008a). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33), 7561–7578.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008b). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33), 7561–7578.

Hoek, G., Boogaard, H., Knol, A., De Hartog, J., Slottje, P., Ayres, J. G., Borm, P., Brunekreef, B., Donaldson, K., Forastiere, F., Holgate, S., Kreyling, W. G., Nemery, B., Pekkanen, J., Stone, V., Wichmann, H.-E., & Van Der Sluijs, J. (2010). Concentration Response Functions for Ultrafine Particles and All-Cause Mortality and Hospital Admissions: Results of a European Expert Panel Elicitation. *Environmental Science & Technology*, 44(1), 476–482. <https://doi.org/10.1021/es9021393>

Holmes, N. S., & Morawska, L. (2006). A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmospheric Environment*, 40(30), 5902–5928.

Idrees, Z., Zou, Z., & Zheng, L. (2018). Edge computing based IoT architecture for low cost air pollution monitoring systems: A comprehensive system analysis, design considerations & development. *Sensors*, 18(9), 3021.

Jacob, D. J., & Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric Environment*, 43(1), 51–63.

Jain, S., Presto, A. A., & Zimmerman, N. (2021a). Spatial Modeling of Daily PM_{2.5}, NO₂, and CO Concentrations Measured by a Low-Cost Sensor Network: Comparison of Linear, Machine Learning, and Hybrid Land Use Models. *Environmental Science & Technology*, 55(13), 8631–8641. <https://doi.org/10.1021/acs.est.1c02653>

Jain, S., Presto, A. A., & Zimmerman, N. (2021b). Spatial Modeling of Daily PM_{2.5}, NO₂, and CO Concentrations Measured by a Low-Cost Sensor Network: Comparison of Linear, Machine Learning, and Hybrid Land Use Models.

Environmental Science & Technology, 55(13), 8631–8641.
<https://doi.org/10.1021/acs.est.1c02653>

Jain, S., & Zimmerman, N. (2024). Exploration of intra-city and inter-city PM_{2.5} regional calibration models to improve low-cost sensor performance. *Journal of Aerosol Science*, 177, 106335.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahuvaroglu, T., Morrison, J., & Giovis, C. (2005a). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, 15(2), 185–204.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahuvaroglu, T., Morrison, J., & Giovis, C. (2005b). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology*, 15(2), 185–204.

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., & Davis, M. (2016). Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, 9(11), 5281–5292.

Johansson, L., Karppinen, A., Kurppa, M., Kousa, A., Niemi, J. V., & Kukkonen, J. (2022). An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation. *Environmental Modelling & Software*, 156, 105460.

Kang, J., & Choi, K. (2024). Calibration methods for low-cost particulate matter sensors considering seasonal variability. *Sensors*, 24(10), 3023.

Korporaal, M., Ruginski, I. T., & Fabrikant, S. I. (2020). Effects of uncertainty visualization on map-based decision making under time pressure. *Frontiers in Computer Science*, 2, 32.

Kwak, K.-H., Baik, J.-J., Ryu, Y.-H., & Lee, S.-H. (2015). Urban air quality simulation in a high-rise building area using a CFD model coupled with mesoscale meteorological and chemistry-transport models. *Atmospheric Environment*, 100, 167–177.

Lahoz, W. A., & Schneider, P. (2014). Data assimilation: Making sense of Earth Observation. *Frontiers in Environmental Science*, 2, 16.

Lanzi, E. (2016). *The economic consequences of outdoor air pollution. Organization for Economic Cooperation and Development.*

Lee, D., & Lee, B. (2024). Building reliable AI for quantifying uncertainty in particulate matter predictions with deep learning. *Journal of Cleaner Production*, 473, 143457.

Leelőssy, Á., Molnár, F., Izsák, F., Havasi, Á., Lagzi, I., & Mészáros, R. (2014). Dispersion modeling of air pollutants in the atmosphere: A review. *Open Geosciences*, 6(3). <https://doi.org/10.2478/s13533-012-0188-6>

Lewis, A., & Edwards, P. (2016). Validate personal air-pollution sensors. *Nature*, 535(7610), 29–31.

Li, R., Zhao, J., Feng, K., & Tian, Y. (2024). Development and application of a multi-task oriented deep learning model for quantifying drivers of air pollutant variations: A case study in Taiyuan, China. *Science of The Total Environment*, 920, 170777.

Li, T., Huang, X., Zhang, Q., Wang, X., Wang, X., Zhu, A., Wei, Z., Wang, X., Wang, H., & Chen, J. (2025a). Machine learning-guided integration of fixed and mobile sensors for high resolution urban PM_{2.5} mapping. *Npj Climate and Atmospheric Science*, 8(1), 95.

Li, T., Huang, X., Zhang, Q., Wang, X., Wang, X., Zhu, A., Wei, Z., Wang, X., Wang, H., & Chen, J. (2025b). Machine learning-guided integration of fixed and

mobile sensors for high resolution urban PM_{2.5} mapping. *Npj Climate and Atmospheric Science*, 8(1), 95.

Liang, L., Daniels, J., Bailey, C., Hu, L., Phillips, R., & South, J. (2023). Integrating low-cost sensor monitoring, satellite mapping, and geospatial artificial intelligence for intra-urban air pollution predictions. *Environmental Pollution*, 331, 121832.

Londonair. (2024). *London Regulatory Monitor Network*.
<https://www.londonair.org.uk/london/asp/publicdetails.asp>

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 29.
<https://proceedings.neurips.cc/paper/2016/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html>

Madrigano, J., Kloog, I., Goldberg, R., Coull, B. A., Mittleman, M. A., & Schwartz, J. (2013). Long-term Exposure to PM_{2.5} and Incidence of Acute Myocardial Infarction. *Environmental Health Perspectives*, 121(2), 192–196.
<https://doi.org/10.1289/ehp.1205284>

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S. P., Zimmerman, N., Kara, L. B., Presto, A. A., & Subramanian, R. (2019a). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, 12(2), 903–920.

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S. P., Zimmerman, N., Kara, L. B., Presto, A. A., & Subramanian, R. (2019b). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, 12(2), 903–920.

Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Robinson, A. L., Presto, A. A., & Subramanian, R. (2020). Fine particle mass monitoring with low-cost sensors:

Corrections and long-term performance evaluation. *Aerosol Science and Technology*, 54(2), 160–174. <https://doi.org/10.1080/02786826.2019.1623863>

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, 42(6), 1359–1369.

Masson, N., Piedrahita, R., & Hannigan, M. (2015a). Quantification method for electrolytic sensors in long-term monitoring of ambient air quality. *Sensors*, 15(10), 27283–27302.

Masson, N., Piedrahita, R., & Hannigan, M. (2015b). Quantification method for electrolytic sensors in long-term monitoring of ambient air quality. *Sensors*, 15(10), 27283–27302.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J. J., McLeod, M. W., Hodgson, T. F., & Dicks, J. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70, 186–203.

Milà, C., Ludwig, M., Pebesma, E., Tonne, C., & Meyer, H. (2024). Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls. *Geoscientific Model Development*, 17(15), 6007–6033.

Minet, L., Gehr, R., & Hatzopoulou, M. (2017). Capturing the sensitivity of land-use regression models to short-term mobile monitoring campaigns using air pollution micro-sensors. *Environmental Pollution*, 230, 280–290.

Mokhtari, I., Bechkit, W., Rivano, H., & Yaici, M. R. (2021). Uncertainty-aware deep learning architectures for highly dynamic air quality prediction. *IEEE Access*, 9, 14765–14778.

- Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D. M., & Fishbain, B. (2015). On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of The Total Environment*, 502, 537–547.
- Mooney, P., & Minghini, M. (2017). A review of OpenStreetMap data. *Mapping and the Citizen Sensor*, 37–59.
- Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., & Dunbabin, M. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International*, 116, 286–299.
- Muñoz-Villamizar, A., Solano-Charris, E. L., AzadDisfany, M., & Reyes-Rubiano, L. (2021). Study of urban-traffic congestion based on Google Maps API: The case of Boston. *IFAC-PapersOnLine*, 54(1), 211–216.
- Murad, A., Kraemer, F. A., Bach, K., & Taylor, G. (2021a). Probabilistic deep learning to quantify uncertainty in air quality forecasting. *Sensors*, 21(23), 8009.
- Murad, A., Kraemer, F. A., Bach, K., & Taylor, G. (2021b). Probabilistic deep learning to quantify uncertainty in air quality forecasting. *Sensors*, 21(23), 8009.
- Nalakurthi, N. V. S. R., Abimbola, I., Ahmed, T., Anton, I., Riaz, K., Ibrahim, Q., Banerjee, A., Tiwari, A., & Gharbia, S. (2024). Challenges and opportunities in calibrating low-cost environmental sensors. *Sensors*, 24(11), 3650.
- Narayana, M. V., Rachvarapu, K. K., Jalihal, D., & M, S. N. S. (2023). *Sens-BERT: Enabling Transferability and Re-calibration of Calibration Models for Low-cost Sensors under Reference Measurements Scarcity*. <https://doi.org/10.1109/JSEN.2024.3362962>
- Nazelle, A. D., Arunachalam, S., & Serre, M. L. (2010). Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: An Application

for Attainment Demonstration in North Carolina. *Environmental Science & Technology*, 44(15), 5707–5713. <https://doi.org/10.1021/es100228w>

Njaime, M., Abdallah, F., Snoussi, H., Akl, J., Chaaban, K., & Omrani, H. (2025). Transfer learning based solution for air quality prediction in smart cities using multimodal data. *International Journal of Environmental Science and Technology*, 22(3), 1297–1312.

Oliver, M. A., & Webster, R. (1990). Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4(3), 313–332. <https://doi.org/10.1080/02693799008941549>

Organization, W. H. (2021). *WHO global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. https://books.google.com/books?hl=en&lr=&id=s5pREAAQBAJ&oi=fnd&pg=PR5&dq=A+new+WHO+air+quality+model+confirms+that+92%25+of+the+world%27s+population+lives+in+places+where+air+quality+levels+exceed+WHO+limits&ots=qASFcUJjym&sig=-4ta1u_KK59vhnQ-I-UzS_dWfUk

Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J., & Batchellier, T. (2017). Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. *Sensors and Actuators B: Chemical*, 240, 829–837.

Patel, M. M., Chillrud, S. N., Correa, J. C., Hazi, Y., Feinberg, M., Kc, D., Prakash, S., Ross, J. M., Levy, D., & Kinney, P. L. (2010). Traffic-Related Particulate Matter and Acute Respiratory Symptoms among New York City Area Adolescents. *Environmental Health Perspectives*, 118(9), 1338–1343. <https://doi.org/10.1289/ehp.0901499>

Peters, D. R., Popoola, O. A., Jones, R. L., Martin, N. A., Mills, J., Fonseca, E. R., Stidworthy, A., Forsyth, E., Carruthers, D., & Dupuy-Todd, M. (2021). Evaluating

uncertainty in sensor networks for urban air pollution insights. *Atmospheric Measurement Techniques Discussions*, 2021, 1–23.

Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R. P., Lv, Q., & Hannigan, M. (2014). The next generation of low-cost personal air quality sensors for quantitative exposure monitoring. *Atmospheric Measurement Techniques*, 7(10), 3325–3336.

Poelzl, M., Kern, R., Kecorius, S., & Lovrić, M. (2025). Exploration of transfer learning techniques for the prediction of PM10. *Scientific Reports*, 15(1), 2919.

Pope, C. A., Ezzati, M., & Dockery, D. W. (2009). Fine-Particulate Air Pollution and Life Expectancy in the United States. *New England Journal of Medicine*, 360(4), 376–386. <https://doi.org/10.1056/NEJMsa0805646>

Preston, A., & Ma, K.-L. (2022). Communicating uncertainty and risk in air quality maps. *IEEE Transactions on Visualization and Computer Graphics*, 29(9), 3746–3757.

