Next-Generation Air Pollution Forecasting: Integrating AI, Spatiotemporal Dynamics, and Privacy-Ensuring Approaches for Urban Areas

Maryam Rahmani






Doctor of Philosophy in Computer Science
MADIS Doctoral School
University of Lille

Next-Generation Air Pollution Forecasting: Integrating AI, Spatiotemporal Dynamics, and Privacy-Ensuring Approaches for Urban Areas

University of Lille & CRIStAL & Inria
Spirals Team

PhD Thesis by:
Maryam Rahmani

Date: December 11, 2024
Place: Inria, Lille

Thesis Defense Jury Members:

Reporters:	Prof. *Sophie* CHABRIDON	Télécom SudParis
	Prof. *Matthieu* PUIGT	Université du Littoral
Examiners:	Prof. *Sébastien* PAYAN (President)	Sorbonne Université
	Dr. *Nadège* MARTINY	Université de Bourgogne
Supervisors:	Prof. *Romain* ROUVOY	Université de Lille
	Dr. *Suzanne* CRUMEYROLLE	Université de Lille






Doctorat en Informatique
École Doctorale MADIS
Université de Lille

Prévision de la Pollution de l'Air de Nouvelle Génération: Intégration de l'IA, des Dynamiques Spatiotemporelles et des Approches Garantissant la Confidentialité pour les Zones Urbaines

Université de Lille & CRIStAL & Inria
Équipe Spirals

Thèse de Doctorat par :
Maryam Rahmani

Date : 11 décembre 2024
Lieu : Inria, Lille

Membres du Jury de la Défense de Thèse :

Rapporteurs :	Prof. *Sophie* CHABRIDON	Télécom SudParis
	Prof. *Matthieu* PUIGT	Université du Littoral
Examinateurs :	Prof. *Sébastien* PAYAN (President)	Sorbonne Université
	Dr. *Nadège* MARTINY	Université de Bourgogne
Directeurs :	Prof. *Romain* ROUVOY	Université de Lille
	Dr. *Suzanne* CRUMEYROLLE	Université de Lille

Acknowledgments

Three years ago, I decided to embark on this academic journey, not fully knowing what lay ahead. Today, as I reflect on this path, I am filled with gratitude and a deep sense of fulfillment for all that has been accomplished. This dissertation is not only a testament to my hard work and perseverance, but also to the invaluable support of the many individuals who have stood by me throughout this journey.

First and foremost, I would like to express my deepest appreciation to my supervisors, Romain Rouvoy and Suzanne Crumeyrolle. Your guidance, expertise, and encouragement have been instrumental in shaping this research. You provided not only academic mentorship but also personal support through the highs and lows of this journey. Thank you for believing in me, even when I faced moments of self-doubt and unpredictable challenges along the way. I also extend my gratitude to Nadège for generously sharing her valuable data and guiding me throughout my PhD project, and to Amir for his collaboration, which allowed me to undertake a one-month internship in Norway.

A special thanks to Lionel Seinturier for being an exemplary team leader; your work ethic is something I admire and aspire to. I also want to extend my gratitude to the Spirals team members, who contributed to many thought-provoking discussions, especially Daniel and Rémy, whose company made the work more enjoyable. To Sihem, Belkis, Imane, Alexandre, Niloofar, and the rest of the team—thank you for the great years together, your help, and all the engaging conversations we've shared. I also appreciate Inria for fostering an environment that provided the resources and support we needed to succeed. I would like to sincerely thank the Fondation I-SITE Université de Lille Nord-Europe (ULNE) for their generous support. Their resources and commitment were crucial to the development and completion of this work, and I am deeply grateful for the opportunity provided to advance my research in such a meaningful way.

To my family, words cannot express the depth of my gratitude. Your unwavering support, love, and encouragement have been my greatest source of strength throughout this journey. Your sacrifices, patience, and endless love have been the foundation of my success. From the bottom of my heart, thank you for always being my biggest cheerleaders—I couldn't have done this without you. Beyond the academic journey, I want to thank my dear friends who made this experience easier. Your mental and spiritual support helped me stay strong and focused. A special thank you to Athulya, Hedieh, Nasrin, and Shakila—you stood by me through every challenge, and I couldn't have done this without you. Your friendship means the world to me.

Last but certainly not least, to my beloved partner Hamid—thank you for your patience, and understanding, and for being my constant source of comfort. Your love has provided me with peace and security, a much-needed refuge during the toughest times.

Finally, to everyone who has helped me along the way—whether through advice, collaboration, or offering a kind word when I needed it most—thank you. Your kindness made this journey so much easier.

Maryam Rahmani
Lille, October 2024

Dedication

This dissertation is dedicated to the loving memory of my mother, whose spirit continues to inspire and guide me each day. Her love and support have made this achievement possible.

Abstract

Air quality is a critical global issue, with air pollution posing serious environmental and public health risks, especially in urban areas where fine *Particulate Matter*s (PMs) especially ($PM_{2.5}$) is among the most harmful pollutants. Despite significant advancements in air quality monitoring and modeling, challenges such as data variability, computational demands, scalability, resolution constraints, and privacy concerns continue to limit the accuracy and effectiveness of current forecasting systems. This dissertation presents a novel approach to air pollution prediction in urban areas by integrating *Artificial Intelligence* (AI), spatiotemporal modeling, and privacy-preserving data collection techniques.

The first major contribution of this research is the development of PMFORECAST, a temporal prediction model specifically designed to forecast $PM_{2.5}$ levels. By utilizing advanced machine learning techniques and temporal attention mechanisms, PMFORECAST effectively captures temporal dependencies in pollutant concentrations, leading to highly accurate predictions for both short-term and long-term forecasting. Additionally, the model demonstrates significant multi-tasking capabilities. It achieves a notable prediction accuracy of 99.7% for 1-hour forecasts and 73.5% for 12-hour forecasts, representing substantial improvements over existing models in terms of precision and computational efficiency.

Spatial and temporal data from underground sensor networks to predict $PM_{2.5}$ concentrations across different geographic regions. The *Graph Temporal LSTM* (GT-LSTM) model employs *Graph Convolutional Network*s (GCNs) to capture the complex interactions between pollution sources and atmospheric conditions at ground level, while utilizing *Long Short-Term Memory*s (LSTMs), as described in the previous contribution, to model temporal dependencies. This approach provides a more refined understanding of pollutant dispersion over time and space. By operating with fixed zone resolutions corresponding to available data resources, the model ensures accurate and localized predictions.

The third contribution is the design of a federated learning architecture called FEDAIRNET, aimed at enhancing air quality prediction using mobile sensor data while safeguarding user privacy. Traditional air quality monitoring stations are often constrained by limited spatial coverage and high costs, whereas mobile sensors offer a more flexible and granular data source. However, the collection of mobile sensor data introduces privacy concerns. FEDAIRNET addresses these challenges by distributing the learning process across multiple devices, ensuring that sensitive data remains on local devices while still contributing to global model updates. This decentralized approach not only improves prediction accuracy but also mitigates risks associated with centralized data collection, such as *point-of-interest* (PoI) attacks.

The models presented in this thesis have been rigorously tested in real-world environments, demonstrating their potential to transform air pollution monitoring systems. The PMFORECAST model provides robust predictions of $PM_{2.5}$ concentrations, making it valuable for public health interventions and environmental policies. The *Spatiotemporal Model* adds a critical layer of understanding by analyzing how pollutants behave across spatial and temporal dimensions, while the FEDAIRNET architecture ensures that privacy is protected as the use of mobile sensors becomes more prevalent.

This research represents significant advancements in air pollution prediction by integrating AI-driven insights with privacy-preserving data collection techniques. Future work should focus on incorporating additional data sources, and refining hybrid models that combine temporal, spatial, and mobile sensing data. These innovations will contribute to more accurate, timely, and secure air pollution forecasting systems, ultimately helping to mitigate the harmful effects of air pollution on human health and the environment.

Keywords: Air Quality, $PM_{2.5}$ Forecasting, Spatiotemporal Analysis, Federated Learning, Data Privacy-Preserving, high-resolution

Résumé

La qualité de l'air est un problème mondial, la pollution de l'air posant de sérieux risques environnementaux et pour la santé publique, surtout dans les zones urbaines où les particules fines (PM), notamment le *Particulate Matter* $(PM)_{2.5}$, sont parmi les polluants les plus nocifs. Malgré des avancées dans la surveillance de la qualité de l'air, des défis comme la variabilité des données, les exigences computationnelles, la scalabilité, et les préoccupations en matière de confidentialité limitent l'efficacité des systèmes de prévision actuels.

Cette thèse présente une approche novatrice pour prédire la pollution de l'air dans les zones urbaines en intégrant IA, modélisation spatiotemporelle, et techniques de collecte de données préservant la vie privée. La première contribution majeure est le développement de PMFORECAST, un modèle de prédiction temporelle conçu pour prévoir les niveaux de $PM_{2.5}$. En utilisant des techniques avancées d'apprentissage automatique et des mécanismes d'attention temporelle, PMFORECAST capture efficacement les dépendances temporelles des polluants, conduisant à des prévisions précises pour le court et le long terme. De plus, le modèle démontre des capacités multitâches, atteignant une précision de 99.7% pour les prévisions à 1 heure et de 73.5% pour celles à 12 heures, représentant des améliorations par rapport aux modèles existants.

Le modèle spatiotemporel intègre des données de réseaux de capteurs souterrains pour prédire les concentrations de $PM_{2.5}$ à travers différentes régions. Le modèle *Graph Temporal LSTM* (GT-LSTM) utilise des réseaux de GCN pour capturer les interactions entre les sources de pollution et les conditions atmosphériques, tout en utilisant des LSTMs pour modéliser les dépendances temporelles. Cette approche permet une compréhension approfondie de la dispersion des polluants dans le temps et l'espace. En utilisant des résolutions fixes correspondant aux ressources de données, le modèle assure des prévisions précises et localisées.

La troisième contribution est la conception d'une architecture d'apprentissage fédéré appelée FEDAIRNET, visant à améliorer la prédiction de la qualité de l'air avec des données de capteurs mobiles tout en préservant la vie privée. Les stations de surveillance traditionnelles, limitées par une couverture spatiale et des coûts élevés, voient en revanche les capteurs mobiles une source de données flexible. Cependant, la collecte de données de capteurs mobiles soulève des préoccupations de confidentialité. FEDAIRNET distribue le processus d'apprentissage sur plusieurs appareils, garantissant que les données sensibles restent locales tout en contribuant aux mises à jour du modèle global. Cette approche décentralisée améliore non seulement la précision des prévisions, mais atténue aussi les risques liés à la collecte centralisée.

Les modèles présentés dans cette thèse ont été testés dans des environnements réels, montrant leur potentiel à transformer les systèmes de surveillance de la pollution. Le modèle PMFORECAST fournit des prévisions robustes de $PM_{2.5}$, essentielles pour les interventions en santé publique. Le modèle spatiotemporel enrichit notre compréhension en analysant le comportement des polluants, tandis que l'architecture FEDAIRNET protège la vie privée à mesure que l'utilisation de capteurs mobiles se généralise. Cette recherche représente une avancée significative dans la prédiction de la pollution de l'air en intégrant des perspectives basées sur l'IA avec des techniques de collecte de données préservant la vie privée. Les travaux futurs devraient se concentrer sur l'incorporation de sources de données supplémentaires et l'affinement de modèles hybrides combinant données temporelles, spatiales et de détection

mobile, contribuant à des systèmes de prévision plus précis, opportuns et sécurisés, afin de réduire les effets nocifs de la pollution sur la santé humaine et l'environnement.

Mots-clés : Qualité de l'Air, Prévision des $PM_{2.5}$, Analyse Spatiotemporelle, Apprentissage

Fédéré, Préservation de la Confidentialité des Données, haute résolution

Table of Contents

List of figures

List of tables

ACRONYMS

R^2 *Coefficient of determination*. 31–33, 38–40, 47, 50, 57, 62, 63, 90, 94, 95, XIII, XIV, XVI

AI *Artificial Intelligence*. 3, 4, 13, 63, IV

ANN *Artificial Neural Networks*. 15, 16, 65

AQI *air quality index*. 27, 64, 65

CNN *Convolutional Neural Network*. 15, 17, 27, 52

DNN *Deep Neural Networks*. 65

FL *Federated Learning*. 8, 18–24, 64, 65, 84, 86, 88, 93, 99, XI

GCN *Graph Convolutional Network*. 17, 50, 52–54, 58, 61, 98, IV, XII, XVI

GT-LSTM *Graph Temporal LSTM*. 7, 51, 52, 54, 55, 58, 61–63, 98–101, IV, VI, XII, XIII, XVI

LSTM *Long Short-Term Memory*. 6, 15–17, 26–30, 32, 40, 42, 44, 47, 50, 53, 54, 58, 62, 65, 70, 73, 95, 97, 98, IV, XI, XII, XVI

MAE *Mean Absolute Error*. 38, 39, 47, 50, 57, 61, 63, 84–90, 93–95, 100, VIII, X, XI, XIII–XV

ML *Machine Learning*. 13, 15, 30, 35, 37, 40, 70

MSE *Mean Square Error*. 32, 38, 39, 62, 90

PM *Particulate Matter*. 1, 3, 5–8, 10, 12, 13, 15–18, 29–31, 34, 35, 37–44, 46, 47, 49–52, 56, 58, 60–64, 70, 76, 97, 98, 100, IV–VII, XI–XIII

PoI *point-of-interest* . 66, 76, 80–84, 93, 100, IV, XIII

RMSE *Root Mean Square Error*. 31–33, 38–40, 47, 50, 57, 59, 61, 63, 84–87, 90, 93–95, 100, 101, VIII, X, XI, XIII–XV

RNN *Recurrent Neural Networks*. 15–17, 26, 30, 49, 50

WMAPE *Weighted Mean of Absolute Percentage Error*. 38, 39, 57, 62, 94, 95, XIII, XIV

Chapter 1

INTRODUCTION

1.1 Air Pollution: A Silent Threat

The air we breathe is a critical yet often overlooked component of our environment, while essential for sustaining life on Earth. However, this invisible layer surrounding our planet is increasingly compromised by air pollution [Org21].

Scientifically, air pollution refers to excessive or harmful substances in the air that adversely affect human health and the environment [Bri23]. These pollutants include gases, like ozone(O_3) and nitrogen dioxide (NO_2), which can be considered 'invisible invaders' pose significant threats [MM04]. Another significant contributor is PM, consisting of microscopic solid or liquid particles suspended in the air [Age20]. PM, classified by size, The finer fraction of PMs ($PM_{2.5}$ and especially PM_1), can penetrate deep into the lungs, posing serious health risks [RSB$^+$11]. Moreover, $PM_{2.5}$, observed in urban areas [GHW24], is enriched in hazardous metals and organic compounds [ZWL20], potentially inducing additional oxidative stress [TSC$^+$10].

Pollutants, particularly (PMs), originate from both natural and anthropogenic sources [KBD$^+$15]. Natural pollutants include emissions from wildfires, volcanic eruptions, dust storms, and sea spray, which contribute to atmospheric particulate matter and can significantly affect air quality, especially during environmental events, such as wildfires [LPG$^+$16]. In contrast, anthropogenic pollutants primarily arise from human activities, such as vehicle emissions, industrial processes, agricultural practices, and residential heating [USS$^+$24]. These emissions are often concentrated in urban areas, leading to elevated pollution levels compared to rural settings. Burning fossil fuels for energy and transportation significantly contributes to fine particulate matter, posing serious risks to public health and urban ecosystems [HLL$^+$21].

While natural and anthropogenic sources impact air quality, anthropogenic emissions are typically more manageable through regulations and technological progress. The interplay between these sources complicates the air quality landscape, as natural events can exacerbate

the effects of human-made pollution. Concentrations of anthropogenic pollutants in urban areas not only threaten immediate health by increasing respiratory and cardiovascular issues, but also jeopardize the long-term sustainability of urban living environments [LZB+24, ZXL+23].

The Industrial Revolution marked a dramatic shift, with industrialization and coal-fired power plants significantly worsening air pollution in cities. This led to the first documented public health concerns [DSSY24, OA]. The devastating Great Smog of London in 1952 [TVC23] served as a wake-up call, highlighting the urgent need for stricter regulations. Thousands perished due to the lethal combination of fog and pollution [Mar, Kel17].

Efforts to combat air pollution began with understanding its presence and extent. Early methods in the late 19th and early 20th centuries were relatively simple. For instance, Ringelmann charts offered a basic assessment of smoke density using a visual scale [QLY+23], while rudimentary deposit gauges collected soot and dust particles to estimate air quality [Val14].

Technological progress in the mid-20th century introduced more sophisticated air pollution measurement techniques. The 1940s saw the development of the first reliable ozone monitors, crucial for quantifying smog levels [BL79]. These early devices employed ultraviolet photometry, a technique that became cornerstone to accurately measure ozone concentrations.

The need for continuous air quality data led to the establishment of automated monitoring stations in the 1950s, providing real-time data crucial for air quality management [BS06]. Modern air pollution monitoring stations are equipped with advanced instruments capable of quantifying a wide spectrum of pollutants.

Air quality monitoring has since evolved beyond ground-based stations. Satellite observations now offer a broader view, enabling the tracking of air pollution plumes across vast distances [Boa]. Additionally, sophisticated atmospheric modeling tools help predict air quality trends and assess the impact of emission control strategies [Dun14].

Technological advancements have also played a critical role in improving air quality management [Zha16]. Air quality monitoring networks enable real-time tracking of pollutants, which informs policy decisions and public awareness campaigns [Pet14]. Public awareness campaigns have become vital in mobilizing support for stricter control measures, particularly as scientific evidence on the health impacts of air pollution has grown [RRBPZ19].

Invisible yet vital, the air we breathe can become a silent threat when harmful substances pollute the atmosphere [KKB+15]. Pollutants wreak havoc on our health, triggering respiratory illnesses, such as asthma, *chronic obstructive pulmonary disease* (COPD), and even lung cancer [LEF+15]. The World Health Organization estimates that over seven million premature deaths annually worldwide are attributed to air pollution [Org23]. Children, pregnant women,

and the elderly are especially vulnerable to these health dangers [LAK12]. Additionally, air pollution adversely affects both terrestrial and aquatic ecosystems, leading to environmental degradation and loss of biodiversity. This highlights the importance of addressing air pollution to mitigate environmental issues effectively [MSSB20].

Monitoring air quality involves measuring the concentration of various pollutants in the atmosphere [CZZ+21]. Government agencies and environmental organizations deploy air quality monitoring stations equipped with instruments that sense and quantify specific pollutants. These stations measure parameters such as PMs, ozone, and nitrogen dioxide [MWC17]. The data collected are used to calculate an *Air Quality Index* (AQI), which provides a standardized score reflecting the health risks associated with current air quality levels [FBB+20].

Today, a sophisticated network of air quality monitoring stations stands guard globally. These stations, equipped with cutting-edge technology, are far more advanced than the rudimentary methods of the past [Tur70]. The evolution of air pollution modeling has seen a shift from simple dispersion models to sophisticated computational models [ANF+17]. Early models provided basic estimates of pollutant concentrations based on emission sources and meteorological conditions but were limited in their ability to account for complex atmospheric processes [XZL21].

Advancements in computing power and atmospheric science have led to the development of *chemical transport models* (CTMs), which simulate the transport and transformation of pollutants in the atmosphere [Sto97]. These models incorporate detailed chemical reactions and physical processes, offering more accurate predictions of air quality [CCHC21].

Air pollution remains a critical global challenge with profound implications for public health and the environment [Hay21]. While significant progress has been made in understanding and addressing this issue, continuous innovation is essential. The integration of *Artificial Intelligence* (AI) has revolutionized air quality monitoring and prediction. By processing vast datasets from diverse sources, AI algorithms can identify intricate patterns and trends, leading to more accurate air quality models and timely interventions during pollution spikes [YWC+24, USSJ23]. These advances offer the potential to reduce computational demands and improve prediction efficiency.

To further enhance air quality management, this research investigates the integration of diverse stationary data collection methods, including cost-effective, stable stations and citizen-generated data from portable sensors. While this approach offers significant potential, it also introduces critical privacy and data security concerns [Che24]. Leveraging emerging technologies, like *Federated Learning* (FL), can address these challenges while maximizing the value of citizen participation [YDL+24a]. By combining multiscale spatial-temporal data processing with AI-driven insights, this study aims to develop innovative solutions for the prediction and monitoring of localized air pollutants, such as $PM_{2.5}$, in urban areas.

1.2 Research Aims and Motivations: Tackling Air Pollution with AI-Driven and Mobile Sensing Insights

Predicting air pollution is vital to safeguarding public health and the environment. Despite advances in monitoring technology and modeling approaches, significant challenges persist that hinder accurate forecasting and effective mitigation, highlighting the need for innovative research and solutions.

- **Complexity of Atmospheric Processes:** Predicting air pollution is challenging due to the complex interplay of atmospheric processes. Factors like meteorological conditions, chemical reactions, and topography significantly impact the dispersion and transformation of pollutants. Traditional models often rely on simplifications that can result in discrepancies between predicted and actual pollution levels, thereby hindering effective mitigation strategies. Our research focuses on identifying the most suitable AI algorithms to accurately analyze the specific characteristics of air pollutant data.

- **Computational Demands of Advanced Models:** Advanced air pollution models are increasingly complex and demand substantial computational resources. Balancing model sophistication with available computing power is essential for practical application. This project aims to develop efficient computational techniques and validation methods to enhance model performance and feasibility, ensuring more reliable forecasts for effective air quality management.

- **Variability and Sparsity of Data:** Air quality data often face significant challenges due to variability and sparsity. Despite the global expansion of monitoring networks, gaps in both spatial and temporal coverage persist, particularly in urban areas where limited infrastructure may rely on a single monitoring station to represent an entire region. Furthermore, constraints in sampling frequency and sensor accuracy can reduce the reliability of forecasts and hinder the effectiveness of targeted interventions.

- **Integration of Artificial Intelligence:** AI presents promising solutions by analyzing large and complex datasets to uncover patterns that traditional methods may overlook. Machine learning models can improve prediction accuracy by learning from historical data and identifying nonlinear relationships. The success of AI-driven forecasting hinges on the quality of training data and the robustness of the algorithms employed.

- **Technological and Infrastructure Limitations:** While monitoring technology has advanced, limitations in sensor accuracy and infrastructure can affect the overall effectiveness of air quality monitoring. Traditional air quality stations are costly and limited in spatial coverage. Mobile sensors offer a flexible, cost-effective alternative, providing

enhanced granularity and coverage. Our project explores the integration of mobile sensor data with existing networks to improve monitoring systems.

- **Preserving Users' Privacy:** Mobile sensor data collection raises privacy concerns, making it essential to ensure data confidentiality and integrity. Technologies like federated learning, which process data locally on devices, offer a way to address these concerns while still enabling valuable insights.

- **Real-Time Prediction and Updating with Streaming Data:** Rapid processing and analysis of large volumes of data are crucial for issuing timely warnings and implementing effective mitigation strategies. Our approach emphasizes the development of resource-optimized models capable of continuously learning and adapting based on incoming data, enabling accurate short-term and long-term pollution forecasting.

Addressing these multifaceted challenges requires innovative approaches to air quality management. By integrating mobile sensors with an existing inexpensive monitoring infrastructure and leveraging advanced data analysis techniques, this research aims to contribute to the development of novel and effective air pollution forecasting models. The insights gained from this study will enhance our understanding of the complex dynamics of air pollution, ultimately leading to improved policy decisions to reduce pollutant emissions and therefore improved public health, as well as environmental protection.

1.3 Contributions

Our research makes significant contributions to the field of air pollution prediction by developing and evaluating models that address the complexity of atmospheric processes and the challenges of data variability and privacy. The contributions are structured into three main areas: a temporal prediction model, a spatiotemporal model, and a *Federated Learning* (FL) architecture.

1.3.1 PMForecast: A Temporal Prediction Model for Air Pollutants

The first contribution is the development of PMFORECAST, a temporal prediction model specifically designed to forecast air pollutants, like $PM_{2.5}$. This model leverages historical data to understand patterns and trends over time, enabling more accurate short-term and long-term predictions. PMFORECAST integrates advanced machine learning techniques to capture temporal dependencies in air quality data, providing insights that can inform public health interventions and environmental policies.

1.3.2 Spatiotemporal Modeling: Integrating Underground Sensor Networks

The second contribution addresses the spatial and temporal dynamics of air pollution by developing a *spatiotemporal model*. This model considers the complex interactions between pollutant sources and the environment across different locations and times. By integrating data from an underground sensor network, the model constructs a spatial-temporal grid that forecasts pollution levels for various geographical areas. This approach allows for a more comprehensive understanding of pollution dispersion, offering valuable information for regional air quality management and mitigation strategies.

1.3.3 Federated Learning Architecture: Enhancing Prediction with Mobile Sensors

The third contribution proposes a novel *Federated Learning* (FL) Architecture (FEDAIRNET) that leverages data from mobile sensors to enhance air pollution predictions while preserving user privacy. This FL architecture addresses the limitations of traditional centralized models by distributing the learning process across multiple devices, ensuring that sensitive data remains local. The FL approach not only improves prediction accuracy by incorporating diverse data sources but also enhances the model's resolution and stability. By enabling real-time updates and reducing the risk of data breaches, this architecture offers a practical solution for integrating mobile sensor networks into air quality monitoring systems.

These contributions collectively advance the state of the art in air pollution forecasting, providing robust models that can be applied in diverse urban environments while addressing key challenges such as data variability, computational demands, and privacy concerns.

1.4 Scientific publications and Vulgarization

- Maryam Rahmani, Suzanne Crumeyrolle, Nadège Allegri-Martiny, Amir Taherkordi,and Romain Rouvoy. *PmForecast: leveraging temporal LSTM to deliver in situ air quality predictions*. Environmental Science and Pollution Research, 2024 [RCAM+24].

- Maryam Rahmani, Suzanne Crumeyrolle, Nadège Allegri-Martiny, and Romain Rouvoy. *Forecasting Urban Air Quality: Integrating High-Resolution $PM_{.5}$ Data with GT-LSTM Modeling*. Under submission

- Presented "Advanced Air Quality Forecasting Using Temporal LSTM (TLSTM) Model" at the 13th Asian Aerosol Conference, November 3-7 2024.

- Participating on GDR RSD Summer School on Distributed Learning, September 19th 2023 and 20th, Lyon, France.

- Presented GT-LSTM in the Air Quality Data Analysis Workshop, University of Burgundy, February 9–10, 2023.
- Completed a one-month internship in Oslo, Norway, collaborating with Amir Taherkordi and his group Networks and Distributed Systems, Department of Informatics, University of Oslo, November 2022.
- Presented FEDAIRNET at the Fed-Malin (Federated Machine Learning over the Internet) workshop, Paris, June 16, 2022.

1.5 Developed Frameworks, Code, and Data

In this project, all the developed code and data are available in the following GitHub repositories:

1. Temporal Model (PMForecast),
2. Graph Temporal Model (GT-LSTM),
3. Simulated Data and PoI Attack Tool,
4. Federated Framework (FedAirNet).

[1]

1.6 Outline

This dissertation is organized into several key chapters, beginning with a literature review and background in chapter 2. This chapter provides insights into the state-of-the-art and existing models related to air pollution prediction, covering research from past to present across various scales and data sources. This foundation leads to three principal chapters, each focusing on a distinct aspect of air pollution prediction and monitoring.