Puntanen, S., & Styan, G. P. H. (1989). The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator. *The American Statistician*, 43(3), 153–161. <https://doi.org/10.1080/00031305.1989.10475644>

pytorch. (2024). *Pytorch* [Computer software]. <https://pytorch.org/>

Qiang, Y., Li, C., Khanduri, P., & Zhu, D. (2025). *Interpretability-Aware Vision Transformer* (No. arXiv:2309.08035). arXiv. <https://doi.org/10.48550/arXiv.2309.08035>

Raaschou-Nielsen, O., Andersen, Z. J., Hvidberg, M., Jensen, S. S., Ketzel, M., Sørensen, M., Hansen, J., Loft, S., Overvad, K., & Tjønneland, A. (2011). Air pollution from traffic and cancer incidence: A Danish cohort study. *Environmental Health*, 10(1), 67. <https://doi.org/10.1186/1476-069X-10-67>

Righi, S., Lucialli, P., & Pollini, E. (2009). Statistical and diagnostic evaluation of the ADMS-Urban model compared with an urban air quality monitoring network. *Atmospheric Environment*, 43(25), 3850–3857.

Ryan, P. H., LeMasters, G. K., Levin, L., Burkle, J., Biswas, P., Hu, S., Grinshpun, S., & Reponen, T. (2008). A land-use regression model for estimating microenvironmental diesel exposure given multiple addresses from birth through childhood. *Science of the Total Environment*, 404(1), 139–147.

Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change*. John Wiley & Sons. [https://books.google.com/books?hl=en&lr=&id=n_RmCgAAQBAJ&oi=fnd&pg=PR5&dq=Seinfeld,+J.+H.+and+Pandis,+S.+N.+\(2006\).+Atmospheric+Chemistry+and+Physics:+From+Air+Pollution+to+Climate+Change.+John+Wiley+%26+Sons.&ots=gTJ6HPsdaO&sig=0jVMf4gptukR1_fZqkkdLaaGZq0](https://books.google.com/books?hl=en&lr=&id=n_RmCgAAQBAJ&oi=fnd&pg=PR5&dq=Seinfeld,+J.+H.+and+Pandis,+S.+N.+(2006).+Atmospheric+Chemistry+and+Physics:+From+Air+Pollution+to+Climate+Change.+John+Wiley+%26+Sons.&ots=gTJ6HPsdaO&sig=0jVMf4gptukR1_fZqkkdLaaGZq0)

Serre, M. L., & Christakos, G. (1999). Modern geostatistics: Computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 13(1–2), 1–26. <https://doi.org/10.1007/s004770050029>

Singh, S., Kumar, M., Verma, B. K., & Kumar, S. (2025). Optimizing Air Pollution Prediction With Random Forest Algorithm. *Aerosol Science and Engineering*. <https://doi.org/10.1007/s41810-025-00292-6>

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017a). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017b). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.

Sokhi, R. S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., Friedrich, R., Geels, C., Grönholm, T., & Halenka, T. (2021). Advances in air quality research—current and emerging challenges. *Atmospheric Chemistry and Physics Discussions*, 2021, 1–133.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., & Bonavitacola, F. (2015). Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. *Sensors and Actuators B: Chemical*, 215, 249–257.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.

Tan, Y., Lipsky, E. M., Saleh, R., Robinson, A. L., & Presto, A. A. (2014). Characterizing the spatial variation of air pollutants and the contributions of high emitting vehicles in Pittsburgh, PA. *Environmental Science & Technology*, 48(24), 14186–14194.

Taştan, M. (2025). Machine Learning–Based Calibration and Performance Evaluation of Low-Cost Internet of Things Air Quality Sensors. *Sensors*, 25(10), 3183.

Thurston, G. D., Kipen, H., Annesi-Maesano, I., Balmes, J., Brook, R. D., Cromar, K., De Matteis, S., Forastiere, F., Forsberg, B., & Frampton, M. W. (2017). A joint ERS/ATS policy statement: What constitutes an adverse health effect of air pollution? An analytical framework. *European Respiratory Journal*, 49(1). <https://publications.ersnet.org/content/erj/49/1/1600419.abstract>

Van den Bossche, J., Peters, J., Verwaeren, J., Botteldooren, D., Theunis, J., & De Baets, B. (2015). Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset. *Atmospheric Environment*, 105, 148–161.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

Vignesh, P. P., Jiang, J. H., & Kishore, P. (2023). Predicting PM_{2.5} Concentrations Across USA Using Machine Learning. *Earth and Space Science*, 10(10). <https://doi.org/10.1029/2023ea002911>

Volk, H. E., Lurmann, F., Penfold, B., Hertz-Picciotto, I., & McConnell, R. (2013). Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatry*, 70(1), 71–77.

Wang, C., Zhu, Y., Zang, T., Liu, H., & Yu, J. (2021). Modeling Inter-station Relationships with Attentive Temporal Graph Convolutional Network for Air Quality Prediction. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 616–634. <https://doi.org/10.1145/3437963.3441731>

Wang, W., Fecht, D., Beevers, S., & Gulliver, J. (2022). Predicting daily concentrations of nitrogen dioxide, particulate matter and ozone at fine spatial scale in Great Britain. *Atmospheric Pollution Research*, 13(8), 101506.

WHO, W. H. (2006). *Air quality guidelines: Global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. World Health Organization. [https://books.google.com/books?hl=en&lr=&id=7VbxUdlJE8wC&oi=fnd&pg=PR9&dq=WHO+\(2006\).+Air+quality+guidelines:+global+update+2005:+particulate+matter,+ozone,+nitrogen+dioxide,+and+sulfur+dioxide.+World+Health+Organization.&ots=w450vQMbuf&sig=3KpvG3oGJigxehyXGC-BXRzXnDA](https://books.google.com/books?hl=en&lr=&id=7VbxUdlJE8wC&oi=fnd&pg=PR9&dq=WHO+(2006).+Air+quality+guidelines:+global+update+2005:+particulate+matter,+ozone,+nitrogen+dioxide,+and+sulfur+dioxide.+World+Health+Organization.&ots=w450vQMbuf&sig=3KpvG3oGJigxehyXGC-BXRzXnDA)

Williams, M. L., Williams, M., Beevers, S., Kitwiroon, N., Dajnak, D., Walton, H., Lott, M. C., Lott, M., Pye, S., & Fecht, D. (2018). Public health air pollution impacts

of pathway options to meet the 2050 UK Climate Change Act target: A modelling study. *Public Health Research*, 6(7), 1–124.

Williams, R., Kilaru, V., Snyder, E., Kaufman, A., Dye, T., Rutter, A., Russell, A., & Hafner, H. (2014). Air sensor guidebook. *US Environmental Protection Agency*. https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=519616&Lab=NERL

Xu, Y., Serre, M. L., Reyes, J., & Vizuete, W. (2016a). Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: A National Application. *Environmental Science & Technology*, 50(8), 4393–4400. <https://doi.org/10.1021/acs.est.6b00096>

Xu, Y., Serre, M. L., Reyes, J., & Vizuete, W. (2016b). Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: A National Application. *Environmental Science & Technology*, 50(8), 4393–4400. <https://doi.org/10.1021/acs.est.6b00096>

Yadav, N., Sorek-Hamer, M., Von Pohle, M., Asanjan, A. A., Sahasrabhojane, A., Suel, E., Arku, R. E., Lingenfelter, V., Brauer, M., & Ezzati, M. (2024). Using deep transfer learning and satellite imagery to estimate urban air quality in data-poor regions. *Environmental Pollution*, 342, 122914.

Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018). Deep Distributed Fusion Network for Air Quality Prediction. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 965–973. <https://doi.org/10.1145/3219819.3219822>

Zhang, X., Chen, X., & Zhang, X. (2018). The impact of exposure to air pollution on cognitive performance. *Proceedings of the National Academy of Sciences*, 115(37), 9193–9197. <https://doi.org/10.1073/pnas.1809474115>

Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. *Alexandria Engineering Journal*, 93, 128–141.

Zheng, Y., Liu, F., & Hsieh, H.-P. (2013). U-Air: When urban air quality inference meets big data. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1436–1444. <https://doi.org/10.1145/2487575.2488188>

Zhong, X., Kealy, A., & Duckham, M. (2016). Stream Kriging: Incremental and recursive ordinary Kriging over spatiotemporal data streams. *Comput. Geosci.*, 90(May), 134–143.

Zimmerman, N., Li, H. Z., Ellis, A., Hauryliuk, A., Robinson, E. S., Gu, P., Shah, R. U., Ye, Q., Snell, L., Subramanian, R., Robinson, A. L., Apte, J. S., & Presto, A. A. (2020). Improving Correlations between Land Use and Air Pollutant Concentrations Using Wavelet Analysis: Insights from a Low-cost Sensor Network. *Aerosol and Air Quality Research*, 20(2), 314–328. <https://doi.org/10.4209/aaqr.2019.03.0124>

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2017). Closing the gap on lower cost air quality monitoring: Machine learning calibration models to improve low-cost sensor performance. *Atmos. Meas. Tech. Discuss*, 2017, 1–36.

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2018a). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1), 291–313.

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2018b). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1), 291–313.

Appendix A. Publications

Peer Reviewed

- Fan, Hongwei, et al. "ViTAE-SL: A vision transformer-based autoencoder and spatial interpolation learner for field reconstruction." *Computer Physics Communications* 308 (2025): 109464.
- Fan Hongwei, et al. "An Efficient ViT-Based Spatial Interpolation Learner for Field Reconstruction." International Conference on Computational Science. Cham: Springer Nature Switzerland, 2023.
- Lever, Jake, et al. "Facing & mitigating common challenges when working with real-world data: The Data Learning Paradigm." *Journal of Computational Science* 85 (2025): 102523.

Appendix B. Data Collection and Data Preprocessing

This appendix presents the methodology for downloading and processing the following datasets: Breathe London data, London Air Quality Network data, land use data, traffic congestion data, and meteorological data. It also details the procedures for spatial alignment across datasets and specifies all relevant parameter settings.

1. Standard Grid Generation

For consistency across analyses, I construct a standard rectangular grid covering the entire study region with uniform spacing. The grid generation process is as follows:

1. **Define Domain Extents:** Identify the geographic bounding box of the study area in longitude–latitude coordinates (WGS 84). Let x_{min} , x_{max} denote the minimum and maximum longitudes, and y_{min} , y_{max} denote the minimum and maximum latitudes, respectively. For Greater London region, I define:

Table B.1: Geographic Bounding Box for Greater London Study Area (WGS 84 Coordinates)

x_{min}	-0.520	The westernmost longitude of London
x_{max}	0.340	The easternmost longitude of London
y_{min}	51.275	The southernmost latitude of London
y_{max}	51.700	The northernmost latitude of London

2. **Project to Planar Coordinates:** Since longitude and latitude are angular measures, their spacing does not correspond to uniform physical distances. To enable construction of a regular grid with fixed spatial resolution (e.g., 100 meters), the bounding coordinates are first projected into a planar coordinate

system. In this study, I use the Universal Transverse Mercator (UTM) Zone 30N projection (EPSG:32630), which preserves local distances and areas over the UK region. Let $(x_{\min}^m, y_{\min}^m)$ and $(x_{\max}^m, y_{\max}^m)$ denote the projected bounding box coordinates in meters.

3. **Compute Grid Dimensions:** Calculate the number of grid intervals required along each axis to span the domain. Using the domain extents and resolution, I compute:

$$N_x = \left\lceil \frac{x_{\max}^m - x_{\min}^m}{R} \right\rceil,$$

$$N_y = \left\lceil \frac{y_{\max}^m - y_{\min}^m}{R} \right\rceil$$

where $\lceil \cdot \rceil$ denotes the ceiling function. This ensures that the grid fully covers the original region, even if the boundary does not align perfectly with the grid spacing. R denotes the grid resolution (in meters).

For the study region, projecting the London bounding box to UTM coordinates and applying a 100-meter grid resolution yields the grid $\mathbf{G}$ of size **[472, 594]**, meaning $N_x = 472$ and $N_y = 594$. This grid spans approximately 47.2~km $\times$ 59.4~km in physical space, providing a high-resolution basis for subsequent spatial modeling and analysis.

2. Breathe London Monitoring Data

This section describes the acquisition and processing of air quality data from the Breathe London network. The BL site monitoring metadata were obtained using the ListSensors and GetClarityData APIs of breathhelondon.org Table B.2 summarizes the usage of the two API endpoints. The methodology is presented in a step-by-step manner for clarity and rigor.

Table B.2: API Used for Retrieving Breathe London Sensor Data

API	URL
ListSensors	https://api.breathelondon.org/api/ListSensors
getClarityData	https://api.breathelondon.org/api/getClarityData/{SiteCode}/{Species}/{Start}/{End}/Hourly

1. **Data Retrieval via API:** I obtained pollutant concentration data using the official Breathe London API, ensuring a systematic retrieval of measurements for each sensor. The initial step involves using the ListSensors endpoint to obtain relevant metadata for all monitoring sites. Then the API's getClarityData endpoint was used to fetch data for each SiteCode (a unique sensor identifier) and Species (pollutant type). For example, SiteCode CLDP001 could denote a particular monitoring node, and Species IPM25 indicates particulate matter $\leq 2.5 \mu\text{m}$ (similarly, INO2 represents nitrogen dioxide). The API call requires specifying the sensor code, pollutant, start time, end time, and an averaging interval. For instance, to retrieve hourly averaged PM2.5 readings from sensor CLDP001 between 11:00 and 13:00 GMT on 11 April 2022, one would call:

```
1. GET https://api.breathelondon.org/api/getClarityData/CLDP001/IPM25/Mon 11 Apr 20
22 11:00:00 GMT/Mon 11 Apr 2022 13:00:00 GMT/Hourly?key={API_key}
```

In this example, CLDP001 is the sensor's SiteCode, IPM25 specifies the pollutant, the next two fields define the Start and End date-times of the query, and Hourly sets the data averaging period. The result is a JSON array of time-stamped concentration values for that sensor in the given interval. All study sensors' data were retrieved to build the full dataset.

2. **Spatial Mapping to Grid Cells:** Each sensor's coordinates were mapped onto a uniform grid covering the study area to facilitate spatial analysis. I assigned each sensor to the cell in which its location fell. This mapping provides a structured spatial framework (grid indices) for comparing and aggregating

data. I performed this assignment for all sensors, yielding a set of grid cells each containing zero, one, or multiple sensors.

3. **Aggregation of Overlapping Sensor Data:** For grid cells containing more than one sensor, I aggregated the overlapping sensor readings to produce a single representative time series per cell. When multiple sensors occupied the same cell, their measured pollutant concentrations at each time step were **averaged** so that the cell is treated as one observation point. Cells without any sensor coverage were recorded as having no direct measurement. After this step, the monitoring data were converted from individual sensor records to spatially gridded time series, suitable for subsequent analysis.

3. Reference Monitoring Station Data

I obtained reference air quality measurements from the London Air Quality Network, which provides hourly mean pollutant concentrations. These data are publicly available via the LondonAir website and can be accessed programmatically. I retrieved metadata (site code, name, type, latitude, longitude) for all LAQN monitoring stations and applied a geographic bounding box to filter sites located within Greater London (see Table B.1). The following R code demonstrates this retrieval and filtering process.

```
1. # Retrieve metadata for all LAQN sites
2. meta_df <- importMeta(source = "kcl")
3. # Filter sites to those within study area bounds
4. meta_df <- subset(meta_df, latitude >= LAT_MIN & latitude <= LAT_MAX &
5.                               longitude >= LON_MIN & longitude <= LON_MAX)
```

Next, I downloaded the hourly concentration data for the selected reference stations using the `openair importKCL` function. This API call connects to the LAQN data repository and returns a time series of hourly mean concentrations for specified site codes and years. All available sites within the study area were queried for the

pollutant of interest (e.g. NO₂) over the study period. The `importKCL` function automatically merges data from multiple sites and years. An example usage is shown below, where `site_codes` is the vector of filtered site identifiers:

```
1. # Download hourly NO2 data for selected sites (e.g., years 2020–2024)
2. site_codes <- meta_df$code
3. ref_data <- importKCL(site = site_codes, year = 2020:2024, pollutant = "NO2")
```

Each reference monitoring site was mapped to the predefined grid *G* used in my analysis.

4. Land Use Characteristic Data

OpenStreetMap (OSM) vector data (OpenStreetMap, 2023) were used to derive land-use characteristics for the standard grid *G*. The OSM dataset (covering Greater London) was obtained through the Geofabrik download service (Geofabrik, 2023), which provides pre-packaged extracts. I extracted **polygon features** (such as building and urban green spaces) and **line features** (such as roads and footpaths) relevant to the study.

Each OSM feature was then mapped to the standard grid *G*. For **polygon features** (e.g. building and parks), I performed an area-overlap operation per grid cell. Specifically, each polygon was intersected with the boundaries of grid cells that it overlaps. The area of the polygon falling within each cell was calculated, and these areas were summed by feature type per cell. For example, if a large park spans multiple cells, the area of the park in each cell is computed and attributed to those cells. Summing the areas of all building polygons in a cell gives the total building footprint area in that cell; similarly, summing all park/green-space polygons gives the total green space area. By dividing these values by the area of a grid cell, I can obtain a fractional land cover. This spatial intersection approach ensures that polygons contribute to every grid cell they occupy (proportionally to overlap), rather

than being assigned only to a single cell, thus capturing fine-grained land-use distribution on grid **G**

For **line features** (e.g. road centerlines and footpaths), I employed a similar approach using line-length within each cell. All road polylines were clipped to the grid cell boundaries – effectively splitting them wherever they crossed a cell border. I then calculated the length of each line segment lying inside each cell (using coordinate geometry in the projected coordinate system). These segment lengths were aggregated by cell and by feature category: summing the lengths of all road segments in a cell yields the total road length (in meters) for that cell. This method of coordinate matching via geometric intersection ensures that if a road runs through multiple cells, each cell gets the portion of that road’s length that lies within it. Minor pathways and major roads alike were handled in this unified way, so each grid cell has a quantified measure of road network density (total road length).

After processing, I compiled the selected polygon- and line-derived metrics into a standardized grid-based dataset. In other words, for each cell in grid **G** I now have a set of land-use attributes as shown in Table 3.2. For instance, building and green space features (originally polygons) are represented by their area contributions to the cell, while road features (originally lines) are represented by their total lengths in the cell. This standardized representation allows direct comparison and combination of disparate land-use features within the modeling framework.

5. Traffic Data

Traffic congestion data were obtained from the Google Maps Traffic layer via the Google Maps API. I used the ‘googletraffic’ R package to automate data collection, which converts Google traffic map imagery into a georeferenced raster dataset. In particular, the `gt_make_raster()` function was employed to capture traffic conditions for the study area at regular intervals. This function requires the center coordinates of the target region, the desired image dimensions in pixels (height and width), a

zoom level (map scale), and a valid Google API key with the Maps JavaScript API enabled. An example usage is shown below:

```
1. # Example: retrieve a 1000x1000 traffic raster at zoom level 11 for a given loca
   tion
2. r <- gt_make_raster(location = c(40.7128, 74.0060), # latitude, longitude of ce
   nter
3.                        height   = 1000,
4.                        width    = 1000,
5.                        zoom     = 11,
6.                        google_key = "YOUR_API_KEY")
```

This call returns a raster object r with traffic data for the specified location and time. Each pixel in r is assigned an integer code representing the traffic level (see Section 3.2.2). The raster comes georeferenced in latitude–longitude coordinates (WGS84), covering the map area defined by the location, height, width, and zoom parameters. I repeated this data retrieval at 1-hour intervals to capture temporal variation. In other words, snapshots of traffic conditions were recorded every 1 hour, each yielding a raster $r(t)$ of traffic congestion at time t . All snapshots were time-stamped and stored for subsequent spatial and temporal aggregation.

Spatial Resolution and Alignment: An appropriate zoom level and image size were chosen to ensure the traffic raster has sufficient spatial resolution for my analysis. The zoom level parameter z (an integer between 5 and 20) controls the map scale: higher values of z produce a more fine-grained map (each pixel represents a smaller area on the ground). For example, zoom $z = 11$ corresponds to roughly 76 meters in actual distance. In my case, I selected a zoom level 11 that provides a ground resolution on the order of only a few meters per pixel, which is fine enough to capture street-level traffic detail.

I align each traffic raster snapshot with the standard grid G . This was done via a spatial aggregation (zonal statistics) procedure: for each grid cell $g \in G$, I identified

all raster pixels whose center falls inside cell g . The traffic values of those pixels were then averaged to produce a single representative value for the grid cell. Formally, let $R_g(t)$ denote the traffic value of cell g at time t after resampling. If $P_g(t)$ is the set of all pixels in raster $R(t)$ that lie within cell g , I compute:

$$R_g(t) = \frac{1}{|P_g(t)|} \sum_{p \in P_g(t)} R_p(t)$$

where $R_p(t)$ is the traffic level value of pixel p at time t , and $|P_g(t)|$ is the number of pixels in grid cell g . This yields a spatially averaged traffic congestion index for each grid cell. If a particular cell g contains no roads, I assign $R_g(t) = 0$ by default, reflecting no congestion in that cell. By performing this resampling, I obtain a sequence of gridded traffic datasets on G , each corresponding to a 1-hour snapshot.

6. Meteorological Data

Data Retrieval via OpenWeatherMap API

Meteorological data were obtained using the OpenWeatherMap (OWM) Historical Weather API. The API endpoint structure used for retrieving hourly data is:

1. GET <https://history.openweathermap.org/data/2.5/history/city?lat={lat}&lon={lon}&type=hour&start={start}&end={end}&appid={API key}>

In this query, $\{lat\}$ and $\{lon\}$ are the latitude and longitude coordinates of the location of interest, $type=hour$ specifies an hourly data interval, $\{start\}$ and $\{end\}$ are the start and end timestamps (in Unix time, UTC) delimiting the requested time window, and $\{API key\}$ is the user's unique authentication key. The API returns JSON-formatted hourly weather data for the specified location and time range. According to the OWM API specifications, a single API call can retrieve at most one week of historical data in one response. If a longer period is required, the time span must be partitioned into multiple sequential one-week segments. In my data

acquisition, the full study period was divided into weekly intervals; separate API requests were made for each interval and the resulting data were concatenated to cover the entire period.

Spatial Sampling Strategy

To account for spatial variability in meteorological data across Greater London, I generated a grid of points spaced 1 km apart and retrieved historical meteorological data for each location. Bilinear interpolation was then applied to each hourly 1 km field for every meteorological variable, producing meteorological grid that is spatially aligned with the standard 100 m grid G .

Bilinear Interpolation

Bilinear interpolation is a simple grid-based interpolation method commonly used for regridding data and image resampling. It operates on continuous two-dimensional functions by performing linear interpolation in one direction, and then again in the other direction. In a rectilinear grid setting, if an unknown value is to be estimated at a point (x, y) that lies within a grid cell defined by four grid points $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$, $Q_{22} = (x_2, y_2)$, where $x_1 < x < x_2$ and $y_1 < y < y_2$, the bilinear interpolation formula is:

$$f(x, y) \approx f(Q_{11})(1 - \alpha)(1 - \beta) + f(Q_{12})(1 - \alpha)\beta + f(Q_{21})\alpha(1 - \beta) + f(Q_{22})\alpha\beta$$

where $\alpha = \frac{x - x_1}{x_2 - x_1}$ and $\beta = \frac{y - y_1}{y_2 - y_1}$. Here $f(Q_{ij})$ denotes the value at the grid point Q_{ij} .

Intuitively, the method first linearly interpolates in the x-direction (horizontal) to estimate values at the intermediate points directly above and below (x, y) , and then linearly interpolates those intermediate results in the y-direction (vertical). The weights w_{ij} for the four corners sum to 1, effectively forming a bilinear weighted average. Bilinear interpolation is computationally efficient and yields a smoother

surface than nearest-neighbor assignment because the output is continuous across cell boundaries (the first derivative is, however, discontinuous at the cell boundaries). It assumes the underlying surface varies linearly within each grid cell. While it does not account for complex spatial correlation structures, bilinear interpolation is often appropriate for smoothly varying fields and is widely used in remapping data from one grid to another.

The interpolation step preserves larger-scale meteorological patterns from OWM while providing localized estimates for the finer grid. The underlying assumption is that within a 1 km area, gradients in variables like temperature or wind are relatively linear. The linear interpolation ensures consistency and continuity of meteorological inputs across the model domain.

Appendix C. Low-Cost Sensor Calibration

The performance of low-cost PM_{2.5} calibration models

Table C.1 presents the PM_{2.5} performance metrics before and after calibration at each co-located Breathe London node.