Chapter 3 explores the development and application of a self-adaptive temporal model for predicting $PM_{2.5}$ levels under real atmospheric conditions. The chapter includes a brief introduction and review of related works in section 3.1, followed by a comprehensive methodology in section 3.2. It details the data acquisition, instruments used, and preprocessing steps in section 3.3, and describes the model architecture, training, and evaluation processes in section 3.4. The chapter concludes with a discussion of the results and achievements.

[1] All repositories are accessible via GitHub: https://github.com/Maryamr92.

Chapter 4 extends the temporal modeling approach by incorporating spatial dependencies through a graph neural network. This chapter investigates the integration of spatial and temporal patterns to enhance $PM_{2.5}$ concentration forecasting. It starts with an introduction and review of related work in section 4.1, followed by a description of the methodology in section 4.2. Details about the dataset used for experimentation are provided in section 4.3, while section 4.4 presents the experimental results. The chapter concludes with section 4.5, which discusses the implications of the findings and outlines potential directions for future research.

Chapter 5 introduces a FL framework aimed at optimizing local-scale air pollution monitoring and prediction while addressing privacy concerns. This chapter examines the potential of mobile sensor data and explores the trade-offs between data privacy and model performance. It begins with a review of the state of the art in section 5.1, followed by a detailed description of the methodology and dataset used for experimentation in section 5.2. The data, including its sources, characteristics, and preprocessing steps, are examined in section 5.3. section 5.4 presents the experimental results, highlighting performance metrics and key findings. Finally, section 5.5 interprets the results, discussing their implications and significance in the context of FL for air quality monitoring.

In conclusion, chapter 6 synthesizes the results from all chapters, emphasizing the benefits of the employed methods and suggesting potential avenues for future research. This chapter provides a comprehensive summary of the research contributions and their implications for air quality management and public health.

Chapter 2

BACKGROUND & CONTEXT

Air quality monitoring and modeling are crucial for understanding and mitigating the effects of pollution on human health and the environment. This chapter provides a comprehensive overview of various aspects of air quality monitoring, from traditional chemical and physical modeling to modern machine learning techniques, spatial and temporal analysis, and the use of crowdsourced data with mobile sensors. Additionally, it discusses distributed learning techniques such as federated learning, highlighting the importance of privacy and data security in this context.

2.1 Evolution of Air Quality Modeling

Climate change, though often regarded as a contemporary issue, has been a subject of scientific inquiry for over a century [Arr96]. In the early 19^{th} century, scientists began to recognize that human activities, particularly those associated with the Industrial Revolution were inducing gradual shifts in the climate of the Earth [Rud03]. By the mid-20^{th} century, with the advent of more sophisticated computer technologies and an enhanced understanding of climate science, researchers were better equipped to model and predict the consequences of anthropogenic actions on global climate [Mac04]. Despite these early efforts, public and governmental recognition of the link between industrial activities and environmental degradation remained limited for many years. The intricate relationship between air pollution and climate change requires the application of advanced modeling techniques. Traditionally, the development of air pollution modeling has been approached through primary methodologies: physical modeling, which examines the transport and dispersion of pollutants in the atmosphere and chemical modeling, which focuses on the composition and reactions of pollutants [SP16]. Both approaches have been developed simultaneously and utilize numerical simulation methods. Physical and chemical models are complementing each other in understanding air pollution dynamics.

2.1.1 Physical Modeling

Physical modeling simulates the transport and dispersion of pollutants within the atmosphere considering factors, like wind patterns, temperature, and topography. These models have evolved significantly over time in terms of complexity and computational demands.

Early models, such as Gaussian plume models developed in the mid-20^{th} century [Zan90], offered a simplified representation of pollutant movement based on meteorological conditions. These models assume that pollutants disperse in a Gaussian distribution, making them computationally light, with typical run times in the range of seconds to minutes on modern machines. While these models provided foundational insights, their limitations in capturing complex atmospheric processes became evident. For example, Bruscaa *et al.* [BFL$^+$16] in a study conducted, at the University of Catania, an experimental campaign using a wind tunnel testing the dispersion of PM_{10} particles. The researchers developed a Gaussian plume model to simulate the experiment's conditions, comparing the model's predictions with empirical data to assess its accuracy and limitations in small-scale systems. This study highlighted how Gaussian models, despite their simplicity, are often inadequate for real-world, small-scale systems where turbulence and variable terrain are critical factors.

To overcome the limitations of Gaussian models, more sophisticated approaches, such as Eulerian grid models were developed [KWSR19]. These models divide the atmosphere into grid cells, allowing them to simulate more detailed spatial and temporal variations in pollutant concentrations. Eulerian models incorporate differential equations to represent both the physical transport of pollutants and chemical transformations they may undergo, adding complexity. The resolution of these models can vary, but typical runs use grids ranging from 1 km to 50 km with vertical stratification, depending on the study area and available computational power. For instance, Egmond *et al.* [vK83] developed an Eulerian grid model to simulate air pollution transport across the Netherlands using a 32×32 grid with a 15 km resolution. While this model incorporated vertical stratification and produced valuable insights, its ability to capture fine spatial variations (e.g., at street level) was limited, likely due to uncertainties in emission data or unaccounted atmospheric processes such as localized turbulence or secondary pollutant formation.

The computational demands of Eulerian models are significant compared to Gaussian models [DC08]. A typical Eulerian simulation for a regional study can take from several hours to days, depending on the grid resolution, domain size, and model complexity. High-resolution simulations, particularly in urban settings, can be computationally prohibitive without access to high-performance computing (HPC) resources. As a result, physical models often rely on input data, such as emission inventories and meteorological reanalysis, that are averaged over space and time to manage this complexity and computational load. This averaging can reduce the

precision of predictions, particularly in heterogeneous and complex environments, like cities, where pollution sources and atmospheric conditions can vary dramatically over short distances.

In addition to Gaussian and Eulerian models, Lagrangian models, which track individual particles or air parcels through the atmosphere, offer another approach to physical modeling [AKH+24]. These models, while highly detailed, can be computationally expensive for large-scale studies, due to the need to track thousands or millions of particles over time. Lagrangian simulations can take several days or even weeks to run, depending on the number of particles and the spatial extent of the study. They are often used for specific applications, like tracking pollutant plumes over long distances, such as volcanic ash or wildfire smoke, where a detailed understanding of transport pathways is crucial [RGL+20].

In general, the computing time and complexity of physical models depend on factors, like grid resolution, domain size, and whether they account for both physical transport and chemical processes. While these models are invaluable for understanding pollutant dispersion and interactions in the atmosphere, they are often constrained by computational limits, requiring trade-offs in model resolution and input data quality. Consequently, even the most advanced physical models may struggle to accurately capture the full complexity of air pollution in densely populated or rapidly changing environments.

2.1.2 Chemical Modeling

Chemical modeling focuses on the transformation of pollutants within the atmosphere through complex chemical reactions. These models simulate the formation and removal of pollutants, like ozone, nitrogen oxides, and particulate matter using intricate reaction mechanisms [MAD89]. However, it is impossible to know and model all the chemical processes accurately, as they depend on a wide range of factors, including the specific compounds present in the atmosphere, as well as environmental conditions such as relative humidity, temperature, and pressure. For example, ozone formation alone can involve 40–60 different reactions depending on the local atmospheric conditions and the compounds available.

While early models were primarily laboratory-based and focused on specific, isolated reactions, advancements in computational power have enabled the development of large-scale atmospheric chemistry models capable of simulating regional and global air pollution patterns [LLM+23]. These models must include a wide array of reaction mechanisms to account for the different pathways pollutants can take in the atmosphere. For instance, ozone formation depends on the interaction of nitrogen oxides (NOx), volatile organic compounds (VOCs), and sunlight, with each compound undergoing multiple reactions, which add layers of complexity to the models.

Unlike physical models, which emphasize pollutant transport and dispersion, chemical models delve into the underlying chemical processes. Accurate chemical modeling depends not only on detailed understanding of reaction mechanisms but also on the availability of reliable kinetic data. Yet, even with the most comprehensive reaction schemes, uncertainties remain due to the sheer complexity of atmospheric chemistry, the interactions between multiple pollutants, and variable environmental conditions.

Furthermore, the representation of atmospheric conditions, including temperature, humidity, and sunlight, is vital for simulating complex chemical processes accurately. For instance, U. Nopmongcol *et al.* [NKT+12] published a study in 2012 employing the CAMx photochemical grid model to simulate ozone (O_3) and PM concentrations across Europe in 2006. The study revealed underestimations for most pollutants, particularly PM_{10}, and identified key factors influencing model outcomes, such as emission inventories, meteorological conditions, and model parameterizations. These results underscore the complexities of air quality modeling and the ongoing need for model refinement.

Notable examples of chemical air quality models include the *Regional Acid Deposition Model* (RADM) [MCB+93], the *Community Multiscale Air Quality* (CMAQ) model [LC12], and the *Comprehensive Air Quality Model with Extensions* (CAMx) [NKT+12], all of which remain highly relevant today for simulating atmospheric processes. These models range in complexity, from 0D models that focus on chemical processes at a single point over time, to 1D models that incorporate altitude profiles, and 2D models that track changes across latitude or longitude in addition to altitude and time. The most advanced are 3D models, which account for latitude, longitude, altitude, and time, offering a comprehensive understanding of pollutant behavior across multiple dimensions.

2.1.3 Integration of Physical and Chemical Processes

The limitations of using physical and chemical models in isolation highlighted the need for an integrated approach, leading to the development of *chemistry-transport models* (CTMs) [VMC21]. CTMs combine meteorological conditions with chemical reactions to provide a more accurate and comprehensive representation of air pollution processes. By integrating both physical transport and chemical transformation components, CTMs offer a holistic view of air pollution.

Chemistry-transport models address the limitations of isolated models by coupling physical and chemical processes. These models simulate the transport, dispersion, and chemical reactions of pollutants, offering valuable insights into the formation and evolution of air pollution. Incorporating detailed chemical mechanisms and meteorological conditions, CTMs are up to now essential for air quality management and policy development.

The advancement brought by CTMs is significant, as they enhance the accuracy and realism of air pollution simulations by combining physical dispersion with chemical reactions. For example, a study by M. Schaap *et al.* [SCH+15] investigated the impact of model resolution on air pollution simulations using the CAMx model across Europe. The study found that increasing model resolution from 50 km to 10–20 km improved the representation of spatial pollution gradients, especially in urban areas. However, the overall model performance—as measured against monitoring data—reported on limited improvement, suggesting that while higher resolution is beneficial, other factors, like emission inventory accuracy, also play a critical role in model outcomes.

This foundational work in integrating physical and chemical modeling is crucial for advancing air quality prediction and management. It also provides a critical basis for applying advanced techniques, such as machine learning, to enhance our ability to predict pollutant levels, identify sources, and understand dispersion patterns.

2.2 The Need for Advanced Techniques

While traditional physical and chemical models have provided valuable insights into air pollution, their limitations in accurately capturing the complexity of atmospheric processes and predicting future trends have become increasingly apparent [SP16].

To address these challenges, the integration of AI and *Machine Learning* (ML) techniques has emerged as a promising approach to advance air quality modeling [ZRY+22]. ML techniques, often employing a variety of statistical methods including both linear and nonlinear algorithms, have been extensively applied in this field as an alternative to traditional numerical modeling approaches.

Statistical predictive methods, which often leverage time series analysis, are based on modeling approaches that predict upcoming air quality by relying on historical data [Wil16, SS17]. In contrast to computationally intensive numerical methods that simulate intricate mechanisms of pollutant emission, diffusion, aging, and deposition, statistical approaches offer a more efficient alternative [Wil19].

2.2.1 Data Collection: The Backbone of AI and ML in Air Quality Modeling

Before exploring the complexities of air pollution modeling, it is essential to establish a robust data collection framework. Accurate and comprehensive air quality data are critical for effective modeling and prediction. Traditionally, ground-based monitoring stations equipped with sensors measure pollutants, such as particulate matter (PMs), ozone (O_3), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). These monitoring stations serve as vital sources of air quality

data, providing important information about local air conditions [AISW24]. However, they have limitations, as they cannot capture the full extent of air quality across larger areas.

Satellite remote sensing provides a comprehensive view of our planet. From their high orbits, satellites scan the Earth's surface, offering valuable insights into air pollution patterns [ZZX+23]. While they provide a broad perspective, there are limitations; clouds can obstruct the view, and distinguishing between different pollutants remains a complex challenge.

To address these gaps, sensor networks have emerged as a promising solution. These networks of low-cost sensors are distributed throughout a city, providing detailed information on air quality variations [KRA10]. They help identify pollution hotspots and temporal patterns with significant precision. However, managing the vast amounts of data generated by these networks presents a considerable challenge.

In recent years, mobile sensors have emerged as a transformative tool in data collection. These devices, integrated into everyday items such as cars, buses, trucks, drones, and bicycles, turn ordinary citizens or mobile devices into pollution detectors, creating a vast, distributed network capable of capturing air quality data on an unprecedented scale. This crowdsourced approach offers a significant opportunity to enhance data collection efforts across diverse geographical areas [SBW+22]. However, ensuring the accuracy and consistency of these measurements from such a widespread sensor network poses complex challenges. Additionally, concerns surrounding the sensitive nature of personal data collected by these sensors must be carefully addressed to protect individual privacy.

This intricate data tapestry forms the foundation upon which air quality models are built. By understanding the strengths and limitations of each data source [PFH24], scientists can develop more robust and reliable models to predict and manage air pollution. To overcome the challenges posed by the limitations of individual data sources, it is proposed to integrate these disparate data streams to create a truly comprehensive picture of air quality. For instance, Yee Leung *et al.* [LZL+19] addresses the challenge of insufficient air pollution data from traditional monitoring stations. To improve data coverage, the authors propose combining data from both stationary and mobile sensors. This integrated dataset is then used to estimate air pollutant concentrations at specific locations and times. The study not only enhances air pollution monitoring but also lays the groundwork for future research in spatiotemporal data analysis.

2.2.2 Air Quality Modeling: Bridging the Gap Between Data and Prediction

Temporal Predictions: Capturing Air Quality Trends

Temporal analysis examines how air quality changes over time. Techniques, such as time series analysis and seasonality detection, are employed to identify trends, periodic patterns, and temporal anomalies in pollutant levels. A diverse range of statistical methods, including ML techniques, encompassing both linear and nonlinear algorithms, have been widely applied in this domain.

Common approaches include the utilization of ML algorithms, such as *Support Vector Regression* (SVR) [ZH19] and *Autoregressive Integrated Moving Average* (ARIMA) [CLAL23], both known for their linear characteristics. For instance, Bassirou Ngom *et al.* [NDS+21] present a novel integration of system observations from various stations with a multi-agent simulation, providing a model for assimilating PM_{10} pollution data through real-time simulation based on the autoregressive ARIMA method.

Additionally, non-linear models, like *Gaussian Process Regression* (GPR) [HLC+23], *Gradient Boosting* (XGBoost) [LAL+22], *Artificial Neural Networks* (ANN) [GHW23], and deep learning algorithms, such as *Recurrent Neural Networks*s (RNNs) [TW22a] offer effective solutions. These methods leverage the power of non-linearity to capture complex relationships in diverse datasets. For example, studies such as [GHL+20] show that combining correlation analysis with ANNs and *wavelet-enhanced ANNs* (WANNs) effectively reveals both linear and non-linear relationships between *air pollution indices* (API) and meteorological variables.

Among these techniques, RNN methods have garnered significant attention from researchers due to their ability to model temporal dependencies in time-series data effectively. Unlike traditional machine learning models, which often treat each data point independently, RNNs are specifically designed to handle sequential data, where past data points influence future outcomes. This makes them particularly suited for air quality forecasting, where pollutant levels at a given time are influenced by historical data due to factors such as meteorological conditions and emission sources [TW22b].

Their capacity to maintain a memory of past data allows RNN-based models to capture long-term dependencies within $PM_{2.5}$ input data, providing more accurate and context-aware predictions. This feature is especially relevant to the research community, as it allows for better forecasting of air pollution trends. Notable research efforts have been directed towards air quality forecasting using models, such as *Gated Recurrent Unit* (GRU) [ZZZ+22, PT23] and particularly LSTM [AQ19, GSMP23, ZH18, BLC18]. Zhang Qi *et al.* [ZHLL22] have made significant strides by integrating domain-specific features with a hybrid *Convolutional Neural Network* (CNN)-LSTM structure, achieving superior accuracy in fine-grained air pollution estimation

and prediction compared to similar baselines. Verma Ishan *et al.* [VAMD18] introduce a bidirectional LSTM model to predict $PM_{2.5}$ severity levels, significantly improving prediction accuracy by leveraging a set of three bidirectional LSTMs and incorporating weather data from multiple locations in New Delhi.

Additionally,Yi-Ting Tsai *et al.* [TZC18] demonstrated the effectiveness of combining RNNs with LSTMs in predicting hourly $PM_{2.5}$ concentrations at 66 monitoring stations in Taiwan using *Environmental Protection Agency* (EPA) data. Yuan *et al.* [EG23] proposed a novel hybrid self-attention LSTM model for accurately predicting long-term $PM_{2.5}$ levels in classrooms, surpassing existing methods in terms of both accuracy and computational efficiency. Moreover, researchers have explored the benefits of combining LSTM with other techniques to enhance predictive performance. For instance, Abimannan *et al.* [ACL20] integrated LSTM with *Multivariate Variate Regression* (MVR) to improve $PM_{2.5}$ prediction accuracy, particularly over a multi-years observation period when seasonal variations are observed. Their comparative analysis highlighted the superior performance of the LSTM/MVR model in predicting hourly $PM_{2.5}$ concentrations compared to traditional LSTM approaches.

Temporal statistical models, such as regression and time series analysis, have been used to predict air pollution levels. However, these methods often struggle to capture the complex, non-linear relationships inherent in air quality data [SKR24]. To overcome these limitations, machine learning techniques, including *Support Vector Machines* (SVMs), ANNs, and RNNs, have been explored [MV23, ABA$^+$20, BAEO20, JBG19]. While these models show promise in capturing temporal patterns, they may overlook spatial dependencies. They may inherently face limitations in capturing the complex spatial variations of pollutants, often necessitating the use of additional modeling techniques.

Modeling Air Pollution in Space and Time

Air pollution levels can vary drastically across short distances due to factors such as traffic, industrial emissions, and meteorological conditions. While temporal models capture changes over time, they often struggle to account for these spatial differences. To address this, spatial analysis techniques, including *Geographic Information Systems* (GIS) and spatial statistics, are employed to map and analyze pollutant distributions.

Belavadi *et al.* [BRRM20] developed a scalable architecture combining wireless sensor networks and government data for real-time air quality monitoring and forecasting. While demonstrating the potential of such systems, the study highlighted the challenges of temporal variations across regions. This underscores the need for adaptive models tailored to specific urban environments. Several studies, including those by Ghufran [DAB22] and Tien-Cuong [BLC18], have employed LSTM models to capture temporal dynamics while emphasizing the importance of meteorological factors.

Recognizing the limitations of purely temporal models, researchers have increasingly focused on spatiotemporal modeling. These models aim to capture complex interactions between air quality measurements across different locations and time points. Combining (CNNs) with LSTMs has emerged as a promising approach for modeling these relationships [HK18, QYZ+19]. Gilik *et al.*[GOO22] proposed a CNN-LSTM model to predict pollutant concentrations in urban areas, leveraging spatiotemporal dependencies for improved accuracy. Similarly, Unjin Pak *et al.* [PMR+20] applied this method to forecast $PM_{2.5}$ concentrations across 384 monitoring stations, covering the entirety of China with Beijing as the central focus, over a 3-year period (January 1st, 2015 to December 31st, 2017). However, the model's complexity and substantial data requirements present challenges to its broader applicability and generalizability.

Zhang Qi *et al.* [ZHLL22] introduced Deep-AIR, a hybrid CNN-LSTM model that integrates domain-specific features, like street canyon effects, to improve air pollution estimation and forecasting. The model uses 1x1 convolutional layers to capture complex spatial interactions between pollutants and urban dynamics, offering a more detailed view of city-wide air quality. However, its data granularity is limited by sensor distribution, which typically covers larger areas, like neighborhoods or cities. Although street canyon effects enhance large-scale predictions, they may not fully capture pollution variations at smaller scales, such as individual streets, where building configurations and micro-climates play a key role. To model these finer details, more sensors and higher-resolution urban data, such as street-level wind patterns, would be necessary.

Graph networks, such as GCNs, have proven to be powerful tools for modeling spatial dependencies in air pollution prediction. By representing monitoring stations as nodes on a graph and defining edges based on geographical proximity or air quality similarity, GCNs effectively capture spatial relationships [HBL15, DBV16]. Building on this approach, Hofman *et al.* [HDQ+22] demonstrate the advantages of combining GCNs with RNNs to enhance forecasting accuracy by addressing both spatial and temporal patterns. Their research highlights the potential of using mobile sensor data to create high-resolution air quality maps. By integrating data from various mobile sources (for example, the one of the datasets consisted of 323,691 NO2 measurements), Hofman *et al.* develop a data-driven model that surpasses traditional interpolation methods in real-time air quality monitoring. Despite these advancements, the accuracy of such models remains dependent on the quality and coverage of the data, and their scalability across different geographic regions and pollutant types.

Ge Liang *et al.* [GWZ+21] introduce the *Multi-scale Spatio-Temporal Graph Convolution Network* (MST-GCN), an advanced deep learning model for air quality prediction. MST-GCN excels at capturing both spatial correlations and long-term temporal dependencies within air quality data. The authors demonstrate its superiority over baseline models, highlighting its potential for tackling other multi-source data challenges. The sophisticated spatio-temporal

attention mechanisms of MST-GCN enhance its applicability to data challenges of multiple sources, although its complexity can be computationally intensive.

By capturing both spatial and temporal dependencies, these models aim to enhance the accuracy and reliability of air pollution forecasts. Emphasizing computational efficiency, scalability, and high-resolution predictions within urban areas, these approaches often leverage inexpensive underground sensors for PM and other environmental data. While seeking to improve prediction resolution without relying on multiscale data, these centralized models may encounter limitations in terms of data volume, computational resources, and privacy concerns.

2.3 Centralized vs. Distributed Air Quality Modeling

Traditional air quality modeling primarily relies on centralized approaches, where data is collected, processed, and analyzed at a single, central location. While effective in certain scenarios, these methods have significant limitations, particularly concerning data privacy, computational efficiency, and scalability [Air, DGL+13]. To overcome these challenges, distributed learning has emerged as a promising alternative [VBT16]. By distributing data and model training across multiple devices or servers, this approach enhances computational efficiency and scalability [MMPJY22].

Most distributed learning efforts leverage IoT devices for data collection, processing, and transmission to a central server. For instance, in [NFZM19], the authors review existing IoT architectures and propose a comprehensive system incorporating both static and mobile sensors for air quality monitoring. These sensors perform basic data processing (edge computing) and validate data through redundancy. Gateways act as intermediaries, conducting additional processing (fog computing) and transmitting data to a cloud-based load balancer. The cloud facilitates extensive data processing, including the application of machine learning and deep learning techniques, while storing raw data in NoSQL databases and collecting data in a data warehouse. While this approach enhances computational efficiency and system stability, it falls short in addressing data sensitivity and privacy concerns.

FL, a prominent distributed learning technique, has gained significant traction in recent years [LSTS19, MMRyA16]. Pioneered by Google in 2017, FL enables collaborative model training without sharing raw data. Instead, devices compute local model updates based on their data and transmit these updates to a central server for aggregation. This process iteratively improves the global model while preserving data privacy.

A study by Nguyen Do-Van *et al.*[NZ21] investigates a FL approach for predicting air pollution in smart cities, with a focus on improving the efficiency of predictive model training using environmental IoT data. Traditional centralized data processing methods often face latency issues, but this research proposes distributing the training process across multiple regions. *Local*

Convolutional Recurrent Neural Networks (CRNNs) are used to predict air quality at each location, with these local models sharing distilled knowledge through a global model. This approach enhances overall accuracy while reducing the need for extensive data transmission. The study shows that CRNN models effectively capture local spatio-temporal data and benefit significantly from knowledge sharing across cities via federated learning. Additionally, this method allows new regions to quickly develop optimized local models using the global model, even without prior data contributions. However, the study also identifies limitations in the current model aggregation process, noting that it does not account for spatial relationships between regions, and the model's reliance on complex data remains a challenge.

The proliferation of interconnected devices, including smartphones, IoT sensors, and wearable devices, has led to an explosion of data generation. These data, often characterized by their geospatial nature, have immense potential for applications in urban planning, traffic management, and environmental monitoring. However, collecting and processing such data raises significant privacy concerns due to the potential exposure of sensitive user information.

2.3.1 Privacy in Federated Learning

Traditional centralized models require the collection and storage of raw data, which can expose sensitive information and raise concerns about data breaches and misuse. FL mitigates these risks by keeping data localized on devices and only sharing model parameters, significantly reducing the potential for privacy violations. For instance, in healthcare, FL can enable the development of personalized treatment plans through decentralized analysis of patient data [HNA24, SSS+24].

Despite these advantages, FL still encounters privacy challenges, including potential vulnerabilities that could lead to the unintended exposure of sensitive information through model updates. Thus, ongoing research aims to enhance FL's privacy-preserving capabilities through the development of robust aggregation methods.

A comprehensive analysis of FL's security and privacy is provided in the paper "*A Survey on Security and Privacy of Federated Learning*" [MPP+21]. The paper covers the core principles of FL, emphasizing its advantages in preserving user privacy and its suitability for managing sensitive data. It also addresses key security threats, such as communication bottlenecks, poisoning attacks, and backdoor attacks, along with privacy risks, such as inference attacks. Notably, the authors offer a detailed classification of FL techniques and explore various strategies to mitigate these risks. Figure 2.1 illustrates the classification of approaches and techniques concerning privacy within existing studies. This figure illustrates the key components and frameworks of FL, highlighting aspects such as data availability, network topology, and the various architectural frameworks utilized. It emphasizes the interplay between these elements in facilitating effective FL implementations.