BL Node ID	LAQN sites	BL Raw		BL Calibrated	
		RMSE	r	RMSE	r
LW2	CLDP0010	6.8801	0.7237	3.7901	0.7845
GR9	CLDP0011	7.0815	0.7954	2.8437	0.7723
BX9	CLDP0013	7.6323	0.8864	3.6939	0.7852
RB7	CLDP0014	9.8108	0.5198	3.2885	0.6860
MR9	CLDP0015	7.2366	0.8304	1.9551	0.8811
MR9	CLDP0016	7.8625	0.8049	1.9252	0.8759
TH4	CLDP0018	6.9457	0.7956	3.2498	0.8022
GR4	CLDP0019	7.7241	0.8119	3.3034	0.7768
GN6	CLDP0021	6.5006	0.7913	2.3836	0.8516
SK6	CLDP0022	6.6972	0.8832	2.1574	0.8932
ST5	CLDP0023	7.0700	0.7172	3.1028	0.7502
GN0	CLDP0024	6.8202	0.7647	2.8911	0.7986
HV1	CLDP0025	8.9969	0.6398	3.4157	0.7215
WM0	CLDP0026	7.3889	0.7809	3.2197	0.8102
KF1	CLDP0027	8.7630	0.8795	3.6774	0.8180
BT4	CLDP0028	7.6981	0.6336	3.6574	0.7421
HP1	CLDP0029	8.5922	0.7642	2.4093	0.8517
HP1	CLDP0030	7.1170	0.8646	2.1460	0.8921

Table C.1 Pearson correlation (r) and RMSE ($\mu\text{g}/\text{m}^3$) of PM_{2.5} before and after calibration at each co-located Breathe London node

Figure C.1 displays the distributions of raw and calibrated PM_{2.5} concentrations from BL nodes alongside the co-located LAQN reference monitors (2020–2024).

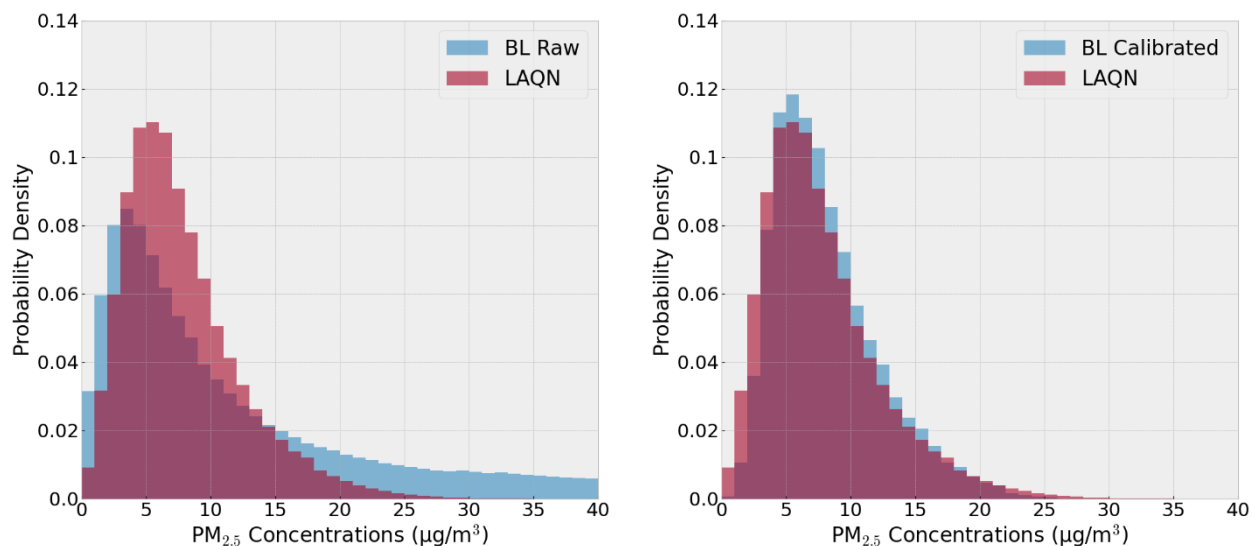
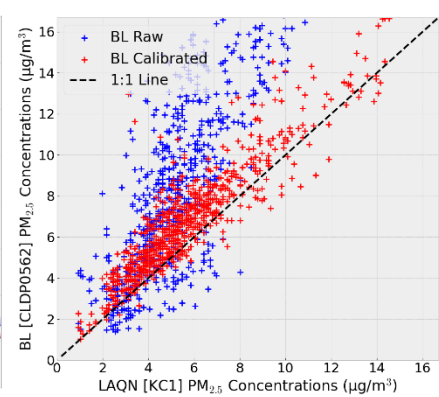
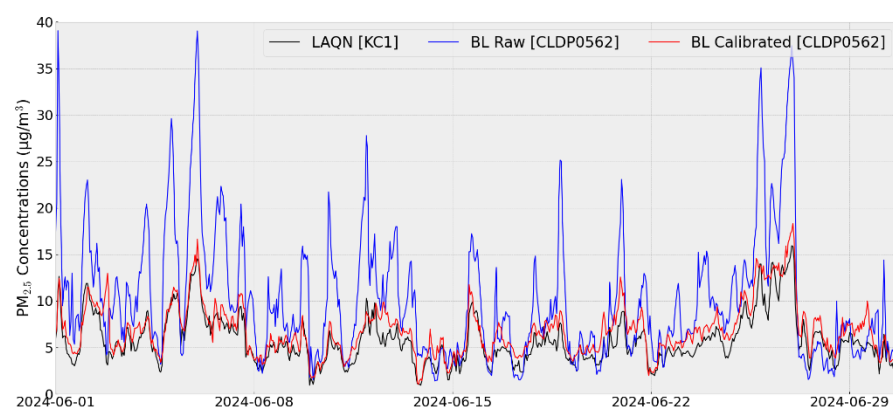
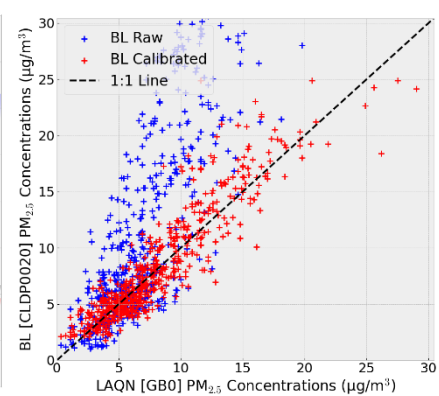
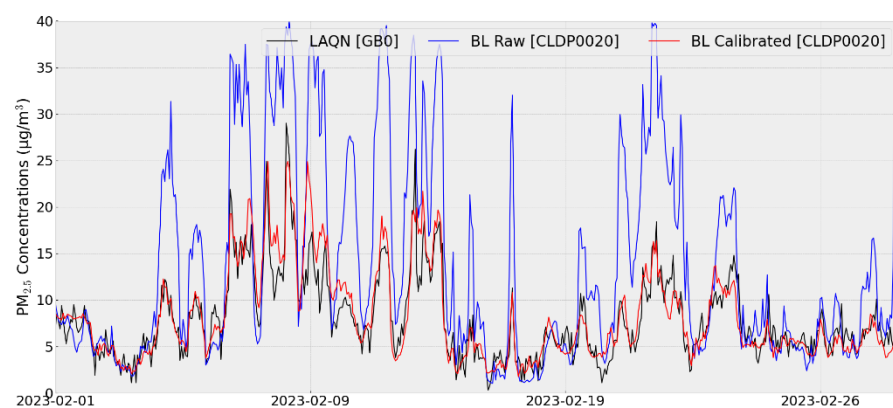
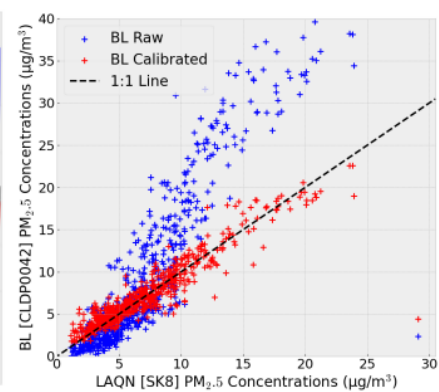
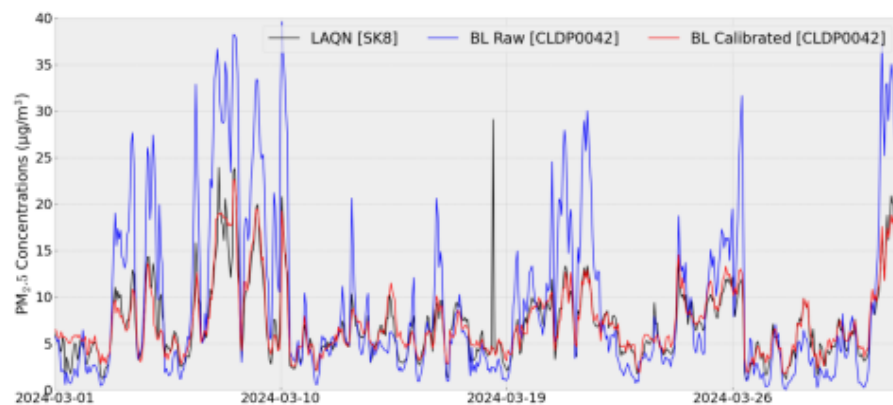


Figure C.1 Distributions of raw (left) and transformer-calibrated (right) NO₂ measurements from Breathe London, compared to London Air Quality Network, across all co-located nodes (2020–2024)

To provide a more detailed and intuitive comparison, I randomly selected one month of PM_{2.5} data for each of five co-located nodes and compared the performance of both raw and Transformer-calibrated Breathe London concentrations, as illustrated in Figure C.2.



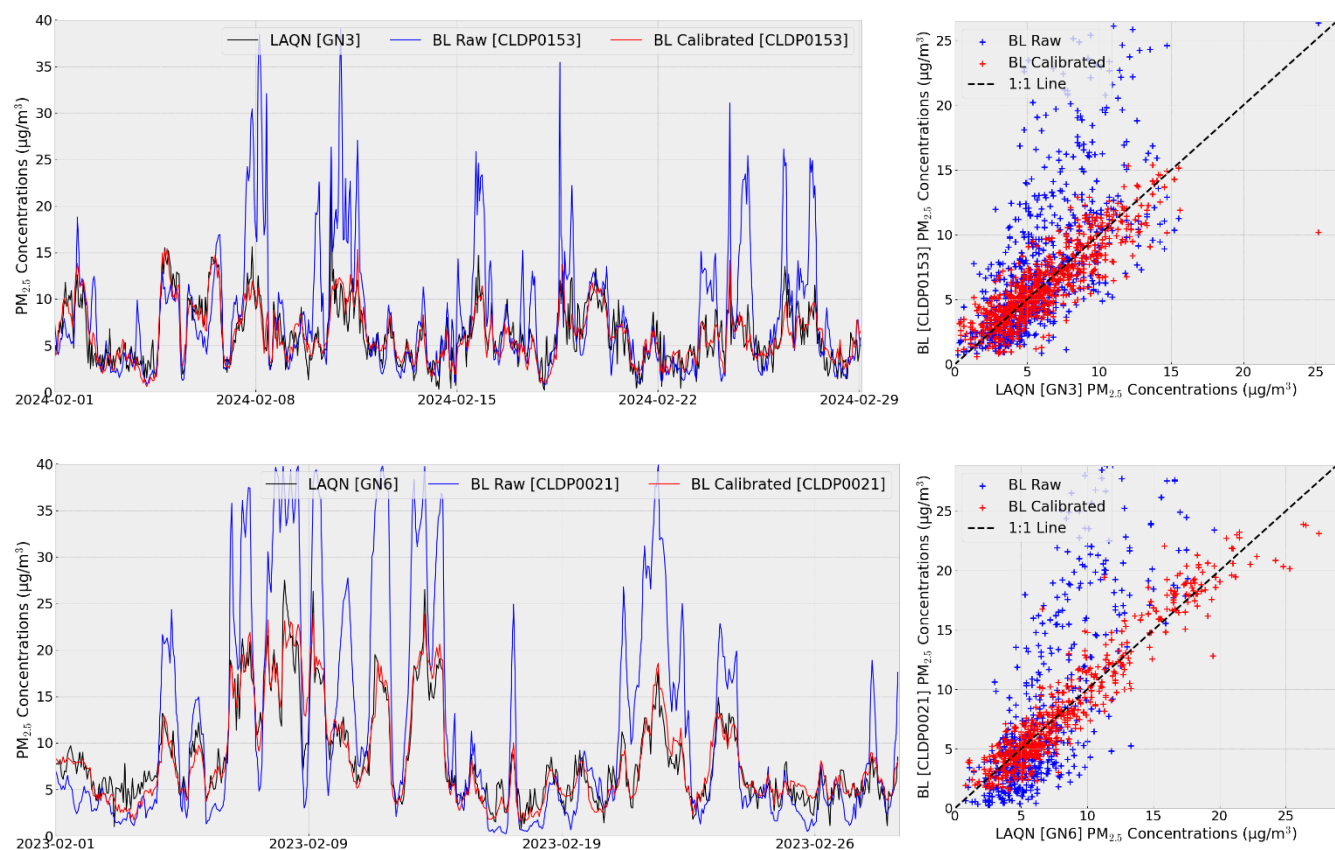


Figure C.2: line (left) and scatter plots (right) of raw and transformer-calibrated NO_2 measurements from five selected Breathe London nodes (CLDP0042, CLDP0020, CLDP0562, CLDP0153, CLDP0021), compared to London Air Quality Network stations

The performance of low-cost NO₂ calibration models

Table C.2 presents the PM_{2.5} performance metrics before and after calibration at each co-located Breathe London node.

Table C.2 Pearson correlation (r) and RMSE (µg/m³) of NO₂ before and after calibration at each co-located Breathe London node

BL Node ID	LAQN sites	BL Raw		BL Calibrated	
		RMSE	r	RMSE	r
LW2	CLDP0010	8.8594	0.7371	8.3801	0.8080
GR9	CLDP0011	9.3674	0.8111	8.4147	0.8858
BX1	CLDP0013	9.1220	0.7615	10.7024	0.8697
RB7	CLDP0014	17.6233	0.4186	10.0949	0.8417
MY1	CLDP0015	22.2279	0.3497	15.5745	0.7131
MY1	CLDP0016	24.5073	0.5903	16.9565	0.7184
TH4	CLDP0018	19.5506	0.8423	10.8171	0.8236
GR4	CLDP0019	9.5087	0.4778	13.7296	0.7608
GB6	CLDP0020	11.9723	0.6475	8.8121	0.8548
GN6	CLDP0021	10.8365	0.7398	7.8513	0.8717
SK6	CLDP0022	10.5993	0.7110	9.1373	0.8754
ST5	CLDP0023	16.3950	0.4792	11.0703	0.7415
GN0	CLDP0024	15.6774	0.5481	8.7389	0.8599
HV1	CLDP0025	12.8543	0.7024	10.0298	0.7764
WM0	CLDP0026	12.5237	0.7246	7.9565	0.8932
KC1	CLDP0027	8.8935	0.7701	10.5680	0.8438
BT4	CLDP0028	22.681	0.4914	13.9697	0.7968
HP1	CLDP0029	11.6745	0.5709	10.7041	0.8257
HP1	CLDP0030	8.9821	0.6806	10.5293	0.8367

Figure C.3 displays the distributions of raw and calibrated NO_2 concentrations from BL nodes alongside the co-located LAQN reference monitors (2020–2024).

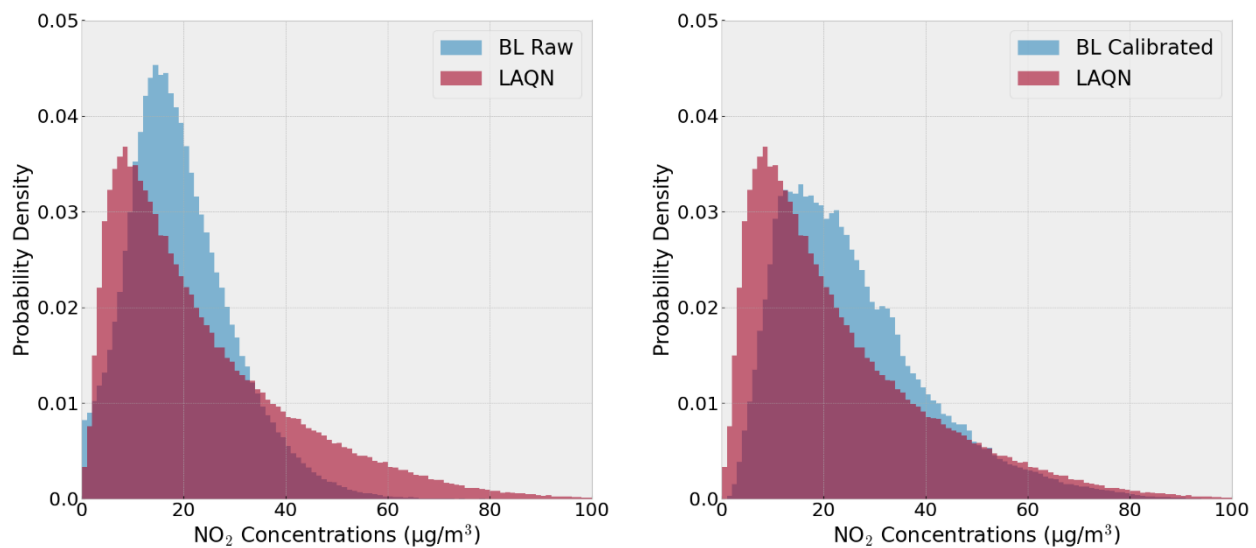
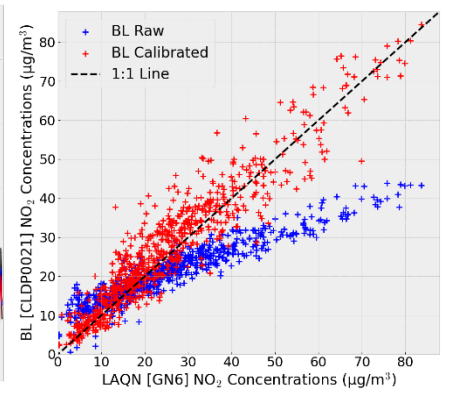
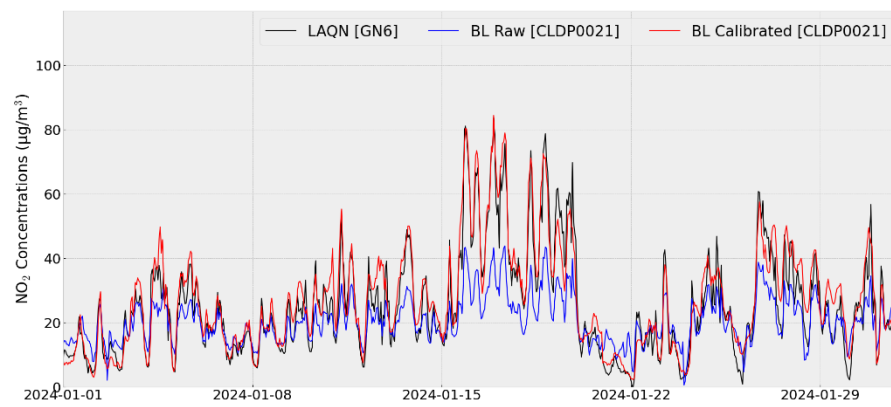
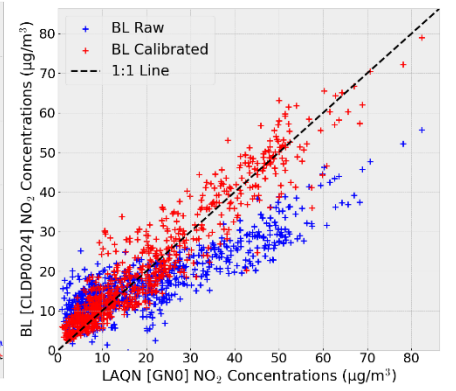
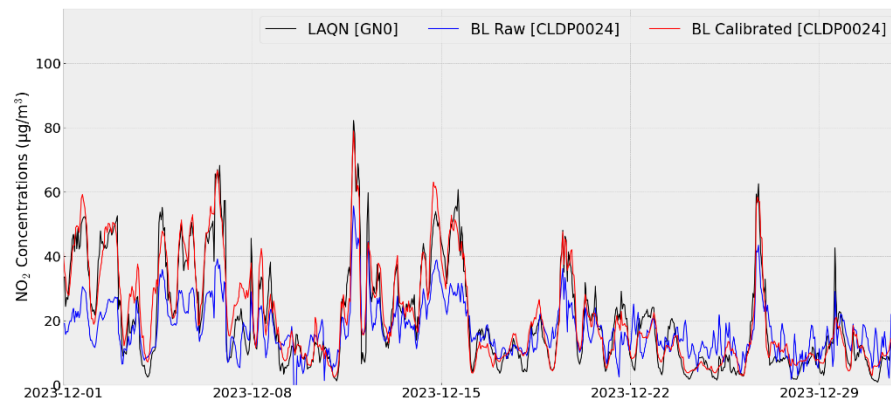
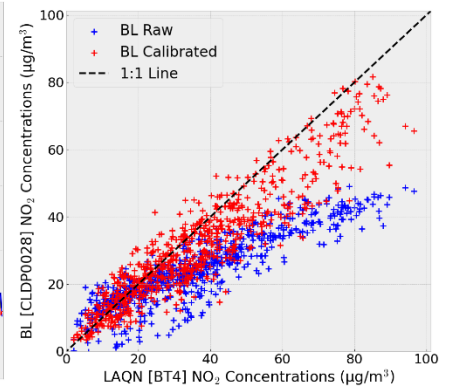
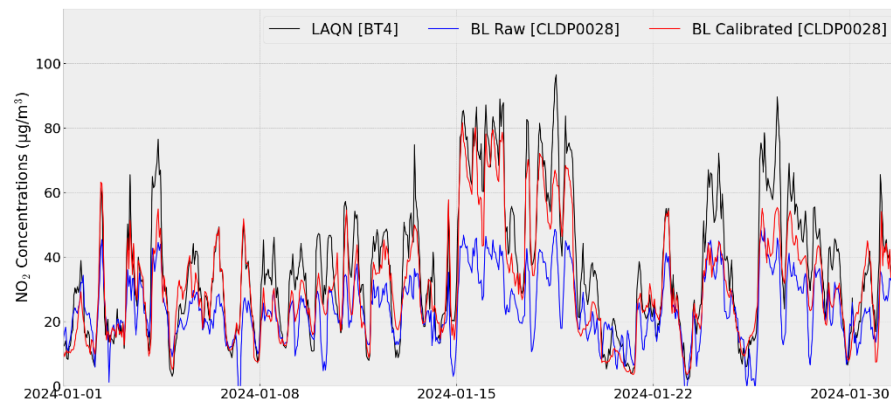


Figure C.3 Distributions of raw (left) and transformer-calibrated (right) NO_2 measurements from Breathe London compared to London Air Quality Network, across all co-located nodes (2020–2024)

To provide a more detailed and intuitive comparison, I randomly selected one month of $\text{PM}_{2.5}$ data for each of five co-located nodes and compared the performance of both raw and Transformer-calibrated Breathe London concentrations, as illustrated in Figure C.2.