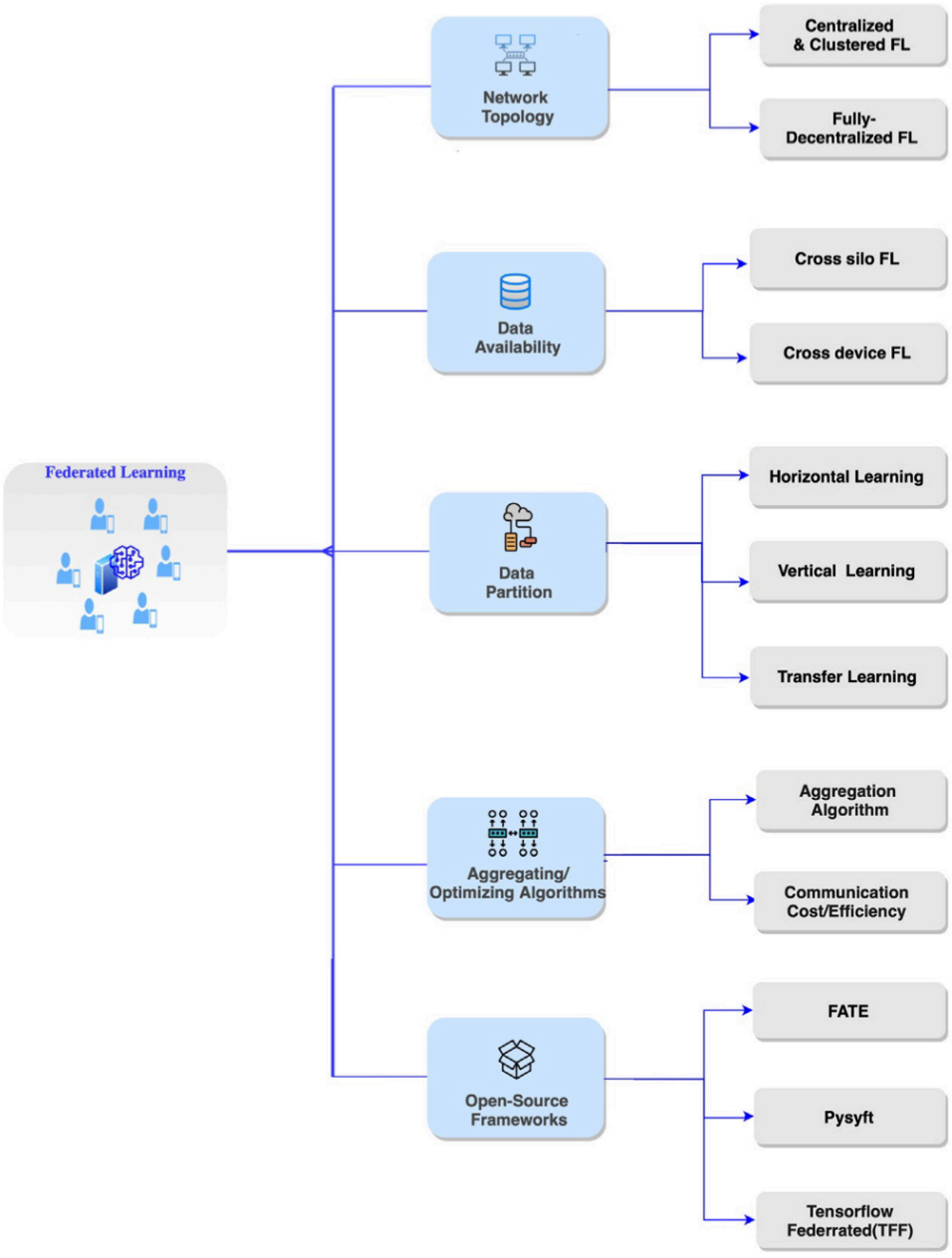


Figure 2.1: Overview of FL approaches and technologies: This figure categorizes various FL methods based on different contexts, highlighting their classifications and applications [MPP+21]

.

The paper conceptualizes the technique, the associated issues, and the appropriate approaches for addressing them. Additionally, the authors emphasize the need for continued research to tackle the identified challenges and improve the security and privacy of FL systems to facilitate their wider adoption.

2.3.2 Collaborative Learning in Human Mobility Analytics

One of the most promising applications of FL is in the domain of human mobility analytics, where privacy concerns are paramount. The use of geospatial data in analyzing human movement patterns, traffic flows, and location-based services requires robust privacy measures.

A recent comprehensive survey, "*Survey of Federated Learning Models for Spatio-Temporal Mobility Applications*," provides a detailed analysis of existing FLs models applied to this domain [BMH+24]. The survey examines 38 studies in 2024, offering valuable insights into the strengths and limitations of current approaches and serving as a foundation for further research.

The reviewed approaches in the document are categorized into the following main groups:

- **Trajectory Predictive Approaches**: These approaches focus on predicting the next point in a user's trajectory while addressing challenges such as non-independent and identically distributed (non-IID) data, which refers to data that is heterogeneous and does not follow a uniform distribution across samples. Additionally, privacy concerns are also taken into account.

  For example, the *Deep Federated Reconstruction* (DFR) model is designed to predict the next point in a user's trajectory while ensuring that individual mobility traces remain on local devices, preserving privacy. The DFR model tackles the challenge of non-IID data by using a shared global model, periodically updated with local model updates from various clients [PCK+23].

- **Traffic Flow Prediction Approaches**: These methods predict traffic patterns, crucial for real-time applications like traffic management and pollution control. They often require real-time data processing and adaptation. The FedGRU model, for instance, utilizes *Gated Recurrent Units* (GRUs) in a federated learning framework to predict traffic flow across different regions. This approach processes large datasets locally, reducing the need for centralized data storage, which is essential for real-time traffic flow prediction [LYK+20].

- **Clustering-Based Approaches**: Clustering techniques in FL aim to group similar data points, such as locations or users, to identify patterns or communities within spatio-temporal data. These methods often focus on organizing devices into clusters for efficient model sharing. For instance, hierarchical clustering using *Convolutional Neural Networks* (CNNs) improves FL performance in non-IID data settings [BFA20]. Another approach,

dynamic GAN-based clustering, enhances time series forecasting, like predicting cell tower handovers, by adapting to evolving clusters [KHH+20]. An example in spatio-temporal FL is the Federated-Deep embedded clustering(F-DEC) method, which clusters urban communities by analyzing heatmap images of mobility trajectories, thus uncovering movement patterns in cities while ensuring data privacy [MSM21].

- **Top-N Location-Based Recommendation Approaches**: Location-based recommendations aim to identify the top-N *Points of Interest* (POIs) for a user based on their historical data. Unlike trajectory prediction, this task often requires fewer real-time updates. Although online learning might not seem essential, evolving user preferences suggest it could be beneficial. Privacy solutions in existing models, which avoid sharing all parameters, may still be inadequate. Additionally, there is a lack of exploration into robustness against attacks, such as shilling attacks or fake profiles. The focus remains on evaluating the main works in this domain, using metrics similar to those in trajectory prediction. For instance, in *Federated Pair-wise Learning* (FPL) method by Vito Walter Anelli *et al.* [ADN+21], the authors adapt the *Bayesian Pairwise Ranking* (BPR) algorithm to federated learning. It allows users to locally train sensitive embeddings and share them selectively, while less sensitive parameters are federated. This flexible framework helps maintain model convergence and is evaluated using the Foursquare dataset, showing improved performance over federated movie recommendations.

- **Privacy and Attacks in Spatio-Temporal FLs Models**: FL aims to protect privacy by sharing model parameters instead of data, but this can still leak sensitive information. To improve privacy, three main approaches have been proposed: (i) sharing less data [AAT22, GLC+21], (ii) using *Differential Privacy* (DP) to add noise [FRS+20], and (iii) employing *Secure Multi-Party Computation* (SMPC) [PDSE23]. Each method has limitations, such as potential exposure through embeddings or vulnerabilities to malicious users. A more robust solution uses Local DP and peer-to-peer secret sharing for sensitive data, offering better privacy while maintaining performance. Additionally, new protocols are being developed to assess and mitigate re-identification risks without requiring a trusted curator. A more privacy-oriented solution was proposed by Chaochao Chen *et al.* [CWF+20]. They aggregated less sensitive embeddings using SMPC and categorized sensitive embeddings into two parts: those related to POIs and those related to users. They used Local DP to add noise at the user level to the POI-related embeddings before sharing them with the server.

The diverse applications of FL in spatio-temporal mobility analytics underscore its effectiveness in addressing privacy, scalability, and real-time processing challenges. FL excels in predicting trajectories, managing traffic flow, and clustering data for community detection, all

while safeguarding sensitive geospatial information. Integrating mobility aspects with crowdsensing and other techniques makes these approaches highly applicable in real-world scenarios.

For example, FedSense [YDL+24b] addresses the challenge of protecting location privacy in machine learning-based crowdsensing applications, where existing privacy-preserving methods often fall short. FedSense combines reinforcement learning (RL) with FL to optimize task allocation without exposing raw data, allowing participants to complete tasks efficiently while safeguarding their location data and minimizing the computational burden on their devices. The global model in FedSense aggregates parameters from local models to avoid task collisions and ensure high task completion rates. Theoretical analysis and simulations demonstrate that FedSense provides strong privacy protection with minimal performance loss compared to centralized systems that compromise location privacy.

Building on these insights, our focus now shifts to applying FL to air quality monitoring by integrating it with crowdsourced mobile sensors—a pioneering effort in this field. Air quality monitoring involves collecting data from numerous mobile sensors while ensuring privacy. FL's decentralized approach is ideally suited for this task, enabling local model training on sensor data collected by citizens and aggregating insights without exposing individual data, thereby addressing privacy and security concerns in spatio-temporal FL models. While several studies have explored the use of FL with fixed stationary data, this approach marks a significant advancement by leveraging mobile data sources.

2.4 Key Federated Learning Frameworks

A diverse ecosystem of federated learning frameworks has emerged to support researchers and developers in building and deploying decentralized machine learning models, each offering different levels of abstraction, scalability, and privacy features. Notable frameworks in this space include TensorFlow Federated, PySyft, Flower, FedLab, PaddleFL, FederatedScope, LEAF, and FedML. Each of these frameworks brings unique strengths to address specific challenges and cater to various use cases.

Among these, Flower stands out as a particularly attractive option due to its flexibility, scalability, and user-friendly features [BTM+22]. It offers high customization, enabling researchers to tailor the federated learning process to their specific needs. Designed to handle large-scale systems, Flower is well-suited for real-world applications. Its ease of use is supported by comprehensive documentation, making it accessible to researchers and developers at all levels. Additionally, Flower's interoperability with various machine learning libraries ensures seamless integration with existing tools and workflows. The framework benefits from an active community of developers and researchers, contributing to its continuous development and providing valuable support.

Table 2.1 provides a summary of existing federated learning frameworks with brief descriptions.

Table 2.1: Overview of Federated Learning Frameworks

Framework	Description
TensorFlow Federated (TFF)	Developed by Google, TFF is a Python framework for federated learning algorithms, providing tools for data management, communication, and model aggregation. *Reference:* TensorFlow Federated Documentation
PySyft	A Python library built on PyTorch that enables secure and private machine learning with cryptographic primitives and federated learning functionalities. *Reference:* PySyft GitHub Repository
Flower	A flexible and scalable federated learning framework that supports various federated learning strategies and can be customized for different use cases. *Reference:* Flower GitHub Repository
FedLab	A PyTorch-based framework designed for efficient federated learning research, offering modularity and extensibility. *Reference:* FedLab GitHub Repository
PaddleFL	Part of the PaddlePaddle ecosystem, PaddleFL supports federated learning across various applications and scales. *Reference:* PaddleFL Documentation
FederatedScope	A flexible framework for federated learning research and deployment, supporting a range of algorithms and integration options. *Reference:* FederatedScope GitHub Repository
LEAF	A benchmarking suite for federated learning algorithms, providing standardized datasets and metrics for evaluation. *Reference:* LEAF GitHub Repository
FedML	A comprehensive federated learning library supporting various machine learning frameworks, offering tools for model implementation and management. *Reference:* FedML GitHub Repository
FedScaleFATE	Integrates with the FATE platform to provide scalable and secure federated learning solutions. *Reference:* FedScaleFATE GitHub Repository

This research aims to bridge the gap in air quality monitoring by applying federated learning (FL) to integrate data from both fixed stationary and mobile sensors for the first time. By leveraging federated learning with crowdsourced data from mobile sensors provided by citizens, we seek to enhance environmental monitoring systems. Our goal is to improve these systems'

responsiveness, accuracy, and scalability while ensuring the privacy of the individuals carrying the sensors.

Chapter 3

PMFORECAST

3.1 Introduction

Several machine learning algorithms have been explored for air quality monitoring and prediction, as described in the previous chapter. Among these, LSTM networks stand out due to their superior flexibility and adaptability compared to traditional models like standard RNNs or autoregressive models. LSTMs are particularly well-suited for dynamic and complex air quality datasets because they can handle a wide range of input features and output formats. Their key advantage lies in their ability to capture complex temporal dependencies, such as the relationships between historical air quality measurements, meteorological factors, and other relevant variables. This long-term memory capability is essential for accurate forecasting of Air Quality values, identifying trends, detecting anomalies, and pinpointing potential sources of pollution.

LSTMs are highly effective in a range of air quality applications, including predicting future AQI levels, uncovering patterns in pollution data, detecting unusual events, and aiding in source identification. These strengths make LSTMs an invaluable tool for developing predictive models that support public health policies and environmental management strategies. By harnessing the unique capabilities of LSTM networks, researchers can build accurate, reliable, and actionable air quality prediction models that address the complexities inherent in real-world environmental data.

3.1.1 Related Works

LSTM networks have become a cornerstone in air quality prediction due to their ability to capture temporal dependencies in sequential data. However, the basic LSTM model often encounters limitations, such as high computational costs, slow convergence, and difficulties in handling complex data patterns. To address these challenges, recent studies have increasingly focused on developing hybrid models that combine LSTM with other algorithms. These hybrid

approaches have demonstrated enhanced predictive capabilities, making them particularly effective in the context of air quality forecasting.

For example, Wu Yang *et al.* [WQH24] developed the DVMD-Informer-CNN-LSTMs model, which tackles nonlinear and unstable patterns and characteristics of *air quality index* (AQI) data. Their approach optimizes *Variational Mode Decomposition* (VMD) parameters using the Dung Beetle Algorithm and incorporates advanced deep learning methods like CNN and Informer architectures. This integration significantly enhances prediction accuracy compared to conventional models, demonstrating the model's effectiveness in managing complex, real-world air quality data. However, the model's high computational complexity and focus on data from specific regions highlight areas for future improvements, such as optimizing computational efficiency and broadening data applicability.

Another innovative approach by researchers proposed an enhanced Vanilla LSTM model [FZL23], named IVLSTM-MCMR, designed with multichannel input and multiroute output mechanisms. This improved LSTM structure reduces parameter count, accelerates convergence, and stabilizes the training process. The multichannel input module employs a linear similarity dynamic time warping algorithm to select optimal data inputs, while the multiroute output model efficiently combines results from various target stations. Tested on air quality data observed over Beijing, the model demonstrated superior performance over traditional and state-of-the-art algorithms. Despite its success, the complexity of the model and its reliance on selected data sources suggest further refinement, such as developing more efficient internal structures and including a broader range of pollution sources, to enhance predictive accuracy.

In another study, S. Gunasekar and his colleagues [SG22] introduced the HALR hybrid model for air quality prediction, combining ARIMA, LSTM, and *Red Deer Optimization* (RDO). This hybrid approach addresses both linear and nonlinear patterns in air quality data, improving performance metrics, like accuracy, precision, and recall over conventional methods, such as KNN, SVM, and standard ARIMA models. The RDO algorithm fine-tunes the LSTM's weights and biases, boosting the model's predictive capabilities. Comparative analyses revealed that HALR outperforms other models with lower error rates and more robust performance metrics. Future research is set to explore hybrid optimization techniques to further minimize local minima issues and improve scalability for larger datasets.

While advanced LSTM-based models significantly enhance air quality prediction, several common limitations persist. These include high computational demands, sensitivity to parameter tuning, and challenges in managing the diverse and dynamic nature of environmental data. Therefore, future efforts must focus on optimizing these models, reducing their complexity, and enhancing their adaptability across different regions and data sources.

3.1.2 Motivations

Despite the notable progress in air quality predictions using LSTM-based models, the field faces persistent challenges and limitations [SKD23]. In particular, the accuracy of predictions hinges heavily on the quality and representativeness of sequential data, with incomplete or biased datasets, potentially compromising model performance. Additionally, while LSTM models excel in capturing temporal dependencies, they may encounter difficulties with abrupt data changes or outliers, necessitating further refinement for robust predictions. Finally, deploying such models at scale requires careful considerations for computational efficiency and online processing, especially in urban areas where a significant volume of data could be generated and collected.

In response to the challenges outlined above, we introduce our *Temporal LSTM forecasting model* (PMFORECAST), specifically designed to address the complexities of urban air quality prediction. Building on the strengths of standard LSTM architecture, PMFORECAST not only captures temporal dependencies effectively, but also integrates mechanisms to mitigate data quality issues, thereby enhancing prediction accuracy.

Our model's embedded temporal mechanisms contribute to robust and sustainable long-term predictions. A key feature of PMFORECAST is its use of locally available, cost-effective sensors from existing devices. This approach not only increases the model's adaptability but also enhances its accessibility.

By meticulously optimizing the PMFORECAST model, we aim to surpass traditional simulation methods, offering a resource-efficient alternative that reduces both time and energy consumption. This ultimately establishes a local real-time framework for air quality monitoring.

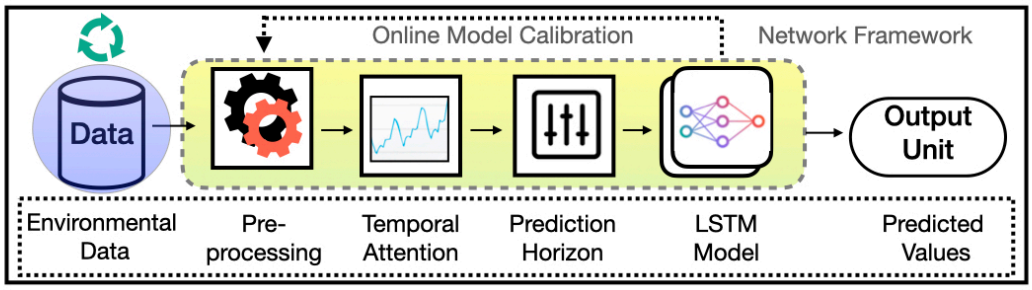


Figure 3.1: The comprehensive framework of PMFORECAST designed for air pollution prediction is outlined, comprising four key steps: data pre-processing, temporal attention to mitigate gradient disappearance, a flexible prediction horizon for dynamic future forecasting, and layers employing Long Short-Term Memory (LSTM)—the trainable component. Further details are provided in Section 2.1. The term 'Environmental data' pertains to data previously collected and utilized by the model for training purposes.

3.2 Methodology

As we embark on the pivotal task of predicting $PM_{2.5}$ concentrations for both immediate and future time frames in urban environments, the foundation of our approach is rooted in the strategic choice of a neural network architecture. Our conviction in the relevance of the LSTM model stems from its exceptional ability to discern and interpret temporal patterns, a crucial aspect in unraveling the intricate dynamics of air quality. By leveraging the strengths of the LSTM architecture, our model, PMFORECAST, delivers reliable predictions with a strong focus on accuracy.

Figure 3.1 depicts an overview of the PMFORECAST framework developed in this study. The proposed model consists of four primary phases: preprocessing, temporal attention, prediction horizon, and LSTM layers. The network is fueled by historical observations, and the outputs encompass the temporal dynamics of the predicted values. Observations are recorded by the sensor every 15 minutes. To focus on more predictable, regional pollution trends, we resample the data to one point per hour. While variations can occur within a 60-minute window due to localized events, such as nearby activity or short-term emissions, these are less relevant to our goal of forecasting broader air quality patterns. This approach allows us to concentrate on significant, region-wide pollution events rather than transient, unpredictable fluctuations. Additionally, the model can operate online by evaluating and retraining at specified intervals, contingent upon available data.

3.2.1 LSTM Model

Our goal centers on optimizing ML algorithms for superior performance, with a specific focus on deep learning techniques, particularly RNNs, which have demonstrated effectiveness in processing sequential data. However, traditional RNNs face challenges in long-term prediction tasks, prominently contending with issues such as gradient disappearance [Noh21]. This phenomenon occurs during the RNNs training process, when the loss function gradients concerning network parameters diminish significantly as they are back-propagated through time.

To address the challenge of long-term dependencies in time-series data, advanced models, like LSTMs incorporate mechanisms, such as gating, to improve the accuracy of long-term predictions—a critical component in our work on air quality forecasting. In an LSTM cell, three gates—input, forget, and output—regulate the flow of information, allowing the model to selectively retain or discard information, which helps mitigate issues like gradient disappearance during training [LZ16]. These gates are key to enabling the LSTM to capture both short- and long-term dependencies within the data.

An LSTM cell processes a sequence of data step by step. The forget gate decides what information should be discarded from the cell state, the input gate determines which new information should be stored in the cell state, and the output gate controls the value that will be passed on to the next time step. This architecture allows LSTMs to excel in sequential tasks, such as air quality forecasting, by capturing temporal patterns effectively [HS97]. By leveraging the LSTM model as our foundational approach, we aim to achieve optimal performance compared to other machine learning algorithms, including *Gated Recurrent Unit* (GRU), *Gaussian Process Regression* (GPR), *eXtreme Gradient Boosting* (XGBoost), and *AutoRegressive Integrated Moving Average* (ARIMA).

3.2.2 Temporal Dynamics Modeling

Afterward, the network structure needs to be optimized in terms of time, cost, and performance. To achieve this, we aim to leverage *Time-Focused Insight Generation*. Inherently considering temporal correlations of historical air pollutant data helps to improve performance.

Temporal Attention We proficiently capture the essential characteristics from historical environmental data through temporal windows whose duration is dynamically adjusted depending on how far ahead the prediction is being made.

Subsequently, we utilize the LSTM layer to extract temporal information from these mapped features. In time series, incorporating time-variant features is pivotal for capturing effective temporal dynamics. We systematically extract supplementary time-related features due to their strong influence on PM predictions, such as (i) weekdays versus weekends related to traffic and

industrial emissions and (ii) hours of a day related to hourly emission strength modification and the rise of the boundary layer height (i.e. the lower part of the atmosphere influenced by the Earth's surface) influencing the dilution of particles in the atmosphere over the course of the day.

Recognizing the significance of these temporal factors, we seamlessly integrate them as essential features in our model's input by merging them with the measurements. When optimizing temporal history for prediction, one question arises: How much of the historical environmental data should be considered?

An extensive sensitivity analysis in which different lag times were tested on a real-world dataset reveals that the lag time should be dynamically adjusted based on the prediction's time horizon to obtain the most accurate PM predictions. Our experimental results indicate that setting a$lag_{constant}$ at 3 hours consistently produces optimal outcomes for one-hour-ahead forecasting. This choice was derived through extensive experimentation, where we tested various lag times, training our model with different historical data windows $[1,3,6,9,12]$ hours, and evaluating their prediction accuracy using metrics like R^2 and RMSE. The 3-hour window consistently produced the lowest prediction errors, indicating optimal performance for capturing relevant temporal patterns. Further validation confirmed that adding more historical data did not significantly improve accuracy, and shorter windows led to higher errors, establishing the 3-hour window as the best balance for accurate predictions. The same experiments were conducted for extended prediction horizons. As the prediction time extends, the lag time incrementally increases, following the empirical relationship we derived, encapsulated in Formula 3.1. Specifically, for every 6-hour extension in the prediction horizon, the increment in the lag_t results in the incorporation of additional past observations, enhancing performance.

$$\mathrm{lag}_t = \mathrm{lag}_{\mathrm{constant}} + \mathrm{round}\left(\frac{\mathrm{pre}_t}{\mathrm{lag}_{\mathrm{rate}}}\right) \tag{3.1}$$

Equation 3.1 is derived from our experimental findings, which suggest a dynamic relationship between the prediction horizon and the optimal lag time. It is crucial to highlight that, in our experiments, pre_t denotes the prediction period and the constant lag_{rate} is set at 6, representing increments based on predictions made every 6 hours.

Prediction Horizon Strategies We formulate a strategy tailored to meet long-term prediction demands. Subsequently, our framework excels at forecasting air pollution intervals based on user preferences, specifically for the next few days with a time stamp interval of 1 hour. The mechanism dynamically updates the ground truth data, lag time observations, and the output unit according to the user's preferences. This process is visually represented as the *Prediction Horizon* in Figure 3.1. When the user modifies preferences, online updates reconfigure the

pre-processing and dynamics of the temporal attention mechanism to align with the new purpose. Then, the model is retrained.

3.2.3 Model Hyper-parameters

In our pursuit of creating an optimal configuration for the LSTM model, the fundamental aspect of our goal is to minimize hardware requirements and employ a lightweight model while maximizing performance. The number of hidden units and layers within a neural network are crucial hyper-parameters that significantly impact the model's capacity and complexity. We strongly emphasize achieving a delicate balance between the model's capacity and the potential risks of over-fitting or under-fitting. In this endeavor, we carefully considered the unique characteristics of the problem and dataset, ensuring that our model is not only efficient but also tailored to the specific challenges posed by the given context. Our investigation comprehensively assesses how varying numbers of hidden layers and units impact the model's performance. Conducting an exhaustive analysis, we explored multiple unit configurations within the range of $[32, 64, 128, 256]$, along with variations in the number of layers ranging from 1 to 5. Model evaluation was performed using R^2 and RMSE metrics for four sites in both train and test datasets. Ultimately, we determined that the optimal architecture for our LSTM model comprises 2 hidden layers with 128 units each, utilizing the *Relu* activation function. The model was trained using the mean squared error *Mean Square Error* (MSE) for the loss function, chosen for its effectiveness in regression tasks.

During the training phase, it is crucial to specify hyper-parameters that significantly influence the performance of deep learning models. Firstly, the *learning rate*, a parameter that determines the size of the steps taken during the model optimization process, was set at 10^{-3}, after evaluating values within the range $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$. Next is the *batch size*, representing the number of data samples processed in one iteration during model training. A carefully chosen batch size of 48 was implemented, indicating that the model processed 48 (equivalent to 2×24 hours) training examples per iteration. Lastly, the *epoch* parameter, denoting the number of complete passes through the entire dataset during model training, was set to an extensive value of 200. This choice allowed the model to iterate through the entire training datasets 200 times, capturing intricate patterns and enhancing overall performance.

To avoid overfitting and promote model generalization, the *Early Stop* technique was used. This involved monitoring the model's performance on a validation set during the training process and interrupting training once the performance ceased to improve or started to degrade, effectively preventing unnecessary further training.

3.2.4 Dynamic Datasets & Online Model Calibration

Our model is designed for monthly updates, as we typically observe minimal changes within a one-month timeframe. Embracing a dynamic approach, the model undergoes regular fine-tuning of its hyper-parameters based on processed information. It continuously assesses performance metrics, such as accuracy (R^2) and root mean square error (RMSE), selectively incorporating updates when improvements are detected. This iterative process ensures efficient training over the specified timeframe, maintaining the model's currency, and optimizing the latest data trends.

Furthermore, any changes in user preferences can be applied online, allowing the model to be quickly recalibrated to better suit user needs. The adaptability of this approach allows our model to respond effectively to evolving patterns and progressively improve its performance over time.

3.3 Sensor Deployment and Data Acquisition

This section provides detailed information on the observations utilized to feed our model, specifically focusing on (QAMELEO network in Dijon). It covers the data acquisition process, detailing the various sources and methods used to collect air quality data, and the subsequent pre-processing steps required to prepare the data for model input. These observations form the foundation of our analysis, ensuring that the model is equipped with accurate and reliable data to enhance its predictive capabilities.

3.3.1 The QAMELEO Network

QAMELEO is a *low-cost* air quality micro-station developed by two research teams in the University of Burgundy and *Institut de Recherches pour le Développement* (IRD) [MNS+23]. QAMELEO microstations measure the mass concentrations within PM_1, $PM_{2.5}$, PM_{10} fractions along with meteorological variables, such as temperature and relative humidity. The measurements are consistently available every 15 minutes, aligning with the time-step of the stations of the official air quality monitoring association (AASQA : Association agréée de surveillance de la qualité de l'air) and operated by regional governments in the Dijon Metropolis.

QAMELEO micro-stations have supported tests in the laboratory and outdoors, in the frame of a national evaluation exercise leaded by the LCSQA (Central Laboratory of Air Quality Surveillance) in July 2018. This proficiency testing of micro-sensors systems, referred to as the *EAμC* field campaign, enabled to compare the QameleO micro-station to 15 other micro-sensors and to the BAM1020 reference analyser, measuring the $PM_{2.5}$ fraction, for 2 entire weeks and at the same location/station. These tests established that the QAMELEO micro-station can satisfactorily reproduce the temporal dynamics of the PM mass concentrations [CRS18, RCH+18].