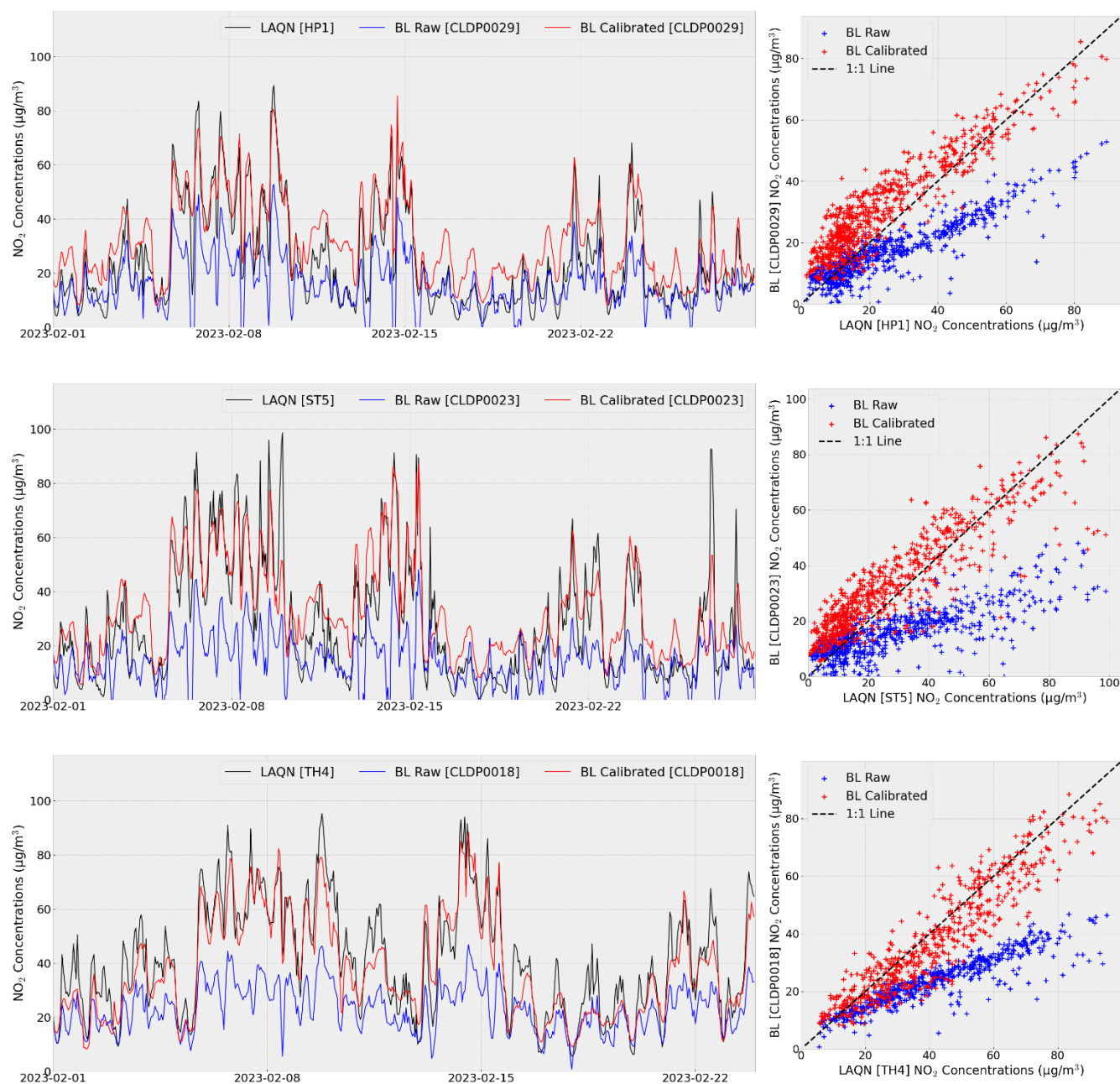
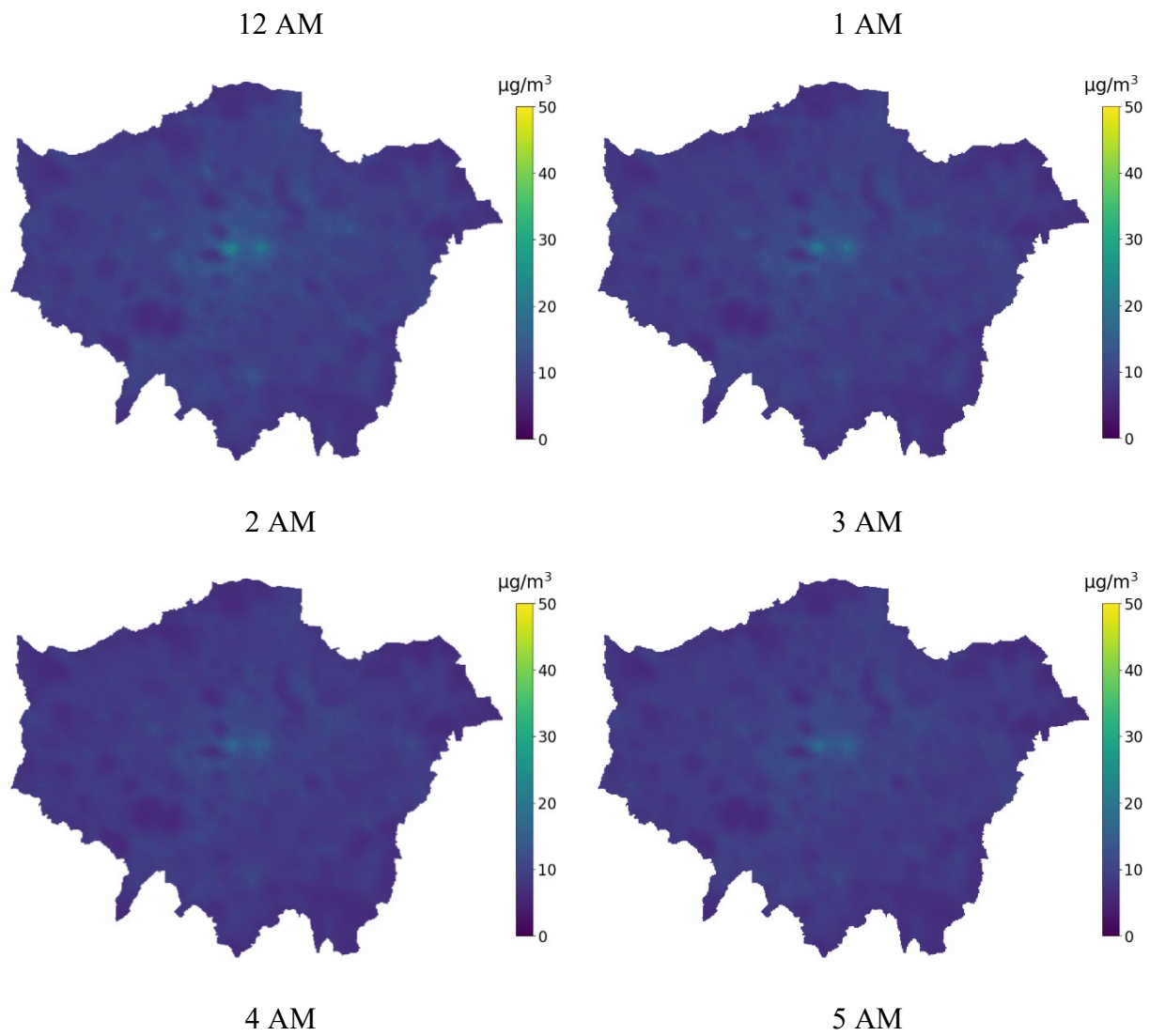


Figure C.4 Line (left) and scatter plots (right) of raw and transformer-calibrated NO₂ measurements from five Breathe London nodes (CLDP0028, CLDP0024, CLDP0021, CLDP0029, CLDP0023), compared to London Air Quality Network stations.

Appendix D. Air Quality Model

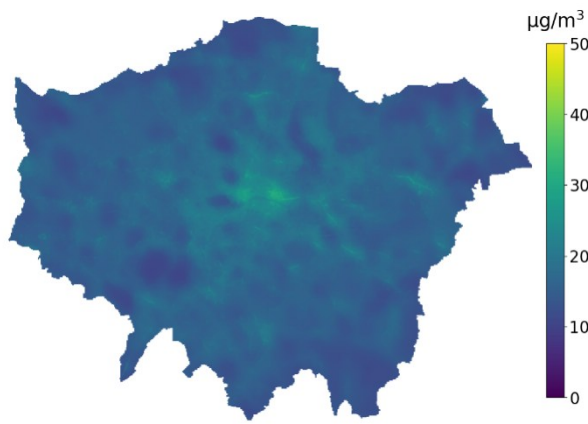




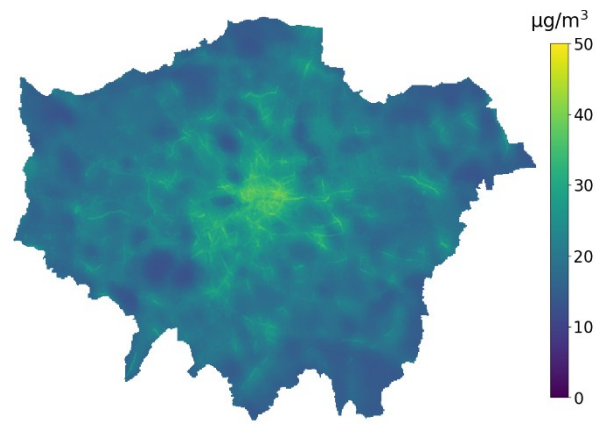
6 AM



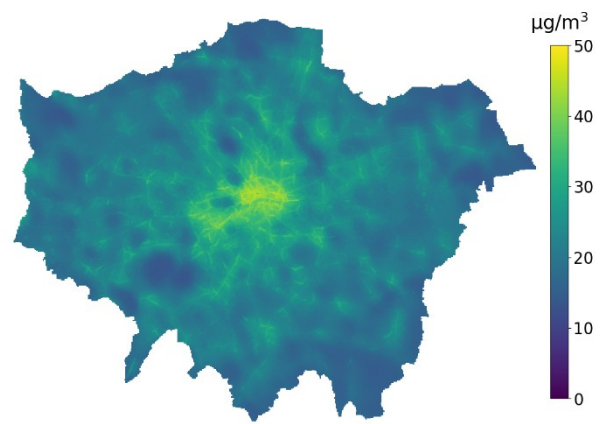
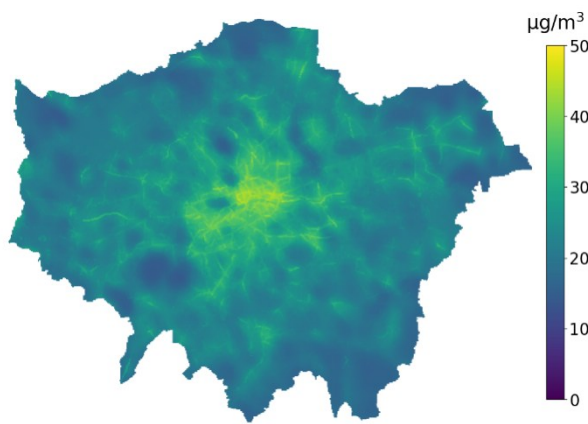
7 AM