In particular, for $PM_{2.5}$, the correlation coefficient between the micro-station and a referenced station is $+0.73$, which is a significant score at a 99% level according to the Bravais-Pearson statistical test, and a mean bias of $-2.71 \mu g/m^3$.

In Dijon Metropolis, the POPSU (*Plateforme d'Observation et de Stratégies Urbaines*) program has been a real opportunity to deploy the QAMELEO network in a real urban environment (cf. Figure 3.2). The QAMELEO microstations were implemented like meteorological stations, all under the same conditions: at 3 meters high, with a similar Sun Exposition. There are 12 QAMELEO micro-stations implemented in Dijon Metropolis. Of the 12 micro-stations, four of them cover a complete year (from November 2020 to October 2021) of measurements as the network has been deployed in progressive phases. These four specific stations are located within the city in *Port du Canal,* Hoche, *Carnot, and* Janin, representative of diversified urban condi-

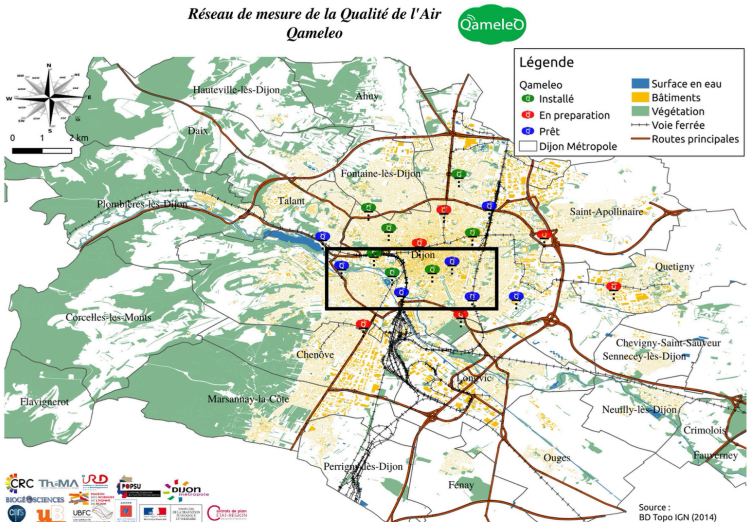


Figure 3.2: Locations of Air Pollution Monitoring Micro-Stations in Dijon. The blue circles in the black box correspond to the four QAMELEO stations used in this study [MNS+23]

tions (traffic: *Carnot and* Hoche, urban background: *Port du Canal and* Janin). PMFORECAST was used on each station independently to test its efficiency.

3.3.2 Data Preprocessing

The data preprocessing stage is crucial in improving the quality and suitability of the data set for comprehensive analysis. It involves a meticulous process of cleaning, transformation, and organization to ensure data accuracy and consistency while eliminating errors. QAMELEO dataset is validated and corrected for the concentrations of the PM mass according to the Isolation Forest (IF) method developed by the University of Burgundy. The Isolation Forest algorithm is an unsupervised machine learning technique specifically designed for anomaly detection. It works by isolating observations through a random selection of features and split values, efficiently identifying anomalies due to their distinct nature. [MNS+23] employed this method to identify and correct anomalous concentration values of the mass of PM, ensuring the precision and reliability of the data set that we used in our study. Addressing missing values is a crucial step before diving into data analysis. Despite the QAMELEO microstations offering a relatively consistent dataset for air quality assessment, with an average of approximately %5 missing values over a year time series, ML methods require a dataset without gaps. To meet this criterion, we applied a 12-hour moving average.

In the final stage, where a few minor missing values persisted around 0.8% to 0.9% percent, we opted for forward filling, replacing each missing value with the most recent observed value in the dataset. Achieving uniformity in the dimension values is crucial for meaningful analysis. To ensure this, we employed *Min-Max Normalization*, a technique that scales the dimension

values to a range between 0 and 1. This normalization process contributes to equitable data representation, a fundamental aspect of robust analysis. The final versions of the datasets are prepared for four sites: Station 1 (Canal), Station 2 (Hoche), Station 3 (Carnot), and Station 4 (Janin). We include measurements spanning 9.5 months in the training sets and 2.5 months in the testing sets for all stations.

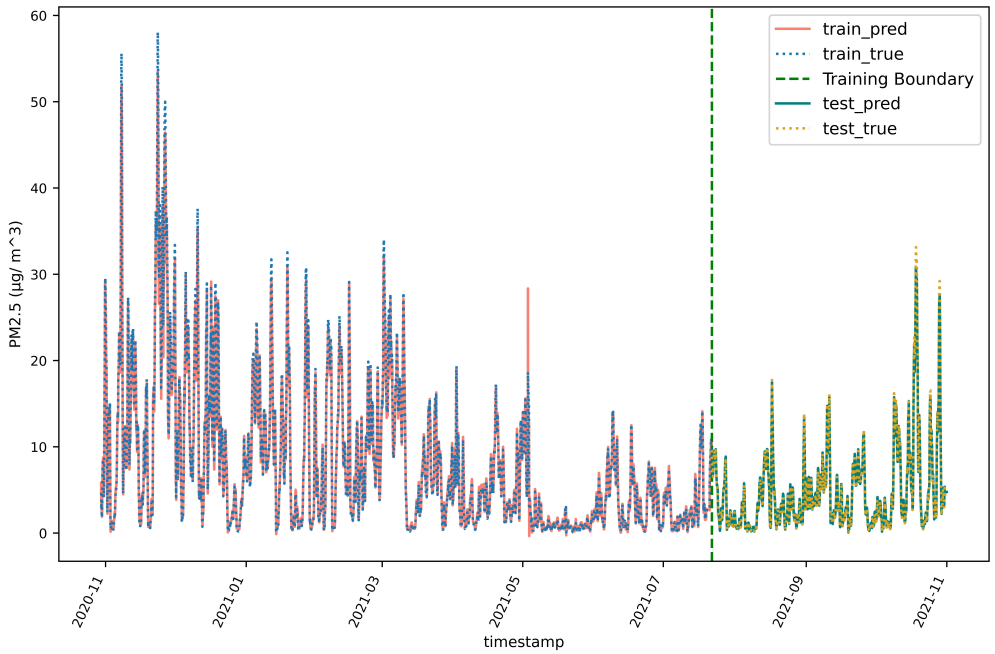


Figure 3.3: Hourly temporal prediction of $PM_{2.5}$ levels over time for Canal site. The dotted lines correspond to the observed values and are representative of the true values during the training (blue) and prediction (golden) periods. The solid lines correspond to the $PM_{2.5}$ predicted during the training (salmon) and the prediction (green) periods. The dashed vertical green line indicates the division between the training and test datasets.

3.4 Results

A thorough series of assessments were conducted to comprehensively present our results. Beginning with an evaluation of hourly prediction precision for the Canal site, we then delve into an in-depth analysis of extending time-frame predictions across all sites. Following this, we undertake a comparative study involving various popular ML algorithms for time-series forecasting. Additionally, we assess the feasibility of multi-task prediction. Finally, we examine the time-consuming aspects across each phase of our framework.

3.4.1 Precision of Air Pollution Forecasting

One of the main objectives of this research is to achieve high precision in the prediction of air pollution, and in particularly $PM_{2.5}$. The experimental results displayed in Figure 3.3 provide a visual representation of $PM_{2.5}$ readings over time for the Canal site, illustrating both predicted and ground truth values from the train dataset and test dataset. To ensure the model's robustness, we conducted training evaluations to confirm that the model performs well not only on the train dataset but also on the test dataset. In a specific time-frame, Figure 3.4 showcases the ground truth and predicted values for the test set on 24^{th} and 25^{th} July 2021.

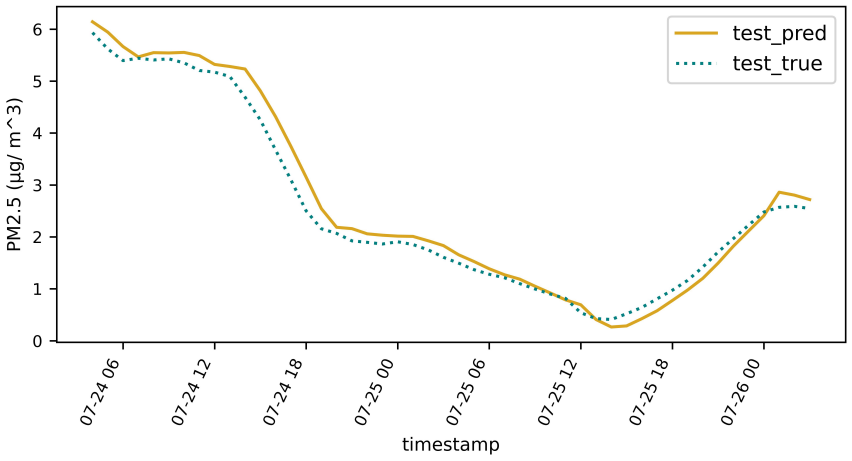


Figure 3.4: Hourly temporal prediction of $PM_{2.5}$ levels over time over the Canal site, forecasting 2-day predictions for July 24th (Saturday) and July 25th (Sunday), 2021. The golden solid line represents the predicted values and the dotted green line represents the truth values for the test set.

Table 3.1: Evaluation metrics (RMSE, MAE, MSE, R^2, WMAPE) for prediction results during the training period for the 4 QAMELEO stations (Canal, Hoche, Carnot, Janin) focusing on 1-hour predictions with a history of 3 hours. Bold values indicate the best performance across all sites.

Train-set	**RMSE** ($\mu g/m^3$)	**MAE** ($\mu g/m^3$)	**MSE** ($\mu g/m^3$)	R^2 (%)	**WMAPE** (%)
Canal	0.566	0.352	0.320	**99.5**	**0.043**
Hoche	0.768	0.428	0.590	98.6	0.078
Carnot	**0.438**	**0.262**	**0.192**	99.3	0.054
Janin	0.626	0.414	0.392	99.4	0.047

Following the assessment of model performance presented in Tables 3.1 and 3.2, it is essential to delve into the specific evaluation metrics employed. The assessment includes not only RMSE and R^2, but also incorporates other key performance metrics, such as MSE, MAE, and WMAPE. RMSE quantifies the average magnitude of prediction errors, providing a comprehensive measure of model accuracy. MSE offers a similar insight without considering the square root, emphasizing larger errors. MAE represents the average absolute difference between predicted and actual values, offering a robust measure of model precision. R^2 gauges the share of correctly predicted instances, providing a holistic view of model effectiveness. WMAPE, calculated as the weighted average of absolute percentage errors, offers a nuanced perspective by considering the significance of errors across different prediction scenarios. These metrics underscore the model's robust performance across diverse evaluation criteria for training and test datasets.

Table 3.2: Evaluation metrics (RMSE, MAE,MSE,R^2, WMAPE) for prediction results during the testing period for the 4 QAMELEO stations (Canal, Hoche, Carnot, Janin) focusing on 1-hour predictions with a history of 3 hours. Bold values indicate the best performance across all sites.

Test-set	**RMSE** ($\mu g/m^3$)	**MAE** ($\mu g/m^3$)	**MSE** ($\mu g/m^3$)	R^2 (%)	**WMAPE** (%)
Canal	**0.353**	0.238	**0.125**	**99.4**	**0.049**
Hoche	0.393	0.241	0.154	98.2	0.106
Carnot	0.357	0.200	0.164	98.9	0.071
Janin	0.444	**0.179**	0.197	97.4	0.061

The predictions show remarkable precision, achieving an impressive R^2 of approximately 100% and substantial RMSE ranging from 0.36 to 0.77 $\mu g/m^3$ in the train and test sets at all stations, respectively. This signifies a robust correlation between predicted and measured $PM_{2.5}$, highlighting the model's exceptional predictive capabilities for the subsequent hour. The combination of high R^2 and low RMSE underscores the reliability and precision of the model in capturing and forecasting target values. Specifically, we note metrics with values less than $0.43\mu g/m^3$, $0.6\mu g/m^3$, and 11% for MAE, MSE, and WMAPE, respectively. In the context of our study, it is essential to acknowledge that our experimental setup involves small test datasets and a model of relative simplicity. In light of these considerations, it is observed that RMSE exhibits a lower value in the test dataset in comparison to the training dataset, which is a general trend observed in machine learning experiments.

While all results demonstrate significance, the notable prominence of the Canal station in the test set and the Carnot station in the train set as the most favorable matches suggesting the best alignment between observed and predicted values at these specific locations (Tables 3.1 and 3.2). Bold values in these tables highlight the highest performance across all sites. Despite local attributes such as unique environment leading to different emission sources as well as their diurnal variations, geographical, or meteorological conditions inherent to each station, the model is surprisingly performing well across all sites. This robust adaptability underscores the model's effectiveness across diverse environmental conditions within this city. More metropoles need to be tested to confirm this behavior with probably more contrasted typology (rural vs urban).

3.4.2 Extended Time-frame Prediction

In Figure 3.5, we assess the accuracy of $PM_{2.5}$ predictions in the train and test sets across datasets from the four stations, ranging from 1 to 72 hours into the future. Figure 3.5a depicts the computed score, representing the average over the prediction period. Leveraging hourly predictions with adaptable horizons for the near future, we conducted an examination to evaluate performance across various timeframes—specifically $[1, 6, 12, 24, 48, 72]$ hours for all stations.

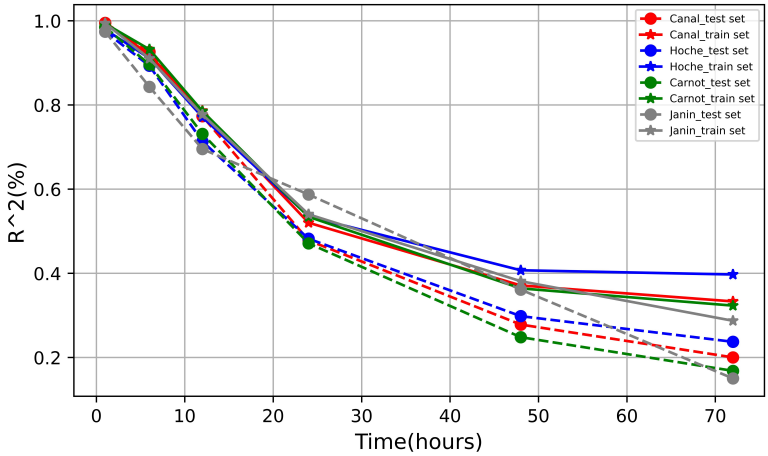
This resulted in $[1, 6, 12, 24, 48, 72]$ hours of prediction values for each time horizon. To provide an example, when forecasting $PM_{2.5}$ levels for the next 12 hours, the model generates a predicted value for each hour within this period, amounting to 12 values for this configuration. The R^2 score for the entire 12-hour prediction horizon is subsequently calculated by averaging these 12 R^2 values.

As the prediction period extends, a noticeable decrease in accuracy for each individual hour is observed, as depicted in Figure 3.5. Another noteworthy phenomenon emerges with the extension of the time horizon, where the performance decreases for a specific time compared to a shorter future prediction. For instance, with a 24-hour forecasting horizon, the accuracy for the first hour drops to approximately 91.5%, while a 1-hour forecasting horizon achieves a higher accuracy of approximately 99%. In particular, for the 48-hour horizon, R^2 drops to 0.4 for all sites in both the train and test periods, indicating a limitation in the model's ability to predict PM levels beyond 36 hours.

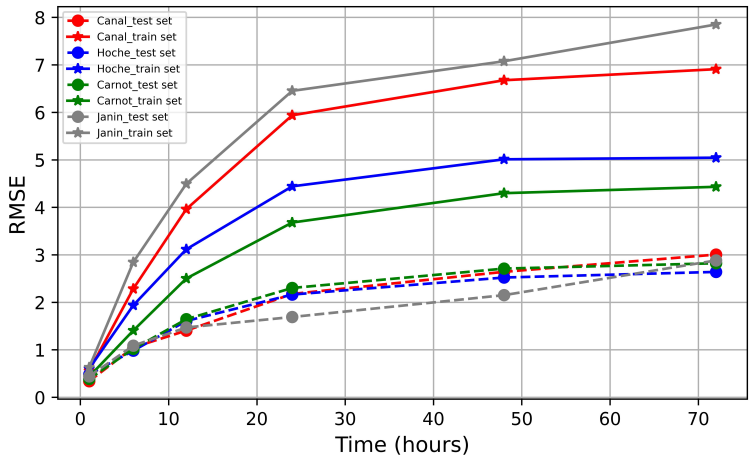
This observed behavior aligns with the common phenomena noted in both physical and numerical models [ZLSS23]. While the results showcase the model's ability to generalize and provide reliable forecasts for the near future, challenges arise when forecasting for more extended timeframes.

3.4.3 Method Comparison Study

Our study considered a diverse set of widely-recognized time series data forecasting algorithms, including *Gaussian Process Regression* (GPR), *Gated Recurrent Units* (GRU), *XGBoost*, *ARIMA*, and Standard LSTM. We evaluate these ML algorithms to forecast $PM_{2.5}$ levels over 1- and 12-hour periods, carefully examining their performance over time across four different sites. The consistent findings across these sites lead to a comprehensive analysis presented in Table 3.3, showcasing the effectiveness of each algorithm in meeting our forecasting objectives. GRU demonstrates performance very close to PMFORECAST, attributed to their similar architectures and shared algorithms in RNN models. However, PMFORECAST consistently outperforms GRU, displaying superior results. ARIMA excels in the short term, providing accurate predictions, but its computational demands increase for longer forecasting horizons, leading to less efficient performance. XGBoost, despite its stability and rapid training times, falls short compared to RNN-based algorithms. GPR achieves high rankings on training data but delivers less favorable outcomes on the test set. While the Standard LSTM demonstrates good performance, PMFORECAST consistently outperforms it, with distinctions becoming more pronounced for extended forecasting horizons beyond 12 hours. Following comprehensive experimentation and evaluation, the PMFORECAST model emerges as the most effective choice. PMFORECAST exhibits better performance achieving shallow RMSE values of 0.357 for 1-hour and 1.635 for 12-hour forecasts on the Carnot site test dataset. The temporal mechanism



(a)



(b)

Figure 3.5: Performance Evaluation of Long-Term $PM_{2.5}$ Forecasting Across Multiple Sites: (a) Accuracy Assessed by R^2 % metrics, and (b) Root Mean Squared Error $RMSE\mu g/m^3$. The solid lines with stars denote the performance on the training sets, while the dashed lines represent the performance on the test sets. Each of the four stations is distinguished by a unique color: Canal (red), Hoche (blue), Carnot (green), and Janin (grey).

embedded in our PMFORECAST framework endows PMFORECAST with superior predictive capabilities among the evaluated algorithms, establishing it as a pivotal asset in our quest for precise PM predictions.

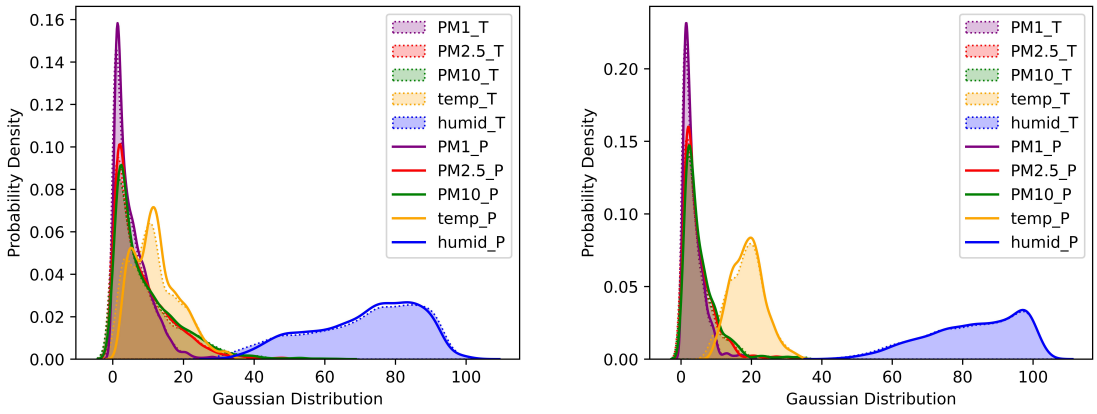
Table 3.3: Assessing our Model's Predictive Performance at the Carnot Site Using Diverse Machine Learning Algorithms on the Test Dataset. **Bold values** indicate the best performance across all methods.

Forecasting	**1-Hour**			**12-Hours**		
Methods	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	R^2 (%)	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	R^2 (%)
GRU	0.424	0.240	98.2	1.648	1.022	70.0
GPR	0.615	0.351	96.2	3.182	1.495	11.0
XGBoost	0.437	0.237	98.1	1.874	1.083	65.1
ARIMA (VARMAX)	0.587	0.483	96.5	2.425	1.675	41.5
LSTM	0.422	0.256	98.2	1.731	1.019	70.2
PMFORECAST	**0.357**	**0.164**	**98.9**	**1.635**	**0.954**	**73.7**

3.4.4 Multi-Tasks Model

Multi-task prediction offers efficiency, optimal resource utilization, and enhanced decision support, establishing itself as a valuable approach to air pollution prediction. Our research involved a comprehensive multi-task forecasting strategy, concurrently addressing the prediction of 3 major pollutants—PM_1, $PM_{2.5}$, and PM_{10}—as well as temperature and humidity simultaneously. Figure 3.6 visually presents these correlations, illustrating the strong alignment between measured and predicted values for the next hour across the three PM fractions at the Canal station. While the single-task prediction model slightly outperformed the multi-task approach with a correlation of about 99% for $PM_{2.5}$ in one-hour prediction, it is important to highlight that the multi-task strategy using the PMFORECAST approach demonstrated remarkable efficiency in capturing the essence of the PM fraction's behavior with an evaluation metric for R^2 around 98% for all fractions. Equally impressive correlations were noted for other features, closely aligning with this value. Moreover, this robust performance was not limited to a specific dataset. Indeed, the correlation results remained consistent across the three additional sites further affirming the effectiveness of our multi-task forecasting methodology.

The overall efficiency and comprehensive insights provided by the multi-task approach, particularly with the PMFORECAST method, underscore its value and efficacy in capturing the complex behaviors of various PM fractions. This intriguing finding emphasizes the potential of the multi-task strategy as an effective alternative, providing comparable predictive accuracy while incorporating multiple parameters in the forecasting process.



(a) (b)

Figure 3.6: Performance assessment through Gaussian distribution for multi-tasking at the Canal Site with varied meteorological data. (a) Examination of the correlation between observed and predicted values for the training set. (b) Investigation of the correlation between observed and predicted values for the test set. The truth and predicted values are illustrated with dotted and solid lines, featuring "T" and "P" in the labels, respectively. The colors represent the five measurements in our data: PM_1 (purple), $PM_{2.5}$ (red), PM_{10} (green), temperature (orange), and humidity (blue).

3.4.5 Time Overhead for Model Training & Inferences

The PMFORECAST model displays variable time consumption across its key steps, as detailed in Table 3.4, utilizing an Apple M1 chip with 16GB of memory. The data pre-processing stage, involving one sample consisting of 5 measurements and 2 temporal features (day of the week and hour of the day), exhibits swift efficiency. Tasks include converting the timestamp from 15 seconds to hourly datasets and handling missing values using a moving average algorithm, achieving a latency of 61 seconds for the entire dataset in one station.

In contrast, the temporal mechanism, which operates on the entire dataset, introduces a favorable latency of less than 1 second, reflecting the simplicity associated with handling temporal aspects. The training phase for one epoch requires 5.0 milliseconds, emphasizing the computational demands involved in optimizing the model parameters. The complete model training process, from raw data to a trained model, takes 250 seconds, with a configuration of 200 epochs and an *early stopping* mechanism set for a patience of 30 epochs.

Recalibration configuration times vary depending on the horizon time. For instance, for 12-hour prediction points, the latency of a fully trained model is 283 seconds. Subsequently, predicting air pollution levels for a one-hour horizon demonstrates remarkable efficiency, with a latency of 375 milliseconds for 2400 samples. For the same number of samples but a longer horizon, such as 12 hours, the latency increases to 481 milliseconds.

Table 3.4: Time latencies for each step of the procedure in PMFORECAST.

Step	Latency (s)
Data pre-processing (8810 samples)	61
Temporal Mechanism (total data)	0.075
Training (1 epoch)	0.005
Train Full Model (full model from scratch to 1 point prediction for 1 station)	212
Online Model Recalibration (from 1 to 12 time points)	253
Prediction (1 time point for 2400 samples)	0.375
Inference Duration per Horizon (12 time points for 2400 samples)	0.481

These temporal benchmarks offer insights into the computational performance of the PMFORECAST model, essential to assess its feasibility in online applications.

3.4.6 Long-Term forecasting Using 20 Years of Satellite Data

We tested the robustness and reliability of our model by utilizing a 20-year dataset of air quality measurements generated using the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis model for the Dijon region. The CAMS model offers global atmospheric composition data with a spatial resolution of 0.75° x 0.75°, where each grid cell corresponds to approximately 83 km x 83 km, thus providing comprehensive data for simulating air quality conditions. Spanning from October 1, 2001, to March 31, 2021, this dataset served as a solid foundation for evaluating the model's long-term forecasting capabilities, enabling hourly predictions.

Following the approach used in PMForecast, we selected three particulate matter (PM) levels—PM_1, $PM_{2.5}$, and PM_{10} —along with humidity and temperature as the primary input features for our model. After preprocessing and addressing missing values, the data was split into 80% for training and 20% for testing, ensuring that our model was rigorously assessed on unseen data while learning from a substantial portion of the available information. The models were trained for 50 epochs, with a batch size of 48, a learning rate of 0.0001, and mean absolute error (MAE) as the loss function.

We utilized our pre-trained PMFORECAST model architecture, unfreezing only the final dense layer. We hypothesize that this lightweight approach will perform well even with large datasets. Table 3.5 presents the results of the long-term prediction metrics, highlighting the model's performance across various time horizons:

These metrics demonstrate the model's robustness in predicting long-term air quality trends based on historical satellite data, with strong performance in LSTM models and low error metrics across both training and testing datasets. Although the testing set shows slightly higher error rates, which is expected in long-term predictions, the model still generalizes well beyond the training data. Performance naturally decreases over time, but there remains a good correlation

Table 3.5: Metrics Evaluation for 1 to 48 Hours Forecasting in 20-year dataset

Hour(s)	**Train RMSE**	**Train MAE**	**Train MSE**	**Train WMAPE (%)**	**Train R^2**
1 Hour	1.547	0.932	2.39	10.5	0.921
6 Hours	2.955	1.934	8.265	21.8	0.753
12 Hours	3.674	2.464	13.503	27.8	0.627
24 Hours	4.265	2.940	18.191	33.2	0.497
48 Hours	5.648	3.981	31.910	45.0	0.231
Hour(s)	**Test RMSE**	**Test MAE**	**Test MSE**	**Test WMAPE (%)**	**Test R^2**
1 Hour	1.592	0.961	2.536	10.7	0.920
6 Hours	3.088	1.991	9.540	22.3	0.747
12 Hours	3.812	2.539	14.532	28.5	0.603
24 Hours	4.433	3.011	19.652	33.8	0.452
48 Hours	5.776	4.089	33.364	45.9	0.202

between predicted and actual values. Figure 3.7 illustrates the predictions on both the training and test sets for upcoming data points.

Figure 3.8 demonstrates near-perfect convergence for a 24-hour prediction horizon, providing further evidence of the stability and robustness of the model we designed. This result supports the claim that our model can be effectively deployed on any device for training with pollution data, ensuring reliable performance across various environments.

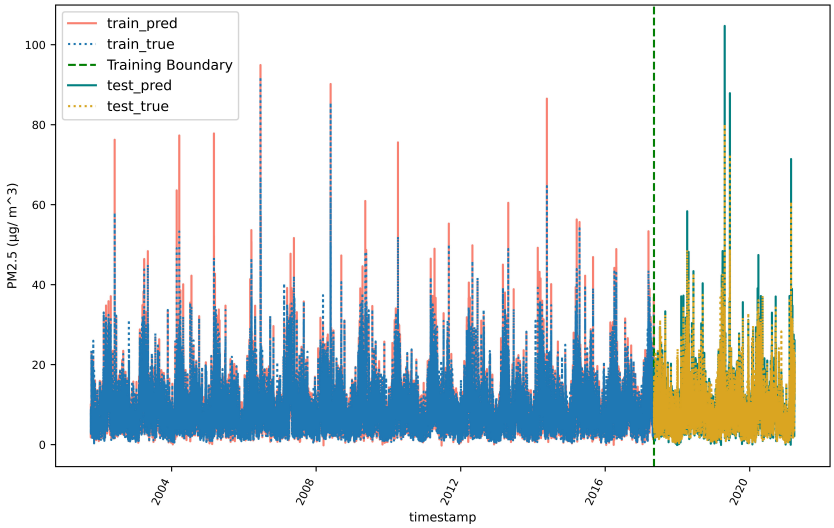


Figure 3.7: Hourly temporal prediction of $PM_{2.5}$ levels over time for 20 years (2001 to 2021).The solid lines correspond to the $PM_{2.5}$ predicted during the training (salmon) and the prediction (teal) periods. The dashed vertical green line indicates the division between the training and test datasets. The dotted lines correspond to the $PM_{2.5}$ truth values during the training (blue) and the prediction (golden) periods.

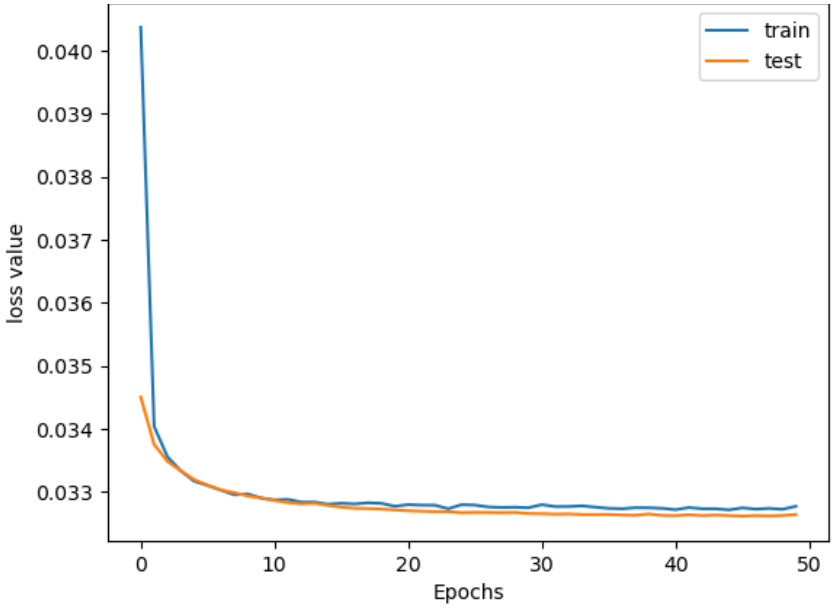


Figure 3.8: Loss Convergence for 24-Hour Forecast: Training vs. Testing Phases

3.5 Discussion

This chapter focuses on developing an accurate air pollution prediction model, with a primary emphasis on $PM_{2.5}$. We evaluated a broad range of forecasting algorithms, including GPR, GRU, XGBoost, ARIMA, and Standard LSTM. Among these, the PMFORECAST model—an advanced iteration of LSTM—emerged as the top performer, achieving remarkable R^2 rates of 98.9% for 1-hour forecasts and 73.7% for 12-hour forecasts. Evaluations conducted across various sites and forecasting horizons consistently demonstrated PMFORECAST 's superior predictive capabilities. Extensive testing and validation further confirmed PMFORECAST 's high accuracy in long-term forecasting, surpassing existing models and advancing the field of sustainable and precise air quality prediction.

In the realm of multi-task prediction, which includes forecasting PM_1, $PM_{2.5}$, PM_{10}, temperature, and humidity, the single-task model slightly outperformed in $PM_{2.5}$ prediction. However, the PMFORECAST multi-task approach stands out for its remarkable efficiency, achieving high correlations of over 98% for all PM fractions. This efficiency is particularly advantageous for dynamic applications, as evidenced by the varying time consumption metrics. PMFORECAST excels in computational efficiency during data pre-processing and provides low-latency predictions, reinforcing its potential for time-sensitive scenarios.

To further evaluate the robustness of our proposed model, we tested our lightweight pre-trained model on a separate dataset comprising 20 years of hourly data. The model demonstrated strong performance, affirming its adaptability and effectiveness across different datasets. It achieved 92% accuracy in both test and training sets for upcoming data points, with superior low error rates—RMSE of 5.776 and MAE of 4.089—in 48-hour horizon predictions.

In conclusion, the PMFORECAST model excels as a robust and versatile solution for accurate particulate matter prediction and offers significant efficiency gains. Its implications for online monitoring and decision-making are highlighted by its superior performance and computational efficiency with low latency, enhancing temporal attention. This makes it a valuable tool for air pollution forecasting applications.

Looking ahead, we plan to incorporate spatial characteristics of air pollution data into the model and implement regular *in situ* retraining strategies to ensure the model remains updated with the latest data, maintaining its relevance. Additionally, our focus will be on developing an integrated spatio-temporal model that considers the interplay of diverse datasets, aiming for a more comprehensive understanding of air quality patterns.

Chapter 4

GRAPH TEMPORAL LSTM

4.1 Introduction

As previously discussed, understanding air quality prediction requires more than just a focus on temporal factors; spatial considerations are equally crucial. Air pollution is not merely a matter of when concentrations peak but also where and how they spread. The spatial distribution of pollutants and their interdependencies across different regions significantly impact air quality patterns. To address these complexities, researchers explore a variety of data sources, including satellite imagery, ground-based sensors, and meteorological data. Each of these sources contributes unique insights: satellite imagery offers broad, real-time observations, ground-based sensors provide detailed, localized measurements, and meteorological data helps explain how factors like wind and humidity influence pollutant dispersion.

Incorporating spatial and temporal considerations into air quality prediction models enhances their accuracy and effectiveness. By integrating diverse data types and employing various spatial representation techniques, researchers can develop more comprehensive systems that better reflect the complex dynamics of air pollution.

Spatial representation in air quality models can be approached in several ways. Grid maps divide the area of interest into a network of cells, each representing a uniform spatial unit. This method facilitates detailed spatial analysis by enabling the modeling of pollutant concentrations within each cell. Grid maps are particularly useful for numerical simulations, providing a structured approach to monitoring and predicting air quality across a defined area. However, this method can be limited by its rigid structure, which may not fully capture the irregularities of real-world spatial interactions.

In contrast, graph maps offer a more flexible representation by modeling the area as a network of nodes and edges. In this approach, nodes represent specific locations, such as monitoring stations or key urban areas, while edges denote the connections or interactions

between these locations. Graph maps are well-suited to capturing complex spatial relationships, such as the movement of pollutants between different regions. They are particularly effective for modeling dynamic systems with irregular spatial structures, reflecting the ways in which pollutants travel along roads, through neighborhoods, and across natural barriers.

4.1.1 Literature Review

Numerous studies have explored both grid map and graph map strategies for air quality forecasting. For instance, recent research by [PRGDS+24] introduces the *Spatio-Temporal Air Quality* Forecaster (ST-AQF), an advanced AI framework employing *Convolutional Long Short-Term Memory* (ConvLSTM) networks. This model enhances the forecasting of pollutant concentrations over multiple time horizons by integrating data from diverse sources and adapting to sensor failures. The ST-AQF framework demonstrates significant improvements in prediction accuracy compared to traditional and state-of-the-art models, particularly in handling multiple pollutants simultaneously. Despite its strengths, the model requires further sensitivity analysis, parameter tuning, and exploration of alternative architectures and imputation methods. Future research aims to increase scalability, integrate additional data sources, and apply the framework to broader environmental monitoring efforts.

Another recent study by Jiaxuan Zhang *et al.* presents a CNN-LSTM model designed to boost air quality prediction accuracy by combining CNN's feature extraction capabilities with LSTM's sequential learning. Applied to Beijing's *Air Quality Index* (AQI), this model outperforms traditional approaches, like ARMA, SARIMA, RNN, and GRU, significantly reducing prediction errors and enhancing accuracy. The CNN-LSTM model effectively captures both spatial and temporal features, addressing the nonlinearity in AQI data. However, it faces challenges in adapting to sudden changes in AQI trends, necessitating further refinements for improved real-time responsiveness [ZL22].

On the graph map side, Hongye Zhou *et al.* [ZZDL21] propose the *Dynamic Directed Spatio-Temporal Graph Convolution Network* (DD-STGCN), which integrates domain knowledge of dynamic wind fields to improve $PM_{2.5}$ concentration forecasting. The DD-STGCN combines a directed graph time-series with wind-field diffusion distances, effectively capturing spatial and temporal dependencies between monitoring stations. Experimental results indicate that DD-STGCN outperforms traditional models such as LSTM, GC-LSTM, and STGCN in prediction accuracy and spatial interpretability, particularly under high-wind conditions. Nevertheless, the model could benefit from incorporating trends and periodicities in $PM_{2.5}$ concentrations to enhance long-term forecasting.

Another notable study introduces the *Spatially Attentive Cluster-based Graph Neural Network* (SA-GNN) for short-term $PM_{2.5}$ concentration forecasting, focusing on Delhi, India. The SA-GNN model utilizes graph neural networks to explore spatial relationships between

monitoring stations and integrates meteorological variables like wind speed and direction. By employing a clustering-based method and a graph attention network (GAT), the SA-GNN model demonstrates significant improvements over baseline models, achieving an R^2 value of 0.75 and reducing RMSE and MAE to 25.13 and 21.28 μg/m^3, respectively. While effective in forecasting high pollution episodes, the model's static graph clustering approach could be further enhanced by exploring dynamic clustering methods for better adaptability and accuracy [MT23].

The latest advancement in this field is the *Spatiotemporal Graph Convolutional Recurrent Neural Network* (Spatiotemporal GCRNN)[Le23]. This model integrates GCNs with RNNs to improve the handling of spatial relationships and temporal learning. Despite being significantly smaller than the ConvLSTM model[LBC20], the Spatiotemporal GCRNN offers superior performance in both short-term and medium to long-term forecasts. However, it remains highly reliant on diverse data sources for optimal performance.

4.1.2 Motivations

Previous researches in air pollution prediction has made significant strides, yet there are notable limitations in current approaches. Temporal methods, such as LSTM networks, are effective at capturing sequential patterns but often struggle with the complexity of integrating spatial dependencies. These methods can miss fine-grained spatial variations in pollutant concentrations, which are crucial for accurate local predictions.

On the other hand, spatio-temporal models, while advanced, can be complex and computationally intensive. Techniques that combine spatial and temporal processing often require multiscale processing and intricate algorithms to manage the dynamic interactions between spatial and temporal factors. This complexity can limit their scalability and practical application in real-time forecasting.

To address these challenges, this research proposes a novel hybrid model that integrates GCNs and LSTMs networks. This model aims to leverage the spatial strength of GCNs and the temporal capabilities of LSTMs to enhance both spatial and temporal prediction accuracy. By focusing on computational efficiency and scalability, the model utilizes cost-effective on-ground sensors for measuring PMs and other environmental variables.

We apply a graph map technique to forecast pollutant levels across urban zones, with a specific focus on Dijon, France. The proposed hybrid model combines GCNs and LSTMs within a temporal attention framework, building on the previous chapter in LSTM-based temporal modeling [RCAM$^+$24]. This approach is designed to improve the localization and accuracy of $PM_{2.5}$ predictions, addressing both spatial and temporal limitations effectively.

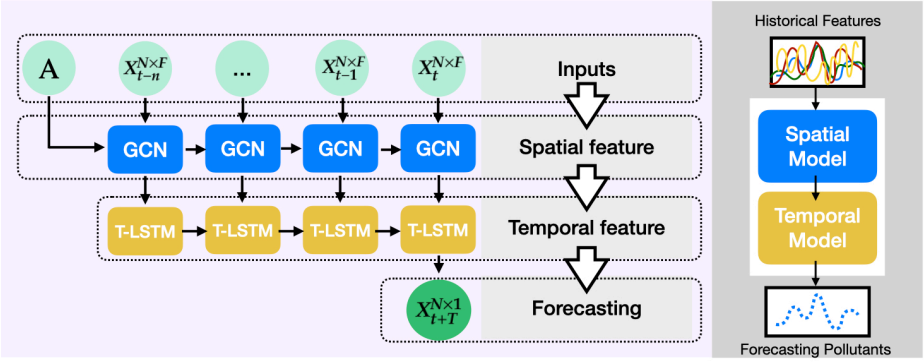


Figure 4.1: Overview of GT-LSTM comprising Spatial and Temporal Models. GCN and T-LSTM, respectively, are shown as the main components

4.2 Methodology

The primary objective of this research is to forecast air pollution levels within a specific time frame by using historical environmental data from various areas within a city. Specifically, our focus is on predicting $PM_{2.5}$ pollutant levels for different areas by analyzing characteristics of spatio-temporal datasets in urban regions.

The problem statement is defined using a weighted graph $G = (N, E)$ to represent the topological structure of the city. Each measurement station is treated as a node, where N represents the set of station nodes $N = [N_1, N_2, ..., N_n]$, n is the total number of nodes. The edge set, denoted as E, defines connections between nodes, illustrating how one node is linked to another. We represent this set of edges with a special adjacency matrix. The adjacency matrix A, denoted as $A \in R^{n\times n}$, visually represents the connections between nodes based on the correlation coefficients obtained from the topology of the city. Pearson's algorithm [KR24] is employed to compute the standard correlation coefficient for each measurement between node pairs, constructing the adjacency matrix A. Equation 3.1 demonstrates Pearson's algorithm, where X and Y represent the measurements of the nodes, Cov signifies the covariance between the measurements, and σ represents the standard deviation of each node.

$$P_{X,Y} = \frac{Cov(X,Y)}{\sigma_X . \sigma_Y} \tag{4.1}$$

A feature matrix, denoted as $\mathbf{X}$, represents air pollution information within the network. This matrix has dimensions $N \times F \times t$, where N is the number of nodes (monitoring stations), F is the number of features (e.g., pollutant concentrations, meteorological data), and t is the time

step. By structuring the data in this way, we can effectively analyze and model the relationships between air pollution levels, node attributes, and temporal patterns within the network.

Spatiotemporal air pollution forecasting involves learning a mapping function f that predicts future air pollution levels given a stationary network topology G and a feature matrix $\mathbf{X}$. The goal is to predict air pollution levels for the next T time steps. The output, denoted as $\mathbf{X}_{t+T}^{N\times M}$, represents the predicted air pollution levels for all N nodes over the next T time steps. In this study, we focus on predicting $PM_{2.5}$ concentrations, setting $M = 1$. However, the proposed framework can be extended to multiple pollutants, treating M as the number of output variables, thereby enabling a multi-task learning approach.

The relationship depicted in our work is formalized by Equation 4.1:

$$X_{t+1}^{N\times 1},...,X_{t+T}^{N\times 1} = f(G, X_{t-n}^{N\times F},...,X_{t-1}^{N\times F}, X_t^{N\times F}) \tag{4.2}$$

Figure 4.1 illustrates the implementation details of our framework, incorporating the parameters described earlier. The figure showcases the architecture and modules of our model, demonstrating how the various elements interact to forecast air pollutant levels for the future. Each component will be elaborated upon in the following sections.

4.2.1 Graph Temporal LSTM (GT-LSTM)

To effectively capture both spatial and temporal dependencies among monitoring stations, we propose a novel spatiotemporal model termed GT-LSTM. Our study focuses on four monitoring stations within the air pollution network of Dijon, France: Canal, Hoche, Carnot, and Janin, designated as node 1 to node 4, respectively. Figure 3.2 provides a visual representation of the spatial distribution of the sensors deployed in this city. Data from four blue sites in the black box are included in this study.

4.2.2 Spatial Model

When it comes to predicting air pollution in an urban area, efficient topological resolution is essential. Relying on a single measurement in a city is often insufficient, particularly in larger cities. Therefore, to ensure accurate and comprehensive prediction, it is crucial to consider multiple measurements distributed throughout the urban area, motivating this study to have sub-zones for monitoring. To extract spatial data from a graph using a neural network, we employ GCN, which is an extension of CNN specifically designed to handle diverse graph-structured data [GDCM23]. In GCN, the process involves multiplying the input neurons by a set of weights, known as *filters* or *kernels*. These filters act as a sliding window across the entire data, allowing GCN to learn the characteristics of neighboring nodes within the graph. As

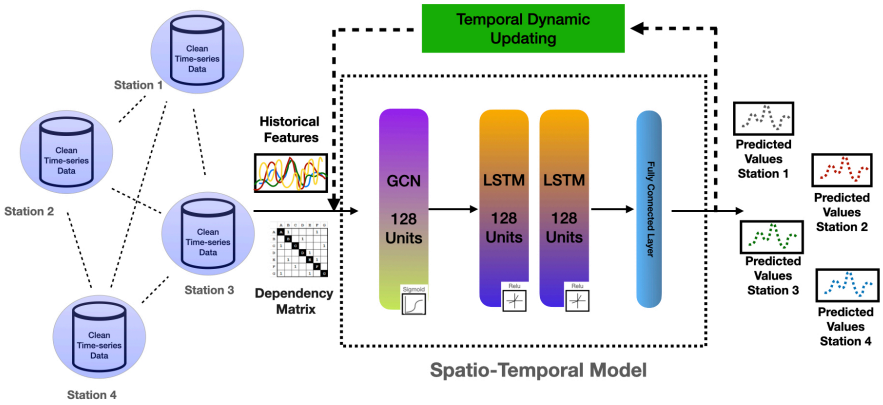


Figure 4.2: Proposed spatiotemporal model. The model incorporates GCN blocks for capturing spatial features, LSTM blocks for capturing temporal features, temporal dynamic updating blocks, input data, a dependency matrix representing spatial relationships, and the predicted outputs.

described earlier, we effectively capture and learn spatial information from the graph structure. A GCN operates on:

- An input feature matrix X of size $N \times F$, where N represents the number of nodes and F denotes the number of input features for each node.
- A matrix representation of the graph structure A of size $N \times N$, such as the adjacency matrix of the graph G.

Figure 4.2 visually represents the spatial distribution of the four monitoring stations and the proposed model architecture. The figure highlights how data from these stations are utilized to predict air quality at four different locations within the network topology and illustrates the model's capability to capture temporal dynamics. The model is lightweight, featuring one GCN layer with 128 units and two LSTM layers with 128 units each.

4.2.3 Temporal Model

To capture temporal dependencies, we employ an optimized temporal LSTM model, as detailed in our previous work [RCAM+24]. This lightweight and compact model has proven effective for various nodes. The comprehensive framework of PMFORECAST designed for air pollution prediction is outlined, comprising four key steps: data pre-processing, temporal attention to mitigate gradient disappearance, a flexible prediction horizon for dynamic future forecasting, and layers employing Long Short-Term Memory (LSTM)—the trainable component. The temporal attention prediction horizon mechanisms are encapsulated in the block *Temporal Dynamic Updating* in Figure 4.2. This block is responsible for using temporal features such as

the day of the week and also adjusting the prediction horizon for longer future forecasts. The LSTM model captures the temporal variations in the data, while the GCN model takes into account the topological structure and dependencies of the nodes.

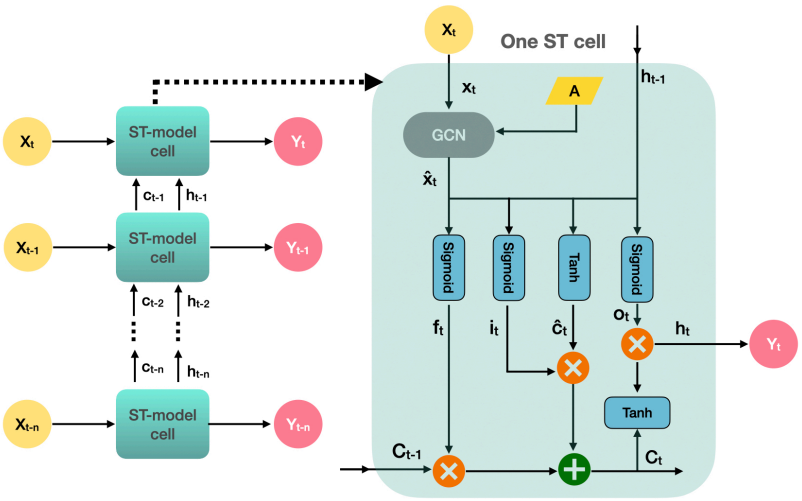


Figure 4.3: A Cell of GT-LSTM

$$\hat{x_t} = f(A, x) \tag{4.3}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, \hat{x_t}] + b_f) \tag{4.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, \hat{x_t}] + b_i) \tag{4.5}$$

$$\tilde{c_t} = \tanh(W_c \cdot [h_{t-1}, \hat{x_t}] + b_c) \tag{4.6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t} \tag{4.7}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, \hat{x_t}] + b_o) \tag{4.8}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4.9}$$

Figure 4.3 illustrates the architecture of a single GT-LSTM cell. The computational process within this cell is outlined by Equations 4.3 to 4.9. The GCN component, represented by Equation 4.3, processes input features x through the adjacency matrix A to generate spatial embeddings $\hat{x}_t$. The LSTM component, defined by Equations 4.4 to 4.9, captures temporal dependencies through the calculation of different gates:

- f_t (forget gate): Decides whether the information can pass through different layers of the network. It takes input from the previous hidden state h_{t-1} and the current input $\hat{x}_t$.
- i_t (input gate): Determines the importance of the information by updating the cell state. It measures the integrity and importance of the information for developing predictions. The information passes through the sigmoid and tanh functions; the tanh eliminates the bias of the network, and the sigmoid determines the weight of the information.
- $\tilde{c}_t$ (cell state candidate): Represents the new candidate values that could be added to the cell state.
- c_t (cell state): The cell state is updated by combining the forget gate and input gate outputs.
- o_t (output gate): The correct information passes through the cell state. Once here, the output of the input gate and forget gate is multiplied by each other. The output gate determines the next hidden state of the network.
- h_t (hidden state): The output gate decides the next hidden state. The updated cell state c_t goes through the tanh function and is multiplied by the sigmoid function of the output state.

The weighted parameters W in these equations are learned during the training process. The final output, h_t, represents the predicted air pollution level for the current time step at a specific node where the time step corresponds to the measurement interval of the station.

4.3 Data

The datasets used in this study are the same as those detailed in section 3.3, collected using Qameleo, an affordable advanced air quality micro-station. These datasets underwent several preprocessing steps to prepare them for analysis. The preprocessing included converting the data from quarterly to hourly intervals, addressing missing values, introducing time-related features such as the day of the week and hour of the day, and standardizing the measurements through normalization.

After converting the datasets from quarterly to hourly intervals, a moving average methodology was employed to address missing data and enhance the overall quality of the data. A sliding window technique with a 12-hour interval was applied to each data point to smooth the series and manage gaps effectively. Additionally, time-related features were integrated into the datasets, and the values were normalized to a range between zero and one to ensure consistency and comparability. A concise summary of these input features can be found in Table 4.1.

Variables	Unit
Pollutants	
PM_1	ug/m^3
$PM_{2.5}$	ug/m^3
PM_{10}	ug/m^3
Meteorology	
Humidity	%
Temperature	°C
Date Time	
Day of Week	0 to 6
Hour	0 to 23

Table 4.1: Summary of dataset variables and their corresponding units. The table includes pollutant concentrations, meteorological measurements, and date-time information.

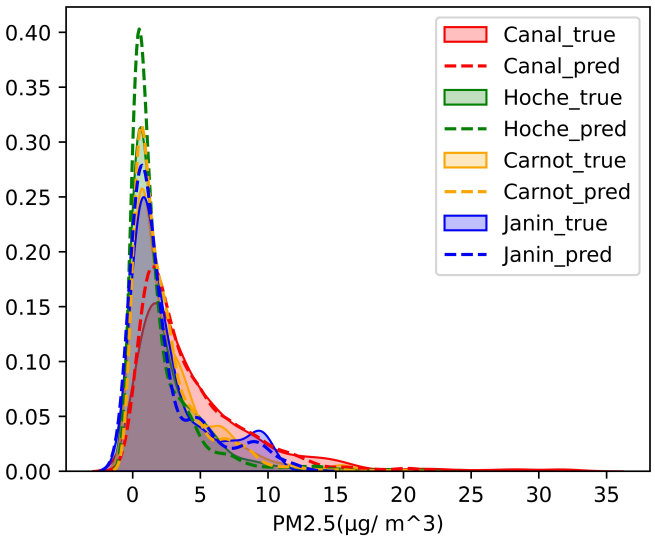


Figure 4.4: Examination of the correlation between observed and predicted values for the test sets, assessed through Gaussian distribution analysis. The observed values are illustrated with dotted lines, while the predicted values are shown with solid lines. Different colors represent the four monitoring sites: Canal (red), Hoche (green), Carnot (orange), and Janin (purple). The distribution of the forecasting values is closely aligned with the actual data, providing insight into the model's reliability across different spatial contexts.

4.4 Model Performance Evaluation and Analysis

To evaluate the model's performance, we employed a combination of visual and quantitative analysis.

Firstly, we visualized the relationship between observed and predicted values using Gaussian distribution plots (Figure 4.4). The Gaussian graph (cf. Figure 4.4) vividly illustrates the model's correlation and remarkable precision across all sites represented by the different colors for test sets. The close alignment of predicted values with the actual measurements across all nodes demonstrates the model's accuracy and reliability. The Gaussian distribution curve further emphasizes the model's accuracy, with its peak closely aligned with the bell shape of the graphs.

Secondly, we quantitatively assessed the model's performance using established metrics: RMSE, MAE, WMAPE, and R^2. The results, presented in Table 4.2, indicate strong performance, particularly evident in the high R^2 values for both training and test sets. The evaluation metrics outlined in Table 4.2 underscore the robust performance of the model, particularly evident in the coefficient of determination (R^2) across both train and test datasets. A high accuracy (ACC) value (0.96) for the training set signifies a strong correlation between predicted and actual measurements, while a value of (0.88) for the test set highlights the model's exceptional

Table 4.2: Performance comparison of the GT-LSTM model against baseline models (LSTM and GCN) using RMSE, MAE, R^2, and WMAPE. Lower values of RMSE, MAE, and WMAPE indicate better performance, while higher R^2 values are desirable.

Methods	Datasets	RMSE	MAE	R^2	WMAPE
LSTM	Train-set	0.805	0.511	0.988	0.074
	Test-set	0.935	0.534	0.934	0.155
GCN	Train-set	1.706	1.058	0.909	0.157
	Test-set	1.599	1.011	0.782	0.306
GT-LSTM	Train-set	1.301	0.697	0.966	0.102
	Test-set	1.139	0.650	0.882	0.211

predictive accuracy. To benchmark our model, we compared our results with those obtained from running the base models independently using identical hyper-parameters for LSTM and GCN, as illustrated in Table 4.2.