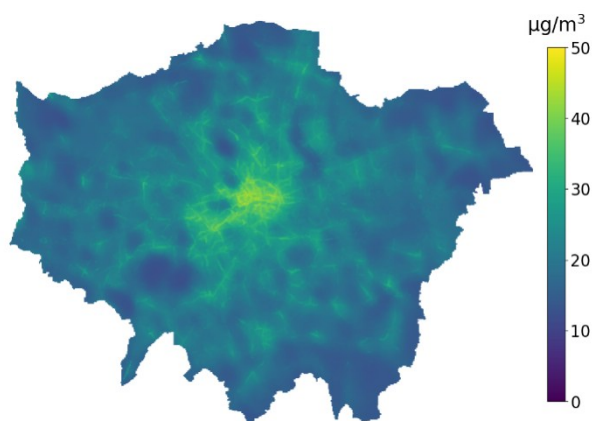
8 AM



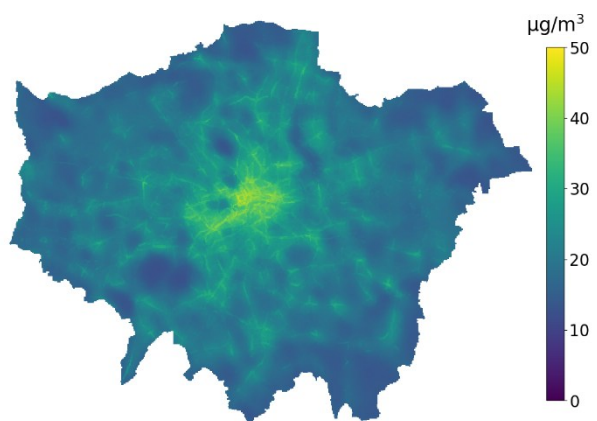
9 AM



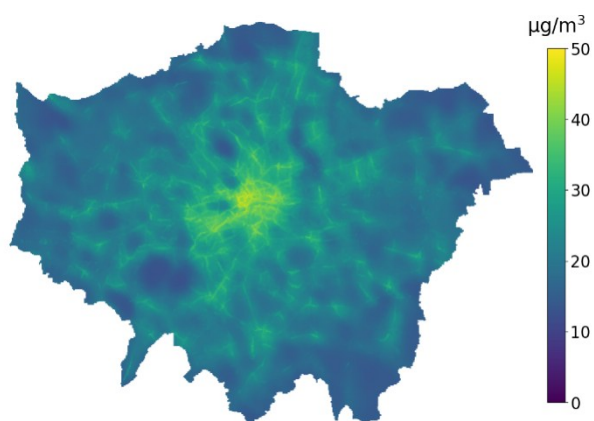
10 AM



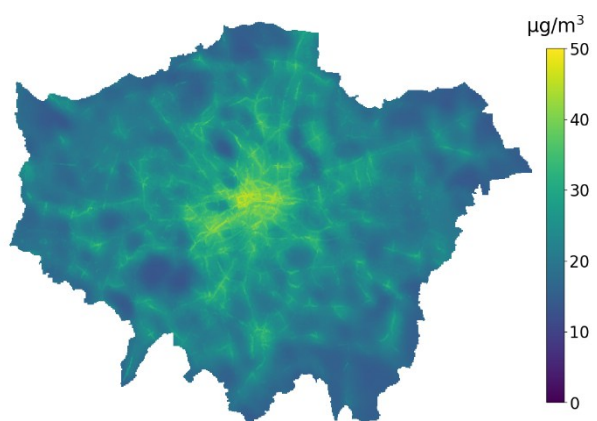
11 AM



12 PM

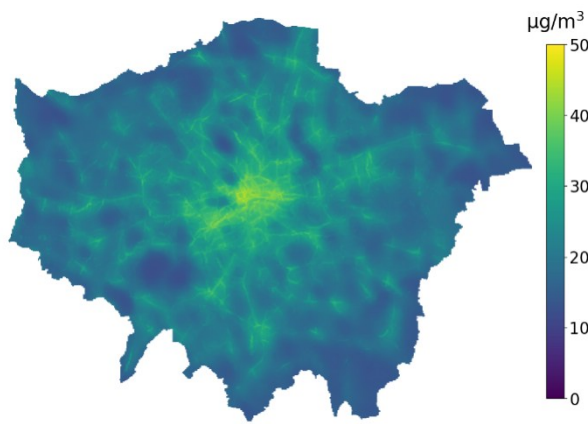


1 PM

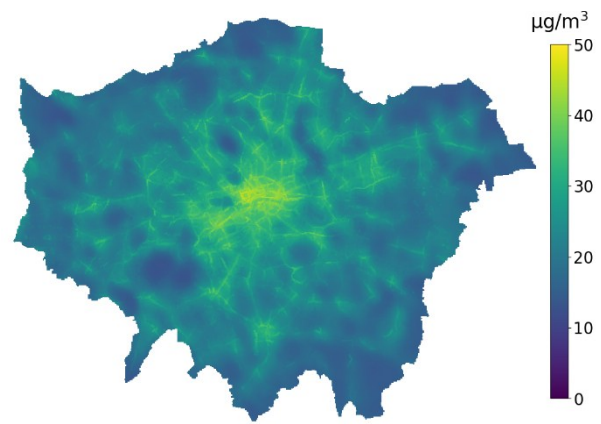


2 PM

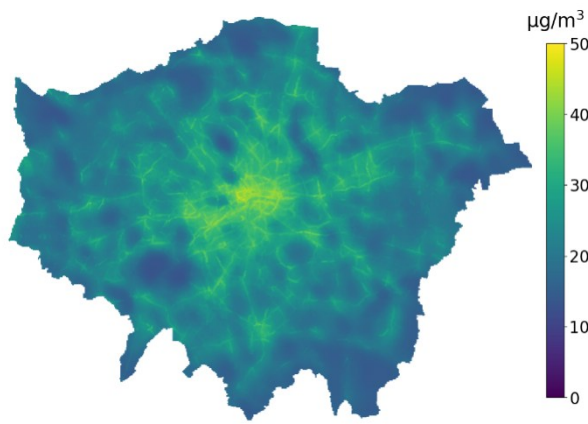
3 PM



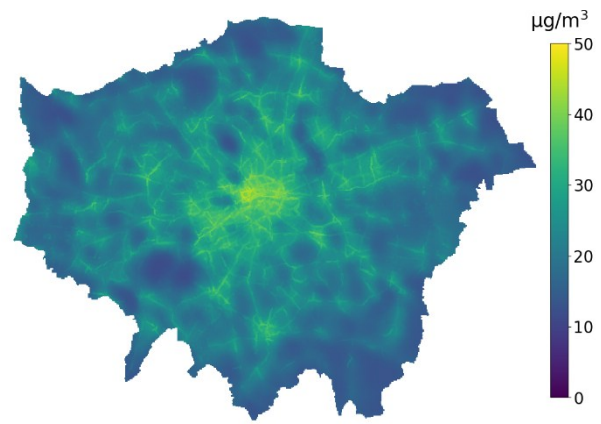
4 PM



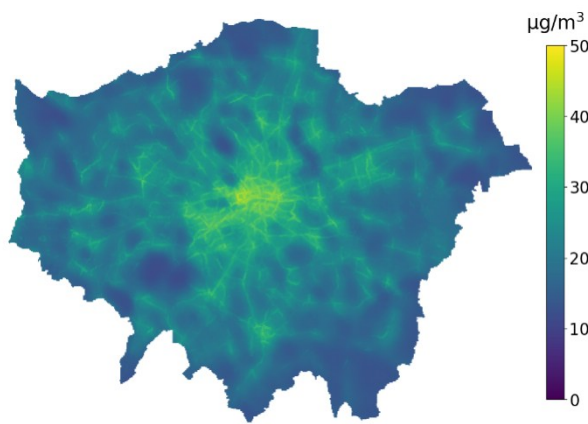
5 PM



6 PM



7 PM



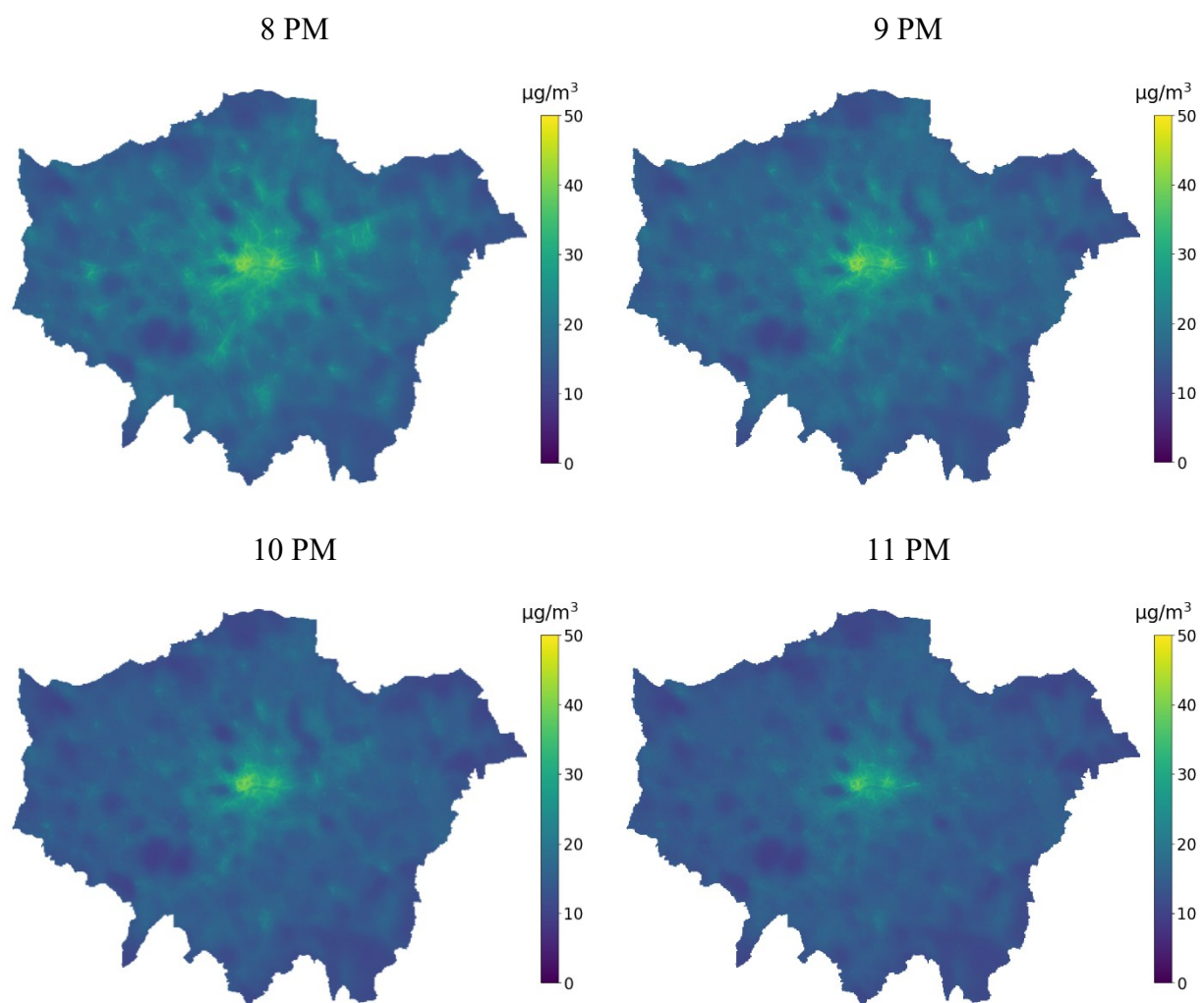


Figure D.1 The hourly estimated NO₂ map from ViT model at 12 February, 2024

Table D.1 reports the inference time of the data-driven models used in Chapter 3 for fine-scale air quality modelling ($\text{PM}_{2.5}$ and NO_2). The comparison includes baseline statistical and machine learning models (Land-Use Regression, Random Forest), convolutional neural networks (CNN), and the proposed Vision Transformer (ViT) framework, all trained on the Breathe London modelling dataset (BL_calib) under a consistent experimental setup.

Table D.1 Inference time of air quality data-driven models

Pollution	Model	Inference time(s)
NO_2	LUR	0.06
	RF	0.2
	CNN	0.06
	ViT	0.06
$\text{PM}_{2.5}$	LUR	0.06
	RF	0.2
	CNN	0.06
	ViT	0.06

Inference time corresponds to the time required to generate one modelling output (i.e., one time step prediction map over the spatial domain) after the model is trained. All experiments were conducted using the same hardware and to ensure comparability.

Appendix E. Uncertainty Analysis in Air Quality Estimation

Figure E.1~4 shows ViT-MCDO (T=100, dropout=0.1) and BME estimated NO₂ map and corresponding uncertainty at different time steps.