Although the results show slightly better performance when using only the LSTM model, the primary objective of utilizing a unified model is to achieve enhanced resolution. Despite the temporal model's superior performance on its respective dataset at each site, the strength of this approach lies in its capacity to achieve high resolution with a single model across diverse databases. The proposed GT-LSTM model improves spatial resolution within the city's topology while reducing processing costs, offering a distinct advantage over employing multiple models for individual nodes, particularly when managing multiple stations within a city.

To further investigate the model's ability to capture spatial dependencies, we conducted a detailed analysis focusing on two closely located sites: Hoche and Canal. As shown in Figure 4.5, these sites are approximately 2 kilometers apart. The model effectively captures the similar behavior of air pollution levels at these sites, as evidenced by the closely aligned predicted values in Figure 4.5(b). Figure 4.5(c) provides a visual representation of the site locations using OpenStreetMap for reference.

4.4.1 Evaluating the Model's Ability to Capture Spatiotemporal Patterns

One of the primary objectives of this study is to generate predicted values for various nodes by feeding the model with different datasets, with a specific emphasis on the $PM_{2.5}$ pollutant. We provide compelling evidence that spatiotemporal modeling, which accounts for inter-node dependencies, significantly enhances the accuracy of pollution level forecasts. This is demonstrated through results from four monitoring sites, notably at Node 4 at the Janin site, where no actual values were available in the test set for a certain period. Remarkably, our model was able to predict values for this site despite the absence of actual measurements. Figure 4.6 illustrates the test-set data across all nodes, with the Janin site highlighted in the lower right section of the graph. This achievement highlights the model's capability to effectively capture

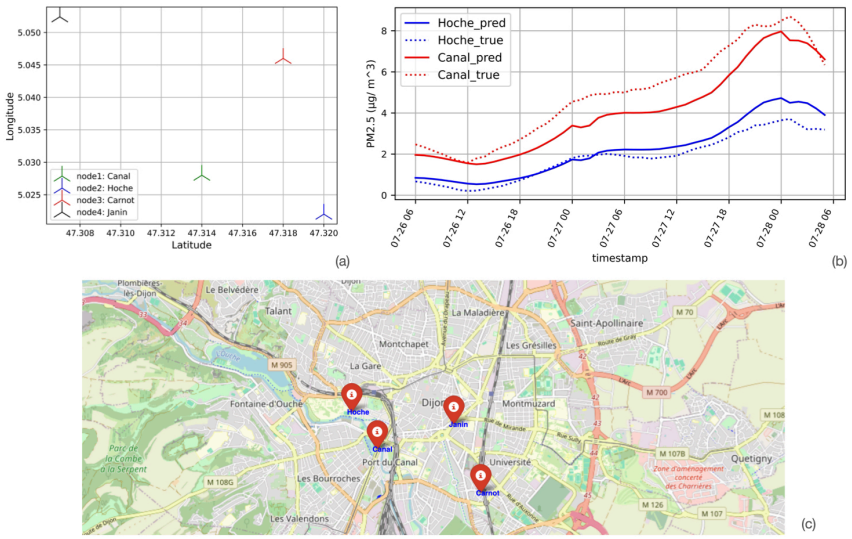


Figure 4.5: (a) Geographical coordinates of all sites. (b) Observed and predicted values for the Hoche and Carnot sites over two days. Dotted lines represent the true values, while solid lines indicate the predicted values. The Hoche data is shown in blue, and the Carnot data is depicted in red.

spatiotemporal relationships and provide robust forecasts, even in the presence of missing data, by leveraging both spatial and temporal correlations.

We conducted another test to assess the stability of the model by using zero values as the input for the Carnot site test set and predicting the values for all four nodes. For the three other sites, there was a slight increase in error. However, for Carnot, the predictions still captured the ground truth data patterns well. Figure 4.7 illustrates two scenarios for the x-test set at the Carnot site: with and without real measurements. The graph shows that in the scenario without real measurements, the predictions, although accurate, had a narrower range of variation in the box bar graph on the right. This indicates the model's limitation in distinguishing outliers when no actual input is provided, which is expected since the model lacks the historical data to predict values accurately.

4.4.2 Model Capability for Long-Term Forecasting

Our forecasting model demonstrates strong efficacy in predicting air pollution levels over extended time spans. Figure 4.8 and Figure 4.9 presents error and correlation metrics for both the training and test datasets across various time intervals, ranging from 1 hour to 36 hours into the future. These values represent the average of predictions across all nodes, aggregated for each specific forecast horizon.

In Figure 4.8, the model shows notable error reduction within the training dataset, achieving an average RMSE value of 2.195 for the first 12 hours in the test set. This indicates a significant

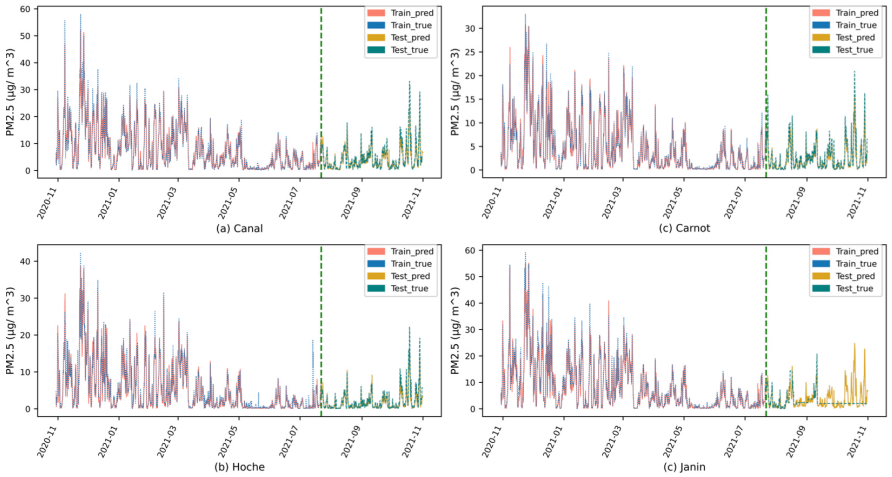


Figure 4.6: Illustrating Model Robustness: predictions for all locations. The dashed lines represent the collected data, reflecting the actual values during both the training (blue) and prediction (golden) phases. The solid lines depict the $PM_{2.5}$ predictions made during the training (salmon) and prediction (green) phases. The vertical green dashed line marks the boundary between the training and testing datasets.

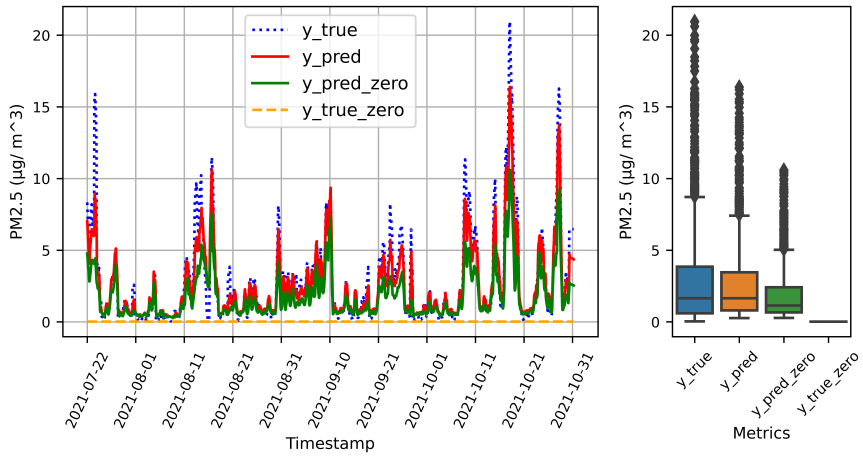


Figure 4.7: Assessing Model Robustness: Two testing scenarios are considered. (i) Without Real Measurements (Using Zero Values): The dashed orange line represents zero values as input, while the solid green line shows the predicted values for this scenario. (ii) With Real Ground-Truth Measurements: The dotted blue line represents the actual ground-truth values, and the solid red line depicts the $PM_{2.5}$ predictions made using the real measurements for this scenario.

correlation between predicted and actual values. The model also exhibits robust performance on the independent test dataset, with an average MAE value of 1.351, demonstrating its ability to generalize effectively and provide reliable predictions even for unforeseen data beyond the 12-hour mark. It is observed that RMSE and MAE values in the training set are higher compared to the test sets. This discrepancy is due to the wider range of measurement values in the seasonal and time series data of the training sets, resulting in larger values compared to the test sets.

Figure 4.9 evaluates the model's accuracy over longer forecast horizons. After 24 hours, the accuracy of predictions decreases by more than 50 percent. Despite this, the model's performance indicates that extending forecasts beyond this period presents challenges. To improve long-term prediction quality, especially with large datasets, it is recommended to consider daily average predictions after the 24-hour mark.

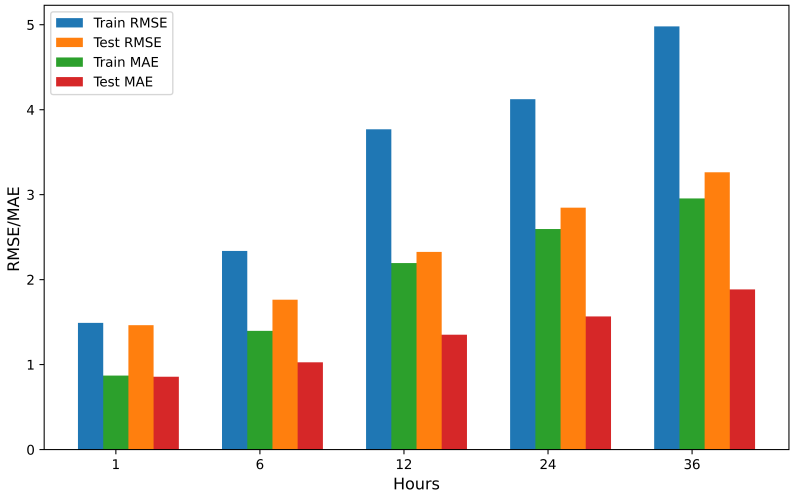


Figure 4.8: Performance Evaluation of Long-Term $PM_{2.5}$ Forecasting with GT-LSTM: The barcharts illustrates the RMSE and MAE errors for both the training and test sets, showing an increase in error over time. Blue and green bars represent the training set metrics, while orange and red bars correspond to the test set metrics.

4.4.3 Experimental Setup

hyper-parameters are crucial in determining the performance of deep learning models, affecting aspects such as learning rate, batch size, training epochs, and the number of hidden units.

For the GCN layer, we systematically explored various configurations to optimize performance. We tested unit counts of $[32, 64, 128, 256]$ with 128 units showing the most promising results. Further analysis indicated that a single layer provided the best performance, proving to

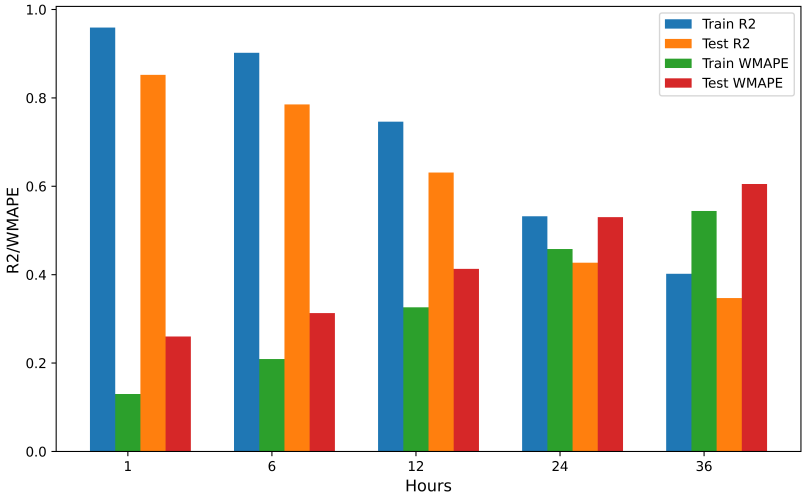


Figure 4.9: Performance Evaluation of Long-Term $PM_{2.5}$ Forecasting with GT-LSTM presents the R^2 and WMAPE metrics, showing a decrease in accuracy over time. Blue and green bars represent the training set metrics, while orange and red bars correspond to the test set metrics. Accuracy is assessed using both R^2 and WMAPE.

be the most effective configuration. The model utilized the Sigmoid activation function, which helped capture complex spatial relationships and dependencies in the data.

In the temporal model, we adopted the same structure as in our previous work, as it demonstrated the best performance. Specifically, we employed two layers of LSTM, each with 128 units and using the ReLU activation function as detailed in 3. This configuration achieved optimal performance while keeping the model structure efficient. The model was trained using the mean squared error MSE for the loss function, chosen for its effectiveness in regression tasks. Throughout our research, careful attention was given to the selection of hyper-parameters. We set the learning rate to 0.001 to regulate the step size for adjusting the model's parameters during training. Additionally, we conducted 200 training epochs with an early stopping criterion to prevent over-fitting, ensuring thorough training and reliable results.

4.5 Discussion

This chapter presents a novel spatiotemporal model for accurately predicting $PM_{2.5}$ concentrations in suburban environments. By leveraging low-cost sensors, advanced modeling techniques, and AI, the model effectively captures the spatial and temporal dynamics of air pollution. It consistently demonstrates superior prediction accuracy and robustness, as indicated by high R^2 and low error rates, even when faced with limited data. Importantly, the model provides reliable forecasts without the need for complex multi-scale dependencies or models.

In addition, this approach offers valuable insights for long-term air quality planning and pollutant prediction across various regions with sparse observational data. The model supports real-time monitoring, making it a useful tool for policymakers and the general public, contributing to improved air quality management and public health outcomes.

The model's performance across diverse temporal and spatial contexts further highlights its robustness and adaptability. Notably, the GT-LSTM model effectively generalizes and maintains high R^2 even when applied to diverse datasets, including those with missing data. This is evident from its success in predicting values at the Janin site, where the model accurately interpolated data despite the absence of actual measurements. This demonstrates the model's ability to exploit spatiotemporal dependencies, ensuring high-quality predictions under challenging conditions.

While the model excels within the first 12 hours of forecasting, with relatively low RMSE and MAE, there is a noticeable decline in accuracy for predictions extending beyond 24 hours. This suggests limitations in capturing long-term temporal patterns, likely due to increasing uncertainty over extended forecast horizons. To address this, future work could explore ensemble learning or hybrid approaches, combining GT-LSTM with other predictive techniques to enhance long-term forecasting accuracy.

In conclusion, the GT-LSTM model represents a powerful tool for spatiotemporal forecasting in urban air quality monitoring. However, its resolution is heavily dependent on the number of nodes and dataset characteristics, indicating that scalability is tied to the predefined sub-zones within the model's architecture.

To improve the spatiotemporal resolution of air pollution monitoring, we propose integrating mobile sensor networks and leveraging crowdsensing resources. Deploying mobile sensors in urban areas will enable more fine-grained air quality data collection, leading to more accurate and localized predictions while reducing the cost of data collection and processing.

Chapter 5

FEDAIRNET

5.1 Background and Context

Introduction

We discussed the importance of capturing both the temporal and spatial characteristics of air pollution data. Traditional centralized systems for monitoring air quality are often costly and provide limited spatial coverage. Establishing numerous sites to collect, analyze, and transmit data to a central server further increases expenses and delays. FL offers a promising alternative by enabling decentralized model training across a network of devices. Through FL, we can leverage a wide array of mobile sensors distributed across various locations to collect real-time environmental data, such as PM levels and meteorological conditions, without the need to centralize sensitive information. This decentralized approach not only improves the granularity and coverage of air quality assessments but also preserves user privacy by keeping data locally in the devices.

State of the Art

The quest for effective air quality monitoring has been revolutionized by the advent of FL, a technology that addresses the pressing challenges of privacy and data management. One innovative approach employs federated learning through a mobile Android application, which enhances air quality monitoring while safeguarding user privacy. By leveraging on-device training, this method avoids data transmission, preserving sensitive information. The system utilizes federated averaging to aggregate updates from multiple devices, resulting in an accurate prediction of the AQI based on image features and weather data. Despite its promising capabilities, this method faces limitations, particularly in low-light or indoor conditions, and requires broader data diversity for improved accuracy and utility [CR21].

In a different realm of air quality monitoring, mobile devices have been employed to apply federated learning for real-time AQI prediction and hazardous zone detection. This

approach focuses on decentralized data collection via mobile devices, allowing for extensive monitoring of large areas while ensuring privacy by keeping raw data on the devices. The use of LSTM networks enables precise AQI predictions, with each mobile device contributing to a global model through federated averaging. While effective in urban environments, this approach faces challenges related to the computational limitations of edge devices and the coordination of multiple mobile devices, which can lead to latency and increased network traffic [CTK+21, LNL+20, LNL+21].

A comprehensive review of federated learning applications in air quality forecasting reveals a significant evolution from traditional machine learning methods. Before 2020, AQI forecasting relied heavily on single-machine learning techniques such as *Deep Neural Networks*s (DNNs) and ANNs, which struggled with local optima. The introduction of federated learning has brought about centralized, decentralized, and hierarchical architectures, with centralized FL being the most common. Multi-model FL approaches, combining various models to leverage their strengths, have shown potential in improving prediction accuracy. The review suggests that future research should focus on refining algorithms and integrating DNNs to better manage partial and temporal data [DKD+22].

Further advancing this field, research on Federated Compressed Learning (FCL) integrates data compression with federated learning and edge computing to address challenges such as data sparsity and high computational costs. This framework reduces data volume, maintains privacy through local model training, and enables real-time analysis on resource-constrained devices. While FCL successfully minimizes data usage and enhances privacy, the reliance on low-cost sensors introduces limitations, including accuracy issues and sensitivity to environmental factors like humidity and temperature, which can affect data reliability, especially in areas with sparse sensor coverage [DP22, PCP+21].

Integrating FL (Federated Learning) with mobile sensors for air quality monitoring addresses several limitations inherent in centralized systems. Centralized approaches often grapple with issues such as data privacy concerns, high operational costs, and limited scalability. Mobile sensors complement this FL approach by offering a broad distribution of data sources, which enhances the accuracy and timeliness of air quality assessments. This integration aims to create a monitoring system that is efficient, scalable, and responsive to dynamic environmental changes. Federated learning represents a significant advancement in air quality monitoring, combining innovative technology with practical solutions to foster more effective and privacy-conscious environmental management.

Our objective is to design a FL system that adapts seamlessly to urban topologies by harnessing the potential of crowdsourced data. Additionally, we need tools to simulate mobility tracking and privacy-preserving techniques. The ACCIO framework [PMB+18], which provides

tools for simulating such techniques, will be utilized to extract potential PoI attacks in our federated learning models.

5.2 Methodology

Our approach focuses on monitoring and forecasting air quality using geospatial data collected from mobile sensors. We aim to achieve high-resolution air quality monitoring in urban areas by distributing data processing and model training across multiple devices. This approach leverages the mobility and density of sensor-equipped devices to gather fine-grained environmental data, offering a superior spatial resolution compared to traditional stationary air quality monitoring stations.

The core idea is inspired by the concept of crowdsensing, which involves collecting real-time and abundant measurements throughout urban areas. In our framework, the city is divided into distinct regions, with each region responsible for locally predicting and monitoring air pollution. Figure 5.1 illustrates the city of Dijon, which is the focus of our research. The city is represented by a grid map where each cell functions as a node equipped with an edge computing device. As shown in the figure, citizens collect measurements as they move through the city.

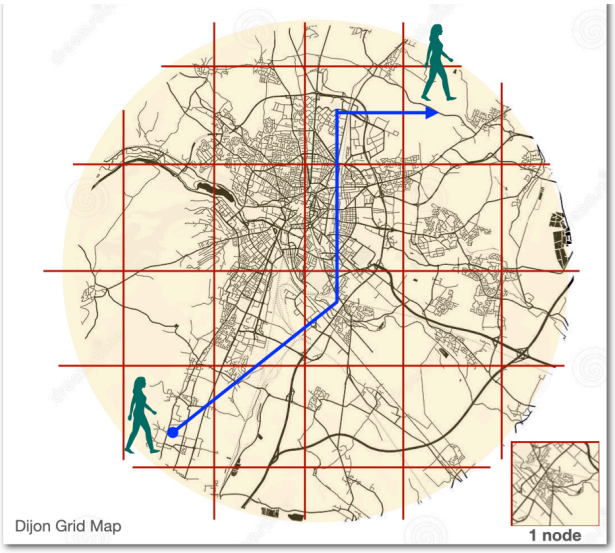


Figure 5.1: Assumed a Grid Map of Dijon: Data Collection by Citizens Using Mobile Sensors

5.2.1 Federated Learning Model Training

This research proposes a federated learning framework utilizing mobile sensors and a decentralized approach. We suppose that the framework enhances prediction accuracy while safeguarding privacy.

The framework relies on federated learning, leveraging data from mobile sensors and cross-device techniques to overcome the limitations of centralized systems. Multiple nodes are distributed across different regions, each functioning independently to train a local model with its collected data. These regional models are then aggregated into a global model, which

benefits from diverse data sources while preserving individual user privacy. The number of regions can be adjusted based on the city's topology and structural complexity. This approach aims to improve model performance and minimize the risk of raw data exposure by keeping data localized within each node.

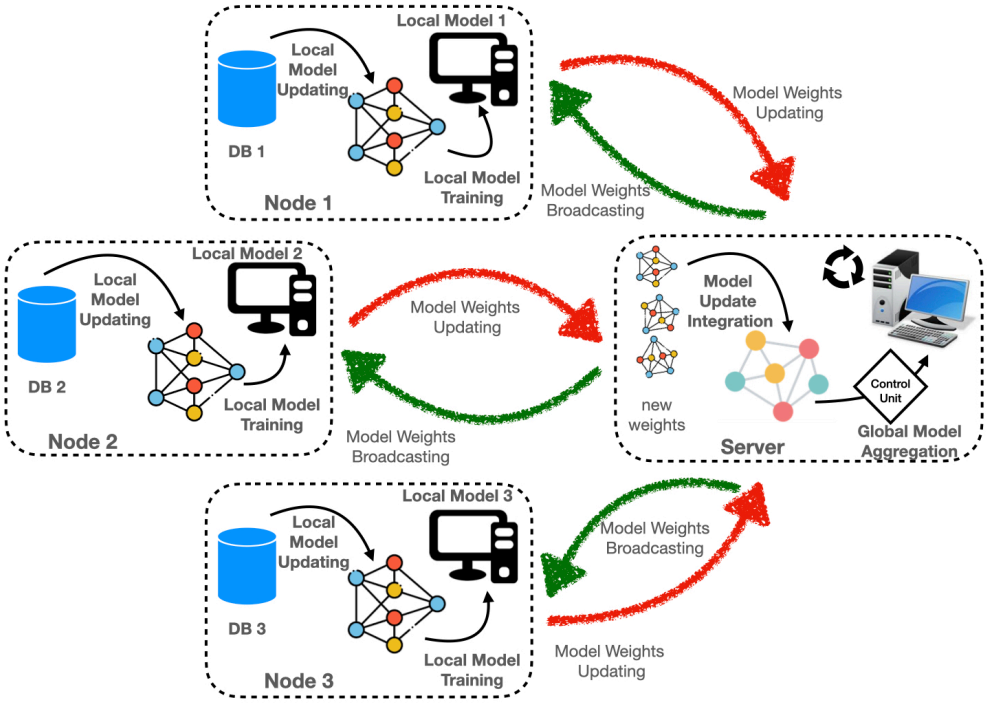


Figure 5.2: Overview of the proposed federated learning framework for secure, decentralized air quality prediction with localized model training on distributed nodes

System Architecture

The system architecture is essential for enabling the decentralized data processing inherent in federated learning. Figure 5.2 represents the proposed framework includes a central server and numerous nodes distributed throughout the city. Each node collects and processes data from mobile sensors in its designated geographic area. The central server aggregates updates from all nodes to refine the global model, thus reducing data transmission across the network.

Local Training: Each node, representing a cluster of user devices within a specific geographic area, collects data from nearby mobile sensor devices. The predictive models are then trained on these edge devices at each node. This local training approach ensures that raw data remains on the edge devices, thereby preserving user privacy.

Model Aggregation: A crucial component of the federated learning framework is the aggregation algorithm, which merges local model updates from various nodes to form a global model. We utilized FedAvg, a widely adopted algorithm developed by Google [MMRyA16],

for this aggregation task. The training process continues until either a predefined number of rounds is completed or a convergence criterion is satisfied, thus ensuring effective global model training while minimizing unnecessary computational overhead.

The nodes can be either active or silent during the aggregation process of the global model. This setup helps mitigate the impact of malicious or faulty nodes contributing subpar data through a control mechanism that filters out those consistently producing measurements with significant deviations from expected norms. By excluding these outlier nodes from the aggregation process, we enhance the overall quality and reliability of the global model, ensuring that only trustworthy data is used.

Our methodology is designed to forecast regional air quality values rather than focusing on specific point measurements. This approach allows us to account for variations in air quality across broader areas, thereby reducing the influence of anomalies caused by localized pollution sources, such as smoking or vehicular emissions.

Global Model Distribution: The updated global model, enriched with insights from various local models, is then distributed back to the individual nodes. This enhanced model empowers each node to conduct real-time air quality monitoring and forecasting with greater accuracy. The high-resolution data collected by mobile sensors within each node's jurisdiction enables detailed and localized assessments of air quality.

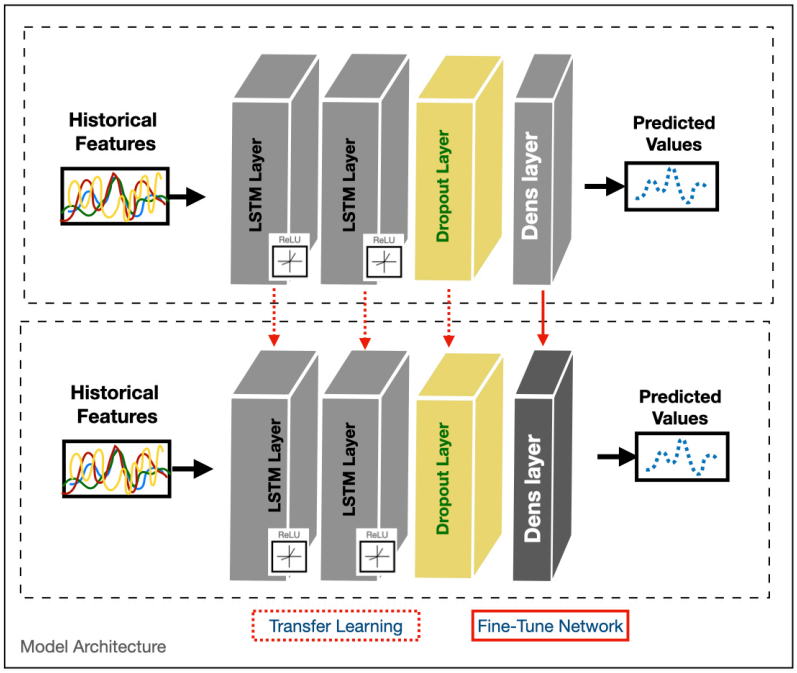


Figure 5.3: Architecture of the proposed transfer learning model for local training on nodes and edge devices within the federated learning framework

Federated Transfer Learning adapts the traditional machine learning technique of transfer learning, wherein a new model is trained using a pre-trained model that has been fine-tuned on a similar dataset to address a related or different problem. Leveraging a pre-trained model in machine learning often results in significantly better performance compared to training a model from scratch.

As discussed in references [CWY+19, YHZC21, LCY18], federated learning can be effectively combined with transfer learning. Given that our dataset consists of limited time-series samples (one month) and requires frequent updates, the available data per node or edge device may be insufficient. To address this challenge, we utilize larger datasets and pre-trained ML models to enhance and accelerate the training process.

For our specific use case, we employed a lightweight, well-trained Temporal LSTM model to forecast $PM_{2.5}$ levels in urban areas, as shown in chapter 3. Figure 5 illustrates the architecture of this model, where the LSTM layers are kept frozen, and only the final dense layer is fine-tuned to adapt to new data. The horizon prediction, as discussed in chapter 3, determines the number of output neurons.

Mathematical Formulation of Federated Transfer Learning

We formalize the federated transfer learning process with the following mathematical definitions:

- ϕ represents the pre-trained feature extractor, which remains frozen during local training.
- ψ represents the final dense layer, which is fine-tuned.
- $w_{\psi,i}^{(t)}$ denotes the parameters of the dense layer ψ at node i after t local training steps.
- $w_{\psi}^{(t)}$ denotes the global parameters of the dense layer after t rounds of aggregation.
- $\mathscr{L}(w_{\psi,i}^{(t)}; D_i)$ represents the loss function at node i for the dense layer, where D_i is the local dataset collected throughout the designated geographic area.

The local update at each node is given by:

$$w_{\psi,i}^{(t+1)} = w_{\psi,i}^{(t)} - \eta \nabla \mathscr{L}(w_{\psi,i}^{(t)}; D_i) \tag{5.1}$$

where:

- η is the learning rate.

- $\nabla \mathscr{L}(w_{\psi,i}^{(t)}; D_i)$ is the gradient of the loss function with respect to the dense layer parameters at node i at training step t.

The global aggregation process is defined as:

$$w_{\psi}^{(t+1)} = \sum_{i=1}^{N} \frac{|D_i|}{\sum_{j=1}^{N} |D_j|} w_{\psi,i}^{(t+1)} \tag{5.2}$$

where:

- $w_{\psi}^{(t+1)}$ is the updated global parameter for the dense layer.
- The aggregation is weighted by the size of the dataset at each node.

After aggregation, the global parameters are redistributed to the nodes:

$$w_{\psi,i}^{(t+1)} \leftarrow w_{\psi}^{(t+1)} \tag{5.3}$$

The overall federated transfer learning process can be summarized as follows:

1. **Initialization**: The server initializes the global dense layer parameters $w_{\psi}^{(0)}$ and distributes them to all nodes, along with the frozen feature extractor ϕ.
2. **Local Update**: Each node i performs local training on its dataset D_i, updating only the dense layer parameters $w_{\psi,i}^{(t)}$ while keeping ϕ unchanged.
3. **Model Aggregation**: The server aggregates the updated dense layer parameters from all nodes to form a new global dense layer $w_{\psi}^{(t+1)}$.
4. **Global Distribution**: The updated global dense layer parameters $w_{\psi}^{(t+1)}$ are distributed back to each node.
5. **Repeat**: Steps 2-4 are repeated until convergence or a stopping criterion is met.

5.2.2 Privacy Preservation

This research proposes a federated learning framework for air quality monitoring that prioritizes user privacy. The core idea involves decentralized data collection and processing, where individual users collect local measurements and share limited data with nearby nodes. While federated learning offers significant privacy benefits, there remains a risk of data inference. Previous studies have demonstrated that user check-ins can be inferred from PoI embeddings. PoIs are specific locations of interest, such as restaurants, shops, or landmarks, that users

frequently visit. To address this, our research employs simulated data to safeguard privacy while accurately monitoring air quality along user routes.

We integrated the ACCIO framework to extract PoIs from user trajectories [PMB+18]. By analyzing these PoIs, we can assess the effectiveness of our privacy measures and identify potential vulnerabilities. For instance, if a curious node can accurately determine a user's frequent PoIs, it may be able to infer sensitive information about that user's habits or lifestyle. By varying the number of nodes, we can evaluate the trade-off between data granularity and privacy protection.

To further enhance privacy, we proposed various levels of node granularity by exploring different topologies tailored to varying privacy requirements. These configurations range from coarse-grained to fine-grained setups for each node. Figure 5.4 visualizes the four levels of node topology, (a) Zone, (b) Sub-Zone, (c) Grid Street, and (d) Pinpoint Location as detailed below.

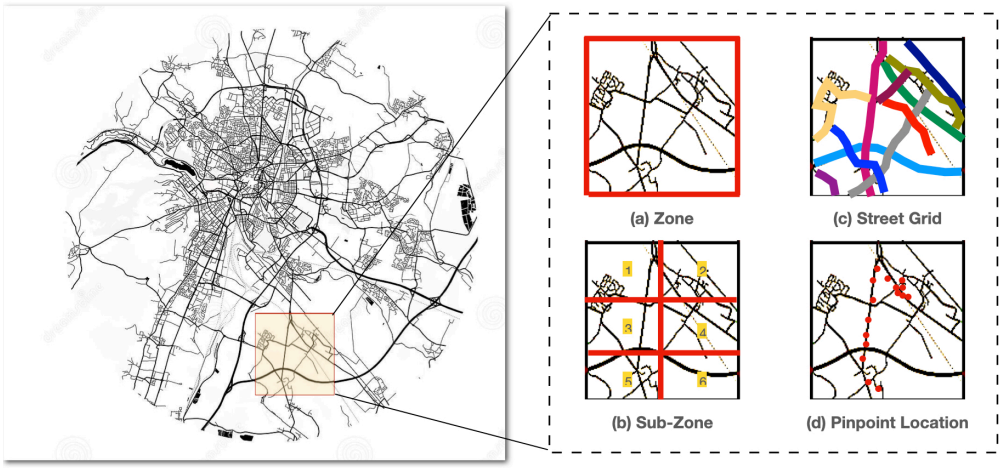


Figure 5.4: Visualization of the four node topology levels within the city scale: (a) Zone, (b) Sub-Zone, (c) Grid Street, and (d) Pinpoint Location.

- **Zone:** Represents the entire geographic area covered by a node. In this configuration, user devices first identify the relevant node and then remove precise geolocation data from their dataset before transmitting it to the corresponding edge device. This method offers the highest level of privacy by generalizing the location information to a broad area. Each node is depicted as a cell in Figure 5.1 and Figure 5.4 (a).

- **Sub-Zone:** To balance data granularity and privacy, we propose dividing each node into smaller sub-zones. This approach enables more localized data collection while limiting shared location data to the sub-zone level, thus maintaining user privacy. Implementing sub-zones requires additional resources to manage the increased computational load. To ensure scalability and efficiency, sub-servers can be deployed within each sub-zone, or

one sub-zone can be designated as the server. This reduces communication overhead and improves system responsiveness. Figure 5.4 (b) illustrates the sub-zones within a node, divided into six sub-nodes.

- **Street Grid:** This approach focuses on specific streets or sets of streets within a node's zone, enabling detailed monitoring along designated routes while protecting broader location privacy. An algorithm within each node identifies the street name and obfuscates the precise location by removing the last digits of latitude and longitude, retaining street-level accuracy. The anonymized data is then transmitted to the corresponding node for processing, ensuring privacy while facilitating localized analysis. Figure 5.4 (c) illustrates the Street Grid concept.

Algorithm 1 Street-Level Location Masking and Data Transmission

1: **Input:** Location data `(lat, lon)` for a user within a node.
2: **Output:** Masked location data sent to the corresponding node.
3: Determine the street corresponding to the user's `(lat, lon)`.
4: Remove the last digits of `lat, lon` to anonymize the location while keeping it within the same street.
5: Send the masked location data to the corresponding node for processing.

- **Pinpoint Location:** This is the most fine-grained level, representing an exact point within the node, such as a specific GPS coordinate. While it offers the highest precision in data collection, it also presents significant privacy risks. In the event of a malicious node, there is a potential for user re-identification through attacks, such as PoI attacks. Figure 5.4 (d) illustrates the concept of Pinpoint Location, where a partial trajectory of a user is shared.

By varying the level of granularity, we can strike a balance between data aggregation and privacy protection. Finer-grained levels allow for more localized analysis, but may increase the risk of data inference. Conversely, coarser-grained levels reduce the risk of inference but may sacrifice data granularity.

5.2.3 Implementation

Our implementation involved developing algorithms to optimize city segmentation into nodes, using Dijon as a case study. We designed an algorithm to determine the optimal number and configuration of nodes across the city map, dividing it into manageable sections for efficient data processing. The simulated data, detailed in the next section, guided our process. We varied scenarios such as the number of epochs, evaluation rounds, nodes per round, time intervals, and prediction horizons, while keeping certain hyperparameters—like activation functions for LSTMs, model architecture, batch size, learning rate and loss function—constant to control for variables. Early stopping was applied after 15 epochs. We used the Flower framework [BTM+22] with a TensorFlow backend, customizing its modules and initializing the

global model with temporal weights and parameters from chapter 3. Various scenarios were simulated to evaluate local and global model performance, and the Hydra package facilitated real-time evaluation and result tracking.

5.3 Data Simulation

5.3.1 Data Collection

The cornerstone of our project in this chapter is the use of mobile sensors embedded in user devices to collect real-time air quality data. The *Appoline* project (Air Pollution and Individual Exposure) [HKV+19] established a robust infrastructure for air quality research and education, leveraging chemical and physical sensor units. These low-cost instruments are capable of measuring key pollutants without disrupting occupants. The devices continuously transmit raw measurements via Ethernet or wireless technologies to a cloud platform, enabling real-time offsite visualization and analysis. Since its deployment in July 2018 across University of Lille buildings, this network has provided valuable insights into both pollution levels and occupancy patterns [CH20].

However, in Dijon, this approach faced several challenges. The main issue was the insufficient participation of volunteers, which limited the quantity and variety of the collected data. Furthermore, managing the mobile sensor network remotely from Lille presented additional difficulties, as it was challenging to coordinate sensor distribution, to oversee data collection, and to maintain smooth operation across different locations. To continue the project, we chose to simulate the data using the same Appoline sensors, maintaining the format and characteristics of the real-world deployment.

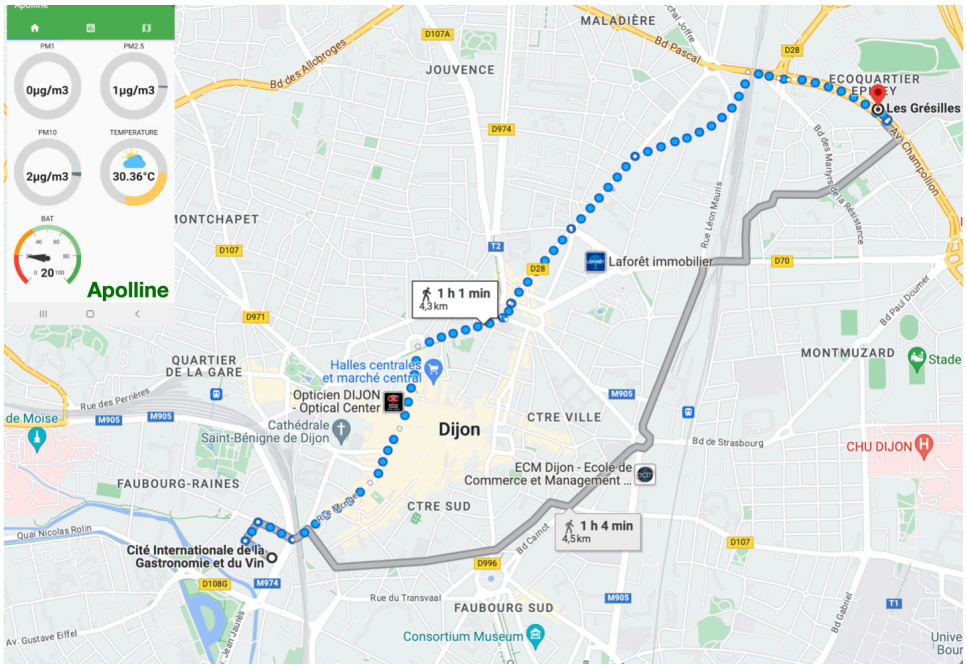


Figure 5.5: Map showing simulated user movements using the Apolline application. The simulation illustrates the user's travel paths and locations over time.

Figure 5.5 illustrates the simulated measurement data collected at various points, represented by blue circles on the map. In our project, the Apolline sensor focuses on monitoring particulate matter (PM_1, $PM_{2.5}$, PM_{10}) in addition to meteorological variables such as temperature and humidity. Furthermore, the sensor captures user movement data, recording latitude and longitude every second. This approach facilitates the collection of diverse environmental data across multiple urban locations, enhancing both the spatial and temporal resolution of our air quality monitoring system.

To tackle the challenge of limited active users for continuous time-series data collection, we utilized an open service map to simulate various user trajectories. Various scenarios were developed, involving up to 100 user paths with random geometric topology and popular PoI, in order to cover as much of the urban area as possible. Each user is modeled to commute to work during weekdays and participate in recreational activities on weekends, with varying amounts of time spent at different PoIs. As noted earlier, PoIs include locations such as homes, workplaces, and recreational venues like gyms and libraries.

We utilized GPX files to simulate user movements and data collection, leveraging the open service map[1] for spatial data visualization. Each user was assigned a set of 7 GPX files, representing a week of travel data from 07:00 to 21:00, which captured typical daily routines, including work hours and weekend activities. The time of departure and arrival home varied individually, with lunch breaks consistently spent at the workplace during weekdays. A custom script was developed to convert these GPX files into CSV format, with columns for date-time, latitude, and longitude. Timestamps were generated at 15-second intervals, utilizing an average walking speed range of 5 to 6 km/h based on estimates from Google Maps. This approach covered the period from 07:00 to 21:00 each day. To unify the datasets while preserving raw data privacy, records before 07:00 and after 21:00 were excluded from the simulated data. The timing for departure, arrival, and duration spent at PoIs varied for each user, resulting in different scenarios for each individual. The time spent at each destination was considered before resuming travel to the next location or returning home. This approach, detailed in Algorithm 2, established a foundational method for simulating user movements and collecting spatiotemporal data.

Haversine Formula for Calculating Distance

---

[1] https://map.project-osrm.org

Radius of the Earth (in kilometers): $R = 6371.0$ (5.4)

Convert latitude and longitude from degrees to radians: (5.5)

$\phi_1 = \text{radians}(\text{lat}_1)$ (5.6)

$\lambda_1 = \text{radians}(\text{lon}_1)$ (5.7)

$\phi_2 = \text{radians}(\text{lat}_2)$ (5.8)

$\lambda_2 = \text{radians}(\text{lon}_2)$ (5.9)

Difference in coordinates: (5.10)

$\Delta\phi = \phi_2 - \phi_1$ (5.11)

$\Delta\lambda = \lambda_2 - \lambda_1$ (5.12)

Haversine formula: (5.13)

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right) \tag{5.14}$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right) \tag{5.15}$$

Distance (in kilometers): $d = R \cdot c$ (5.16)

5.3.2 Data Integration and Enrichment for Air Quality Simulation

To simulate a dataset with real-world environmental measurements, we first generated geodata for users' paths. Each user's data was processed by comparing their latitude, longitude coordinates, and time steps against reference points from the Observation Qameleo Network (section 3.3). These reference points correspond to measurements taken from four Qameleo stations located throughout Dijon. The integration process involved identifying the nearest Qameleo station for each coordinate pair and merging the corresponding air quality as well as environmental measurements from the Qameleo datasets for each user. The distance between two points on the Earth's surface was calculated using the Haversine formula, as detailed in Equation 5.4 to Equation 5.16. The spatial and temporal data collection and integration process is illustrated in Figure 5.6. Consequently, the generated values represent mobile Qameleo values, derived from the real measurements collected at the mobile coordinates corresponding to user movements.

Each user's dataset is uniquely crafted, featuring a single date-time column that simplifies temporal analysis, along with detailed measurements of various air quality indicators. This comprehensive structure, as detailed in Algorithm 3, facilitates seamless integration with analytical tools and models.

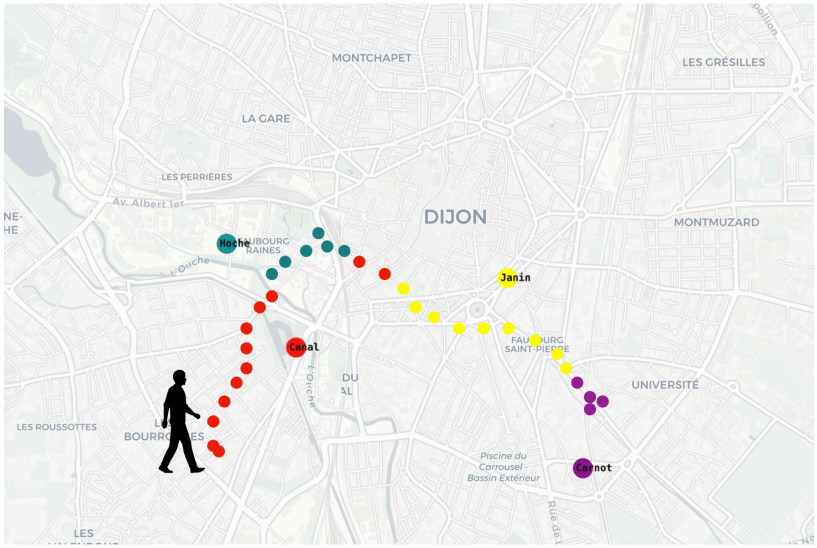


Figure 5.6: Map showing simulated user movements using the Apolline application. The simulation illustrates the user's travel paths and locations over time.

Then, we maintained the same spatial trajectories for users weekly and we replicated the latitude and longitude data for 100 users over the course of an entire month. Similar processes were applied to other temporal features. To protect user privacy, data collection was halted after 21:00 for all users. However, this approach is not a comprehensive solution and may not uniformly prevent reidentification in all cases.

The one-month dataset for 100 users was divided into 80% for training (80 users) and 20% for testing (20 users). The data was distributed across nodes based on geographical regions defined by the network topology, where each node received a subset corresponding to users who passed through its specific region, determined by latitude and longitude. For example, Figure 5.7 illustrates the allocation of a random user's data in a network with 9 nodes, with each color representing spatial data linked to a particular node. The temporal features span from April 1, 2021, to April 30, 2021. To generate a distinct dataset for each node, partial datasets were merged using overlapping and averaging techniques, ensuring a seamless combination of user data. During nighttime hours, from 21:00 to 07:00 the following day, we used Qameleo stationary sensors nearby to obtain the necessary values. Additionally, to accommodate varying temporal resolutions—minutely, quarterly, and hourly—the data was averaged over intervals of one minute, 15 minutes, or one hour, respectively.

Our goal was to consolidate data from multiple users into a comprehensive dataset for each node. Following this, based on the network topology configuration, the training data was distributed across the nodes. Each node's dataset was then divided into a training set (80%) and a validation set (20%). This thorough preprocessing ensured that the data was well-structured and properly distributed, enabling robust training and evaluation of both global

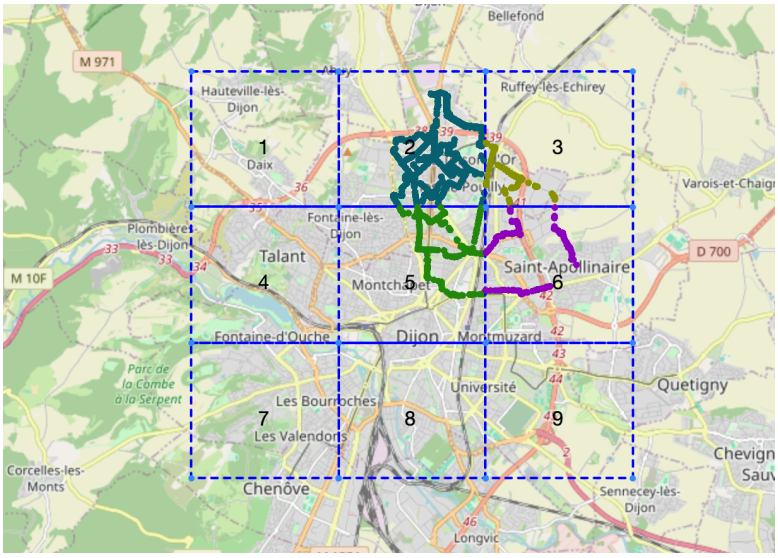


Figure 5.7: Data distribution of a random user's movements across a network configuration with 9 nodes, where each color represents spatial data corresponding to a specific node.

and local models within a federated learning environment. While we recognize that the analysis relied on simulated data—which simplifies real-world complexities—the insights gained can still inform practical applications.

5.4 Results

In this section, we explore how various characteristics of our proposed model, such as grid segmentation, user data distribution, the number of nodes, and the number of rounds, affect both model performance and user privacy. We evaluated our model from two primary perspectives: privacy and performance.

5.4.1 Privacy Analysis

To assess user privacy, we conducted experiments using a PoI attack, which aims to reveal sensitive information by identifying each user's most frequently visited locations and daily patterns. We analyzed the impact of different node segmentation scenarios on privacy, focusing on configurations with 1, 4, 9, and 16 nodes. First, we extracted the PoIs for each of the 100 users based on the node configuration using the ACCIO method [PMB+18]. Then, we computed the distribution rate of PoIs per node for each configuration. This analysis provides insights into how malicious nodes or attacks could potentially compromise user data by simulating scenarios where an attacker gains insights from data within a compromised node.

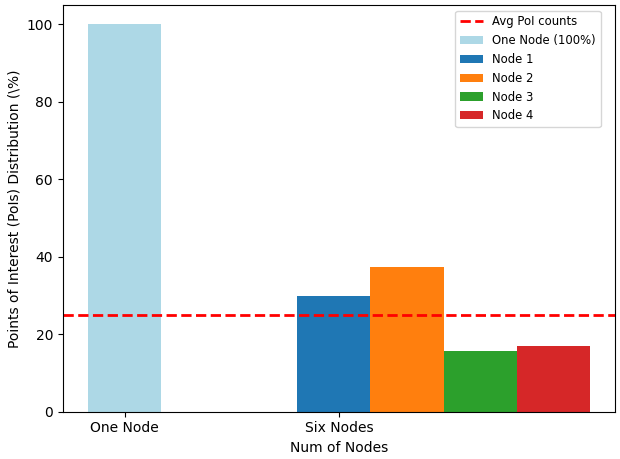


Figure 5.8: Distribution of average PoI counts across a single-node and a four-node configuration. In the single-node setup, all PoI data is centralized, resulting in a 100% concentration in one node. When the data is distributed across four nodes, the PoI counts are more evenly spread. The red dashed line indicates the average PoI percentage per node.

The bar charts in Figure 5.8 and Figure 5.9 depict the distribution of PoI across various node configurations—specifically, one node, four nodes, and nine nodes. For the single-node configuration, all PoI data is centralized in that single node, resulting in a 100% PoI count for that node. This concentration of data, while straightforward, introduces significant privacy risks as all user activity is aggregated in one easily targetable location.

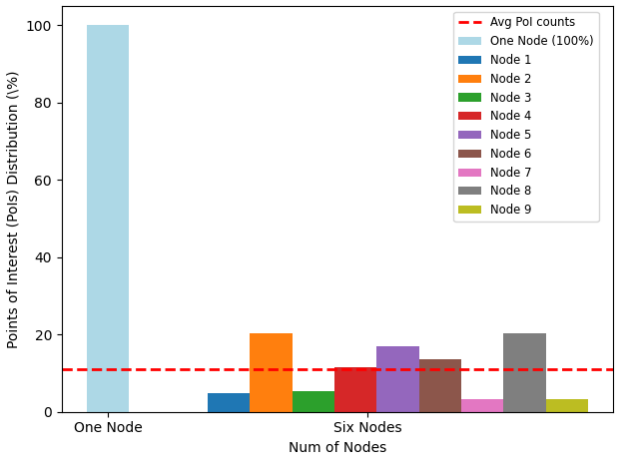


Figure 5.9: Distribution of average PoI counts between a single-node and a nine-node configuration. In the single-node setup, all PoI data is centralized, leading to a 100% concentration within one node. In the nine-node configuration, the PoI counts are more evenly distributed across the nodes. The red dashed line represents the average PoI percentage per node.

As the number of nodes increases to four, the distribution of PoI data among these nodes results in significant variability in the amount of data captured by each one. Some nodes gather a substantially larger proportion of PoIs than others, leading to an uneven distribution. The red dashed line indicates the average PoI count of 25% across the four nodes, emphasizing this disparity, as some nodes surpass the average while others fall short. This imbalance implies that certain nodes may still harbor considerable amounts of sensitive information, making them potential targets for attacks. While this analysis is based on simulated data, it offers valuable insights into potential real-world scenarios.

In the nine-node configuration, the distribution of PoI data results in each node capturing a smaller share of the total PoI count. The average PoI percentage per node is approximately 15% lower than in the four-node setup, with most nodes' percentages clustering around this average, indicating a more even distribution of users data. This improved distribution enhances privacy by minimizing the amount of data held by any single node, thereby reducing the risks associated with potential data breaches or attacks. However, some nodes still capture slightly more PoIs than others, demonstrating that increasing the number of nodes effectively lowers the risk of PoI attacks.

These charts highlight the trade-offs between privacy and the complexity of node configurations in a federated learning system. While increasing the number of nodes generally results in a more balanced and privacy-preserving data distribution, it also necessitates more sophisticated

strategies for data aggregation and management. Additionally, the costs associated with edge devices and network communications must be considered.

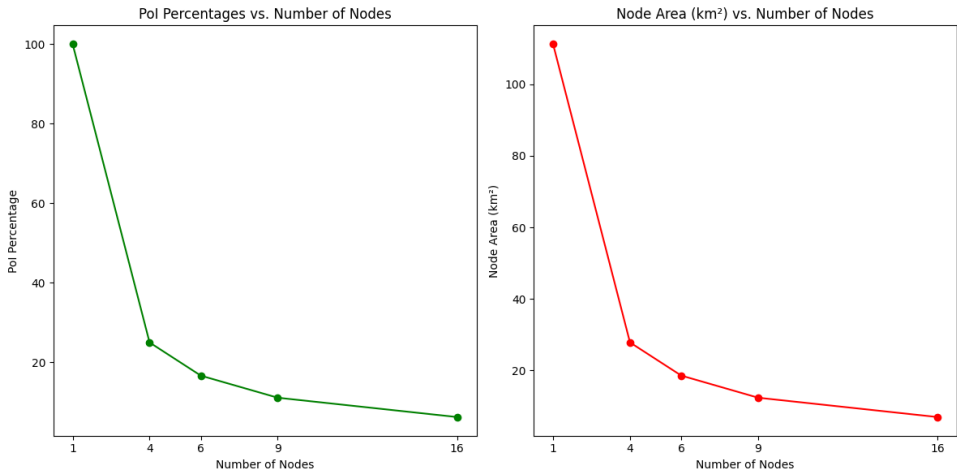


Figure 5.10: displays two subgraphs illustrating the relationship between the number of nodes, the average percentage of glspoi counts, and the area per node. The left subgraph shows how the average percentage of glspoi counts varies with the number of nodes. The right subgraph illustrates the area covered by each node under different segmentation schemes, ranging from approximately 111 km² to around 7 km².