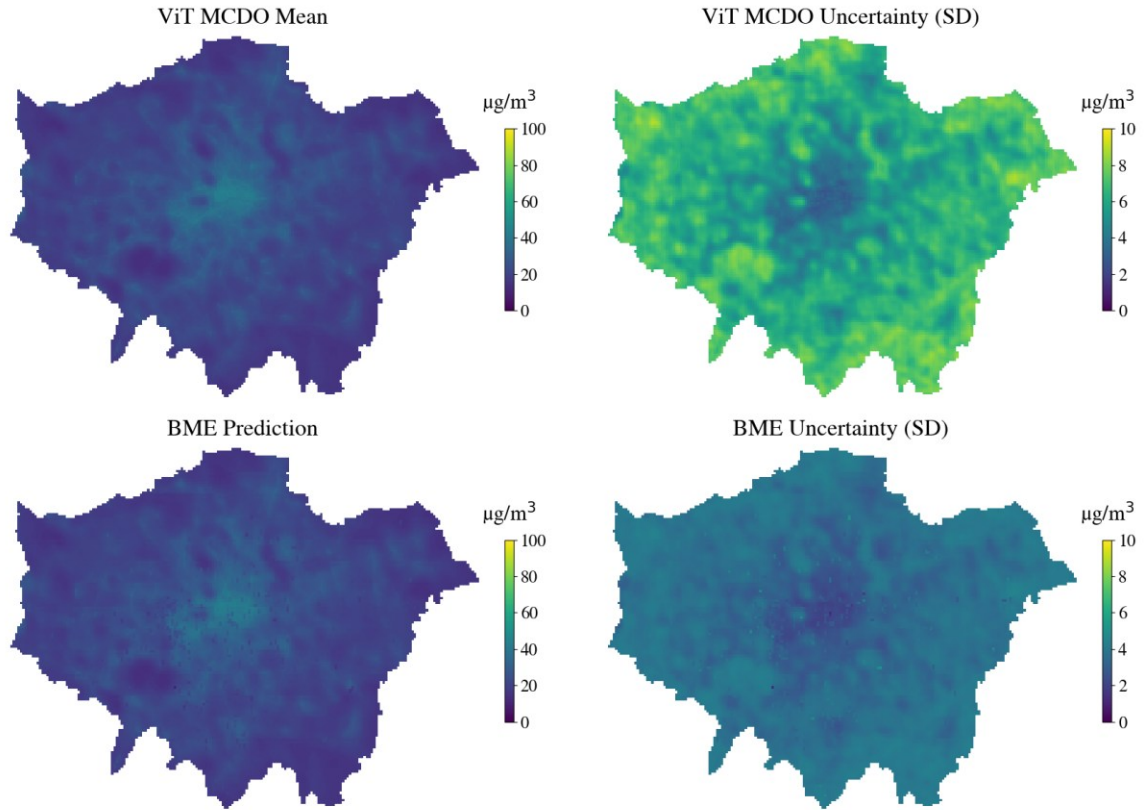


Figure E.1 NO₂ estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, T=100) VS BME and at 9 AM, February 14, 2024

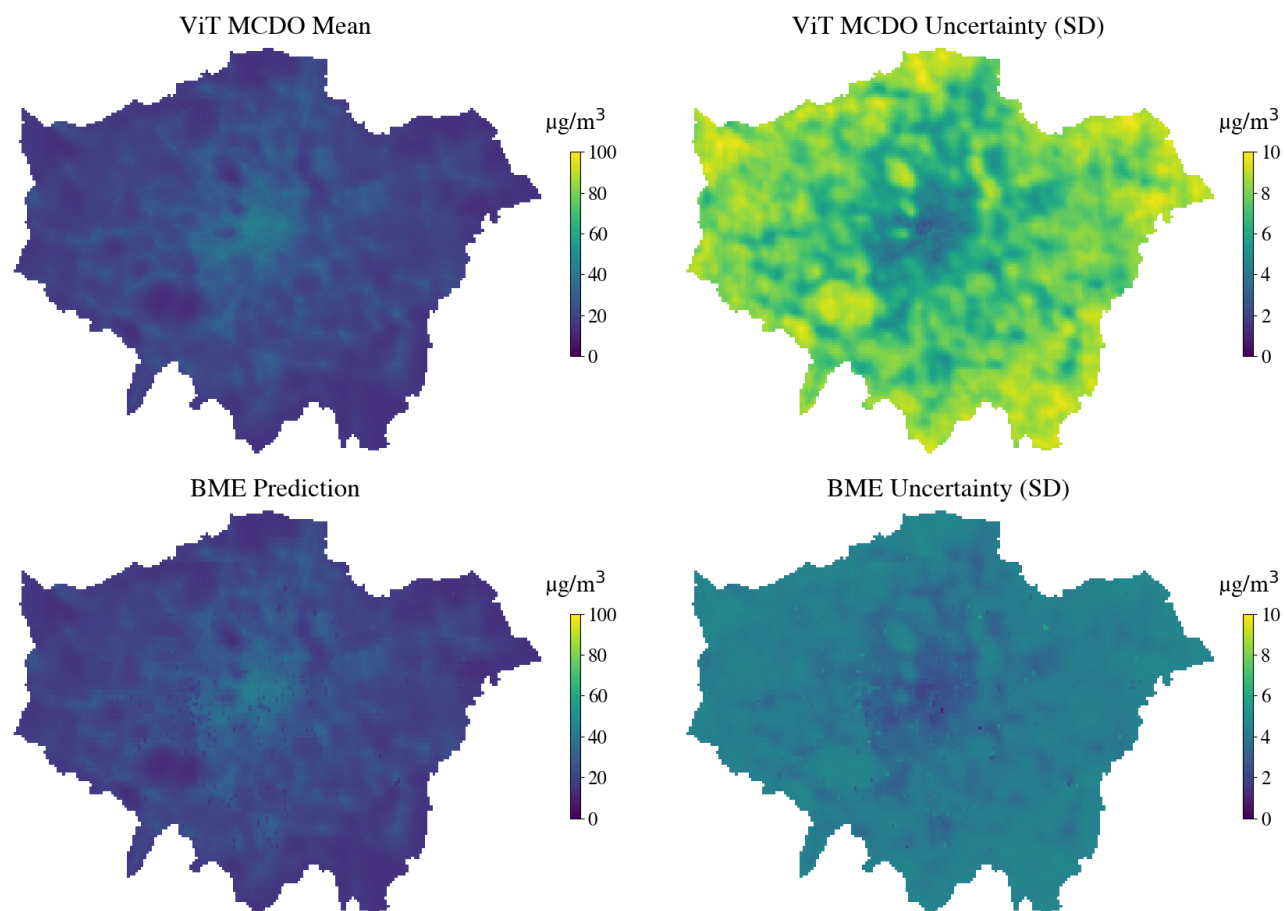


Figure E.2 NO₂ estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, T=100) VS BME and at 12 PM, February 14, 2024

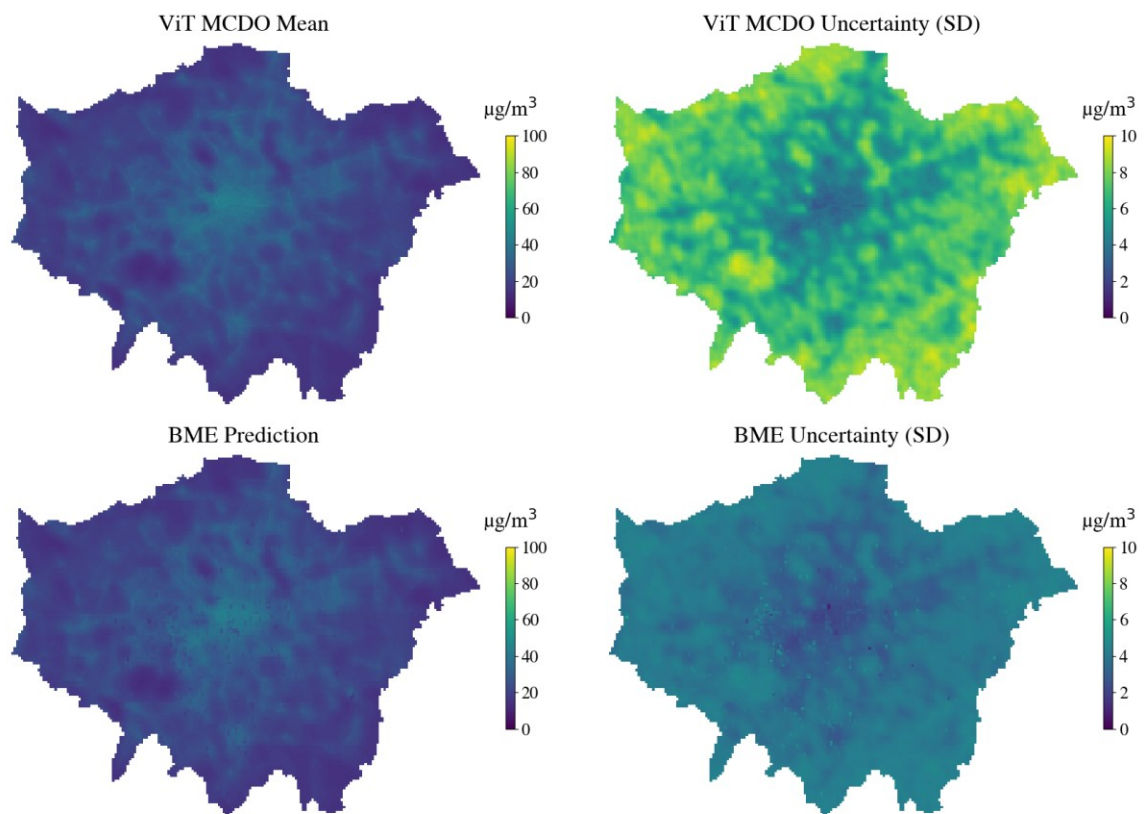


Figure E.3 NO₂ estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, T=100) VS BME and at 3 PM, February 14, 2024

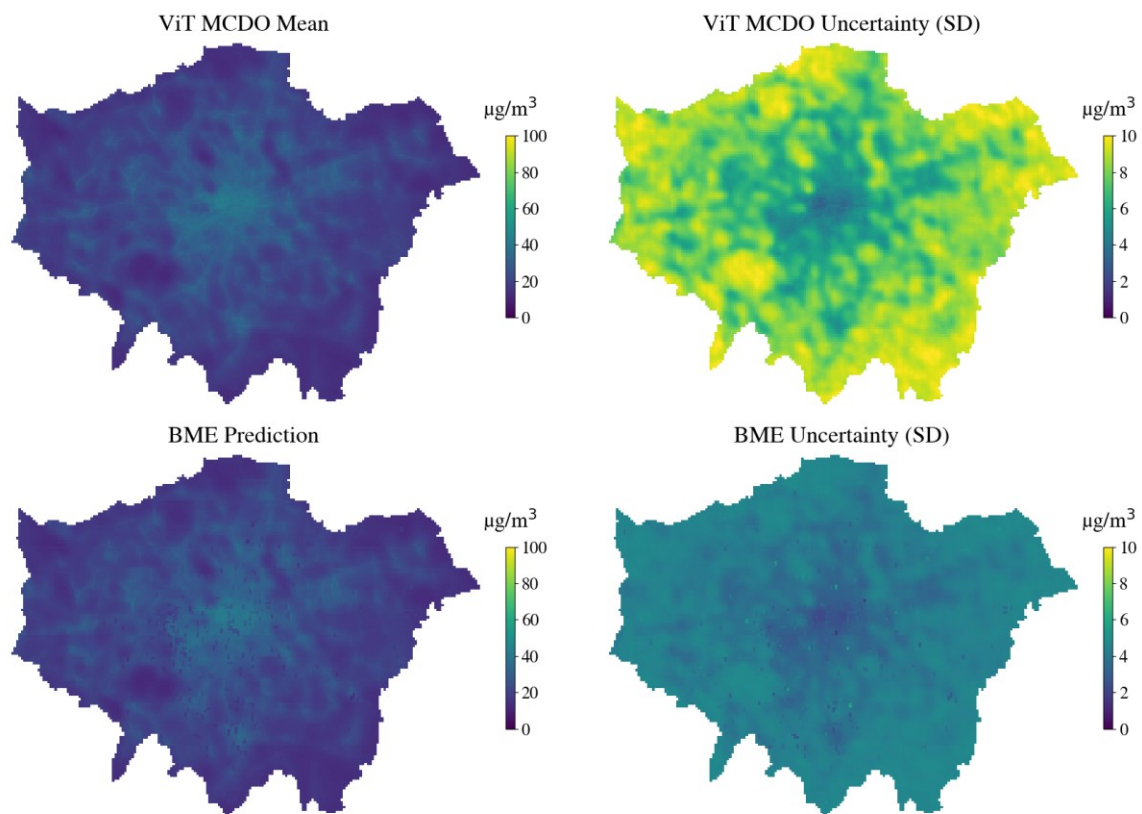


Figure E.4 NO₂ estimation and corresponding uncertainty (standard deviation) of ViT-MCDO (Dropout rate=0.1, T=100) VS BME and at 6 PM, February 14, 2024