Impact of Node Count on PoI Distribution and Spatial Coverage: Figure 5.10 illustrates the relationship between the number of nodes, the average percentage of PoI counts, and the area covered per node. The figure is divided into two subgraphs: the left subgraph displays the average percentage of PoI counts relative to the number of nodes, while the right subgraph shows the area per node across different segmentation schemes.

The left subgraph indicates that the average percentage of PoI counts is directly proportional to the number of nodes. This means that as the number of nodes increases, each node tends to handle a smaller proportion of the total PoI data, leading to a more distributed data load.

In the right subgraph, the area per node is depicted across various segmentation configurations, ranging from approximately 111 km² to around 7 km². This variation demonstrates how different segmentation schemes impact the spatial area assigned to each node. As the number of nodes increases, the area covered by each node decreases, enhancing the granularity of data distribution and improving localized insights. However, this reduction in area per node can also decrease the number of volunteers contributing data, potentially impacting privacy by increasing the risk of individual re-identification. While our analysis primarily focuses on PoI attacks, having fewer users per node may heighten privacy risks, underscoring the importance of considering a wider range of privacy threats in future research.

These visualizations underscore how increasing the number of nodes affects both the distribution of PoI counts and the spatial coverage per node. As the node count grows, each node captures a smaller share of PoI data and covers a smaller area, which can lead to a more balanced and detailed distribution of data across the network.

horizontal and vertical segmentation:

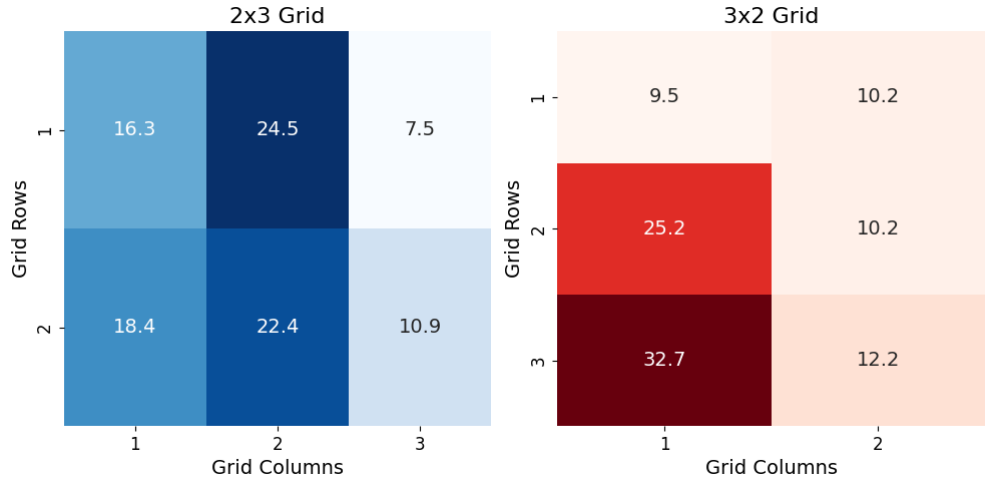


Figure 5.11: Impact of Node Count on PoI Distribution and Spatial Coverage. The left heatmap visualizes the average percentage of PoI contributions per node in a 2x3 grid configuration. The right heatmap depicts the spatial coverage per node in a 3x2 grid configuration, highlighting variations across different node configurations.

We investigated the effects of horizontal and vertical segmentation on data distribution using six nodes arranged in two different grid configurations: 2 rows by 3 columns and 3 rows by 2 columns. By analyzing datasets from 30 randomly selected users with varying geometric paths, we examined how PoIs are distributed across these nodes in different grid dimensions, while maintaining the same total number of nodes. The goal was to understand how different grid structures impact data handling and distribution.

Given that cities are inherently non-homogeneous, with some areas experiencing more concentrated activity than others, this segmentation revealed uneven distributions of PoIs across the nodes. Figure 5.11 illustrate the PoI distribution across the 6 nodes in both configurations, showing higher concentrations in central urban areas, which align with the presence of shops, libraries, parks, and other activity centers.

A heatmap could visualize PoI intensity and frequency across different grid areas, highlighting regions with high concentrations and providing a clearer view of PoI distribution. Additionally, using non-symmetric node segmentation to analyze high-density areas, such as downtown, could reveal detailed patterns and imbalances not apparent with uniform segmentation.

Integrating real data and asymmetric segmentation could improve this analysis, offering a detailed view of the PoI distribution and addressing data privacy and distribution issues. We discussed various topology for examining different levels of user privacy in section 5.2. The following segmentation levels were considered:

- **Zone**: Analyzing PoI counts under various conditions suggests that if a malicious actor gains access to a node, they could obtain a fraction of user information. To mitigate this, a mechanism can be implemented to remove geodata when users enter a zone, sending only raw data to edge devices.

- **Sub-Zone**: This approach is useful for high-density and complex urban areas and requires more resources and communication. Privacy levels are determined by the number of sub-zones on average. Strategies for removing location data can also be applied.

- **Street Grid**: This topology is particularly suited for high-density urban areas with complex building layouts. It involves truncating the last digits of latitude and longitude coordinates to ensure that users are aligned within defined streets. For example, precision levels can vary significantly, ranging from ±10 centimeters (6 digits) to ±1 kilometer (2 digits). Due to constraints in accessing a free service map containing street names and mapping information, we were unable to provide statistics for the extracted PoIs.

- **Pinpoint Location**: This approach involves users sharing their exact time and location with the corresponding nodes, providing the finest level of detail but with the least privacy. As each node covers a larger area, a malicious actor or attacker could potentially access a greater portion of user trajectories, increasing the risk of privacy breaches.

5.4.2 Model Evaluation

In evaluating model performance, we considered local models, the global model, the stability of different node configuration on the performance, and long-term forecasting. We kept certain hyper-parameters constant: the learning rate was set to 0.0001, the batch size was 48, and the temporal mechanism was implemented as described in section 3.2.

Local Models Assessment: In this study, firstly, we evaluated the performance of our neural network model across different node configurations within a federated learning (FL) setup. We analyzed the model's training and validation performance using standard metrics across three distinct datasets: minute-level, quarterly, and hourly data. The focus was on MAE and RMSE over 50 epochs to determine how effectively the local models, trained independently on their respective datasets, could generalize to unseen validation data before aggregation. We selected a configuration of 4 nodes to align with the previous chapters, while additional plots representing a finer scale with 12 nodes are provided in the appendices for further analysis.

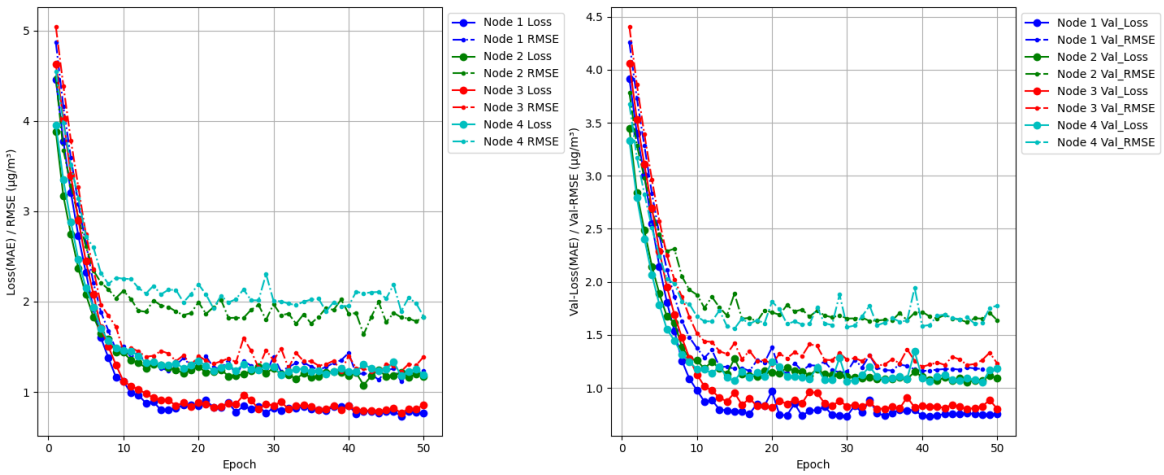


Figure 5.12: Training and validation loss (measured by MAE) and error (represented by RMSE) across 50 epochs for four federated learning nodes using hourly datasets. The left panel illustrates the convergence of training loss over time, while the right panel highlights the variability in validation performance among the four node configurations.

Figure 5.12, Figure 5.13, and Figure 5.14 show the comparison of loss (MAE) and error (RMSE) for local models across hourly, quarterly, and minutely datasets. Initially, a rapid decline in MAE and RMSE is observed, indicating effective early learning, followed by a convergence phase where loss values stabilize, suggesting that the models have captured most data patterns.

Some nodes consistently outperformed others, exhibiting lower final loss values and more stable validation metrics. This suggests a better understanding of local data distribution and generalization within the federated learning (FL) framework. However, certain nodes displayed greater variability in validation loss, reflecting differences in the spatial or temporal characteristics of the data. For example, a node capturing data from an industrial area with higher pollution variability may show higher error rates, while a node in a stable residential area might demonstrate lower errors. The maximum error across all nodes was 1 $\mu g/m^3$ for hourly datasets and 0.5 $\mu g/m^3$ for quarterly and minutely datasets, both in loss and validation loss values. This variability also underscores the challenges in capturing finer temporal patterns, particularly in granular datasets like the minutely dataset.

It's crucial to consider the number of samples per node: for the hourly and quarterly datasets, each node nearly contains the complete datasets, with $[717, 717, 702, 717]$ samples for hourly and $[2750, 2877, 2877, 2877]$ for quarterly (out of 2877 samples). In contrast, the minutely datasets have fewer complete samples per node, with [39845, 43185, 43170, 42557] out of 43200 samples for the month.

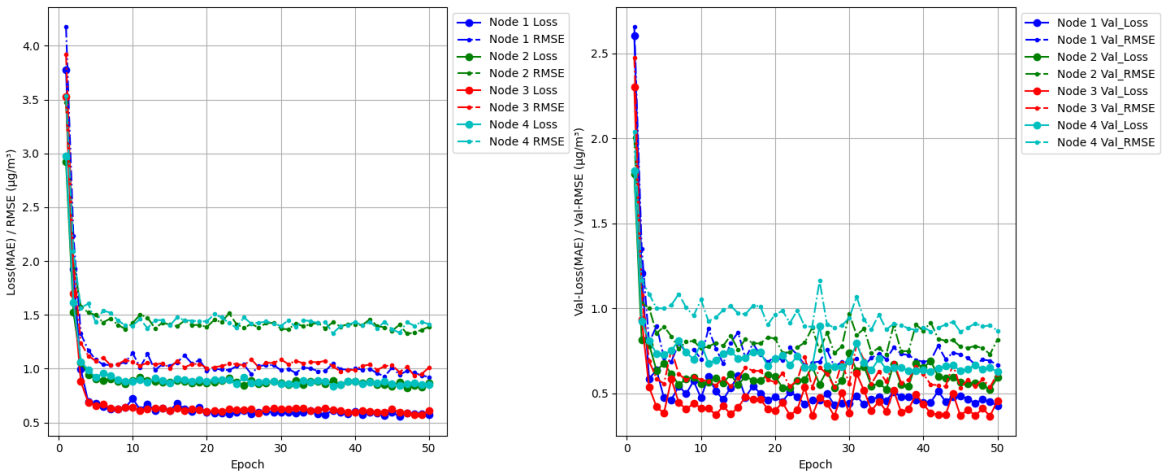


Figure 5.13: Comparison of loss (measured by MAE) and error (represented by RMSE) for local models across four nodes, trained with quarterly datasets. The left panel depicts the consistent reduction in loss over time during training, while the right panel emphasizes the variability in validation performance among the four nodes.

The node with missing data may experience slight overfitting on the smaller, simplified dataset, leading to deceptively lower error rates during training and validation. In contrast, nodes with complete datasets encounter a more diverse range of examples, resulting in slightly higher errors but improved generalization to unseen data. For instance, Node 1 in the minutely dataset (see Figure 5.14) exhibits the lowest error; however, it lacks the variation necessary to capture the data's complexity fully. In our analysis, a data limitation—around 30% or more—can lead to significant errors, as demonstrated by Node 1 with 515 samples in the hourly dataset (depicted in Figure A.1) or Node 2 with 1,942 samples in the quarterly dataset (shown in Figure A.2) and in the appendices.

The performance differences between nodes underscore the inherent challenges in FL, where local data heterogeneity can significantly affect the overall effectiveness of the model. These findings highlight the need for improved aggregation strategies or local model training processes to ensure effective contributions to the global model, particularly for datasets with finer temporal resolutions. Notably, minutely data exhibited more fluctuations in convergence, especially in validation data, though the variation range is approximately half that observed in hourly datasets.

We trained the models for 50 epochs with full participation of nodes for both training and evaluation. We evaluated convergence by examining the training process in a single round of FL.

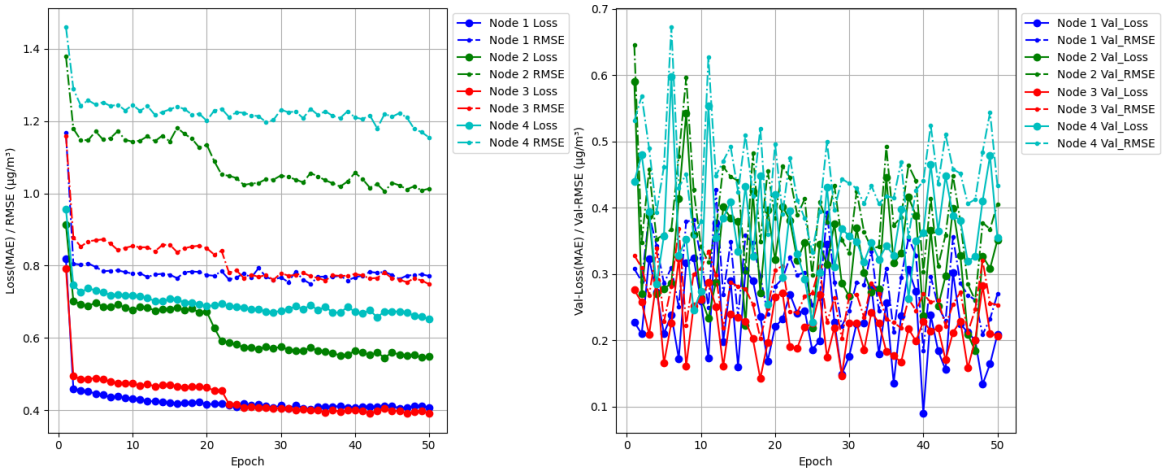


Figure 5.14: Comparison of loss (measured by MAE) and error (represented by RMSE) for local models across four nodes, trained with minute-level datasets. The left panel illustrates the consistent reduction in loss over time during training, while the right panel highlights the variability in validation performance across the nodes.

Global Model Assessment: Evaluating the global model's performance after training local models is crucial for understanding its effectiveness. Several factors can influence the global model's performance, including the number of distributed models, the number of contributing nodes, the number of evaluation rounds, and the choice of hyper-parameters. In our study, we controlled several parameters while focusing on key variables central to our project. Specifically, we investigated the impact of the number of evaluation rounds and the number of local models (nodes).

For evaluating the global model, we utilized a greater number of nodes to more effectively assess the impact of participating nodes on the global model. We examined two scenarios based on six nodes: one scenario utilized a pre-trained model (see Figure 5.15), while the other employed a trainable model (refer to Figure 5.16). Both scenarios were evaluated over 10 rounds of global model evaluation.

Figure 5.15 shows the performance with full and half contributions of the total nodes. Each round consists of training local models, aggregating them, updating the global model, and redistributing the updates. An additional round is included to initialize the network weights. The pre-trained model parameters lead to a significant reduction in error in the second round, with local models quickly achieving very error low values. In the full node contribution scenario, the error metrics stabilize with minimal fluctuations after the initial drop, indicating effective learning and convergence to lower error values, particularly after four rounds. In contrast, the half-node contribution scenario, where only half of the nodes contribute randomly, shows similar

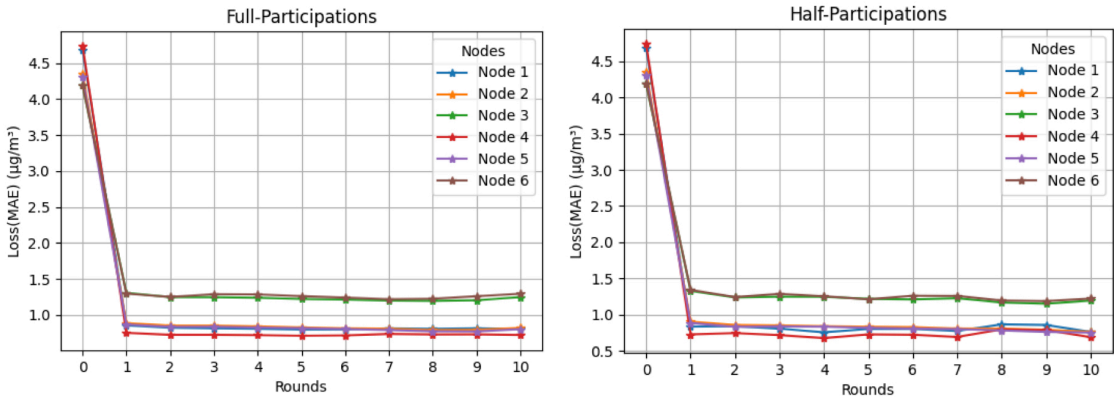


Figure 5.15: Loss (MAE) of the global model measured over 10 rounds of evaluation, with full and half contributions from all six nodes, using a pre-trained model. The updated global model, sent back by the server to the local models at each round, is used to test the performance of the local models on their respective local test sets.

trends but with slightly higher and more variable error metrics. This variability suggests reduced data diversity from fewer contributing nodes, affecting the global model's ability to generalize effectively. Both scenarios display a sharp decline in error during the first round, demonstrating that the global model quickly adapts to aggregated data after initialization with the pre-trained model.

In Figure 5.16, we conducted a similar examination with a trainable model. The results show a notable trend where the global model also quickly adapts to aggregated data, though with variations for two rounds. With full contributions from all six nodes, the model exhibits more stability after 10 rounds, while the half-node contribution scenario displays considerable error variability and fluctuation, indicating that reduced data diversity from fewer nodes impacts generalization.

Overall, the pre-trained model with half-node participation performs similarly to the trainable model with full-node participation, providing a cost-effective solution for resource management in federated learning (FL). However, the slightly fluctuations in the half-node scenario indicate that missing data from some nodes impacts model robustness. These findings suggest that a reduced number of contributing nodes can result in less effective learning and increased generalization error, which is a critical consideration in the design of FL systems. Additionally, the total error across all nodes is lower in the pre-trained model, even with only half of the nodes participating.

We trained the models for 50 epochs using hourly datasets. Convergence was evaluated by examining the training process with a learning rate of 0.0001 and a batch size of 48.

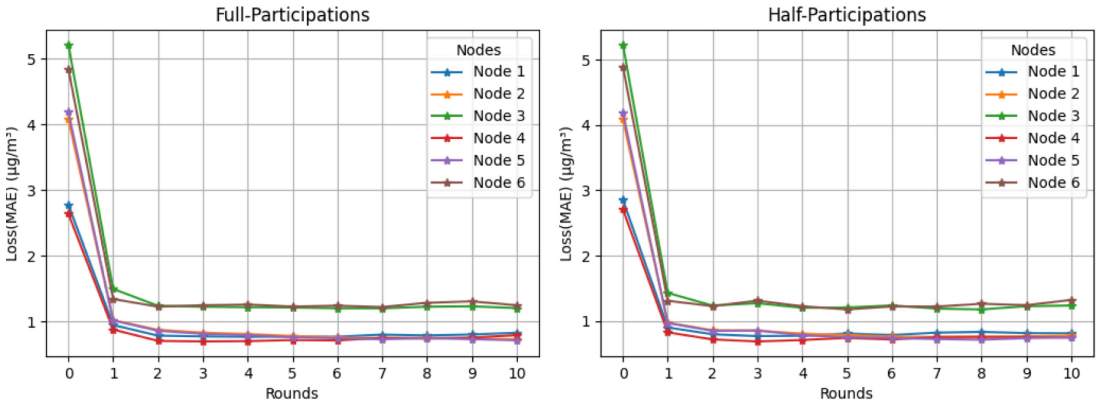


Figure 5.16: Loss (MAE) of the global model measured over 10 rounds of evaluation, with full and half contributions from all six nodes and with a trainable model. The updated global model, sent back by the server to the local models at each round, is used to test the performance of the local models on their respective local test sets.

Scalability Assessment:

Achieving an optimal balance between node count and data distribution is crucial for both model effectiveness and user privacy protection. In this study, while the total number of users remains fixed at 100 (with 80% for training and 20% for testing), we assessed the model's scalability by distributing the data across varying numbers of nodes. This approach ensures stable performance as the network topology changes and data is segmented differently. Even with a constant number of users, adjusting the number of participating nodes allowed us to examine how data distribution impacts model accuracy and privacy.

Table 5.1: Evaluation of Forecasting Metrics for Varying Numbers of Nodes in 1-Hour Forecasting)

Nodes	**MAE**	**MSE**	**RMSE**	**WMAPE (%)**	**R^2 Score**
2 Nodes	1.398	5.145	2.264	20.724	0.903
4 Nodes	0.878	2.180	1.453	21.975	0.870
6 Nodes	0.910	2.792	1.630	22.784	0.854
8 Nodes	0.998	3.263	1.784	25.004	0.810
10 Nodes	1.068	3.710	1.854	27.400	0.778
12 Nodes	1.056	3.829	1.907	26.348	0.777
14 Nodes	1.112	4.312	1.990	29.739	0.721
16 Nodes	1.564	8.978	2.578	34.058	0.562

Table 5.1 presents an evaluation of forecasting metrics across varying node counts for 1-hour air pollution forecasting. A clear trend emerges as the number of nodes increases, the model's performance improves up to 4 or 6 nodes, after which the accuracy declines.

With 4 nodes, the model achieves its best performance, with the lowest MAE (0.878), MSE (2.180), and RMSE (1.453), along with a strong R^2 score of 0.870. This indicates that the model can explain 87% of the variance in the data, providing accurate predictions with minimal error. As the number of nodes increases to 6, performance remains competitive, with a slight increase in error, reflected by an MAE of 0.910 and an RMSE of 1.630, but with a comparable R^2 score of 0.854.

However, after 6 nodes, the performance starts to degrade. At 8 nodes and beyond, the MAE, RMSE, and MSE steadily increase, indicating that the model struggles to maintain accuracy as the data becomes more fragmented. For example, at 16 nodes, the MAE increases to 1.564, and the RMSE reaches 2.578, while the R^2 score drops significantly to 0.562. This suggests that the distribution of data across more nodes introduces greater complexity and error, reducing the model's effectiveness.

In summary, the optimal balance appears to be between 4 and 6 nodes, where the model performs efficiently. Beyond that, increasing the number of nodes leads to diminishing returns in terms of accuracy, likely due to over-segmentation of the data.

We prepared hourly datasets from 80 users for training over one month. The local models were trained for 50 epochs using the hourly datasets and pre-trained weights. Convergence was assessed by monitoring the training process, employing a learning rate of 0.0001 and a batch size of 48. The global model was evaluated across all node test sets, comprising data from 20 users, through three rounds of evaluation, with full participation from all nodes. We structured the nodes to form as close to a square shape as possible. When this wasn't feasible, horizontal splitting was prioritized over vertical segmentation.

Figure 5.17 visualizes the predicted versus actual values for the test sets across a four-node segmentation. The blue solid line represents the actual values, while the orange dotted line illustrates the predicted values for one month. The close alignment of the lines demonstrates the model's accuracy in capturing temporal trends across different nodes.

Long-Term Predictions

We performed an analysis of long-term forecasting using the same performance metrics across four different node configurations. This segmentation was designed to ensure each node had a sufficient amount of data for robust testing. The datasets were derived from 20 out of 100 users, creating four-node datasets based on user contributions. All nodes had complete datasets, except for one node, which had less than 3% missing values; this was handled using the *fillna()* method. The models were trained for 50 epochs, with convergence monitored through the training process, utilizing a learning rate of 0.0001 and a batch size of 48. All local models participated actively in both the training and evaluation phases, providing a comprehensive

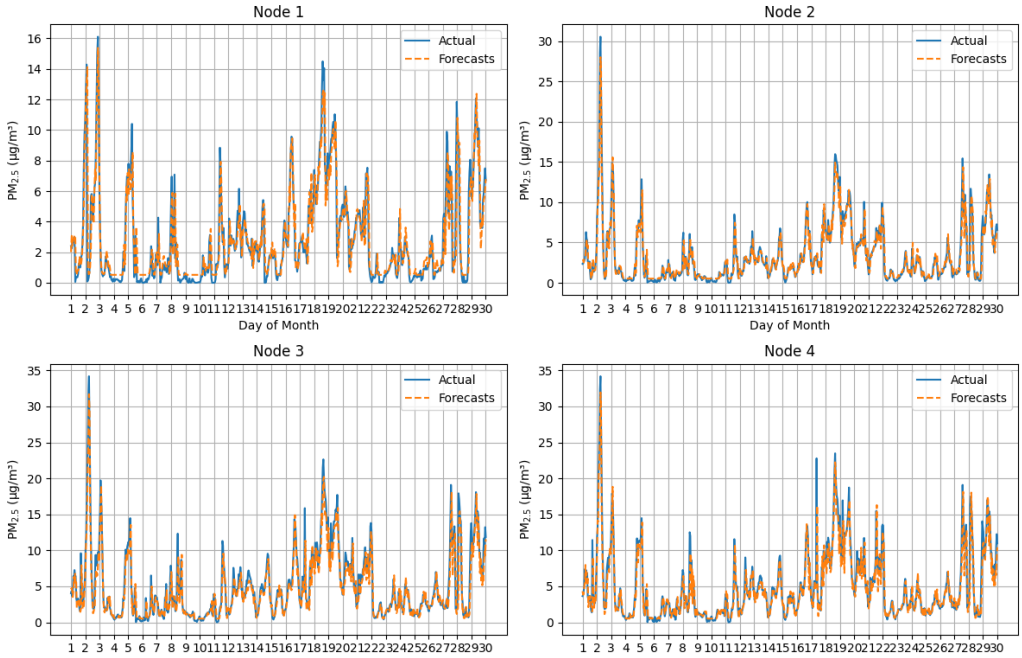


Figure 5.17: Comparison of actual versus predicted values for a four-node segmentation over a one-month period. The blue solid line represents the actual hourly data, while the orange dotted line indicates the model's predictions.

assessment of model performance across the various nodes, each predicting locally with its distinct dataset.

The analysis of forecasting performance across different nodes provides valuable insights into the model's generalizability and accuracy over varying prediction horizons. By examining multiple metrics, we gain a clearer understanding of how the model's effectiveness changes as the forecast duration extends from 1 hour to 24 hours. Figure 5.18 illustrates this experiment, highlighting the variation in error and accuracy metrics across all nodes over time.

The results underscore the challenge of maintaining consistent performance across nodes and forecast horizons. While the model performs well for short-term forecasts, errors increase notably for longer-term predictions, particularly at specific nodes. This suggests that node-specific factors, such as localized environmental conditions, may influence prediction accuracy over extended periods.

These findings show that while the model is effective in some scenarios, its scalability across different nodes and forecast durations is limited. For practical applications—especially where nodes exhibit higher error rates—further model refinements may be needed, including enhanced data pre-processing, feature engineering, or alternative modeling techniques.

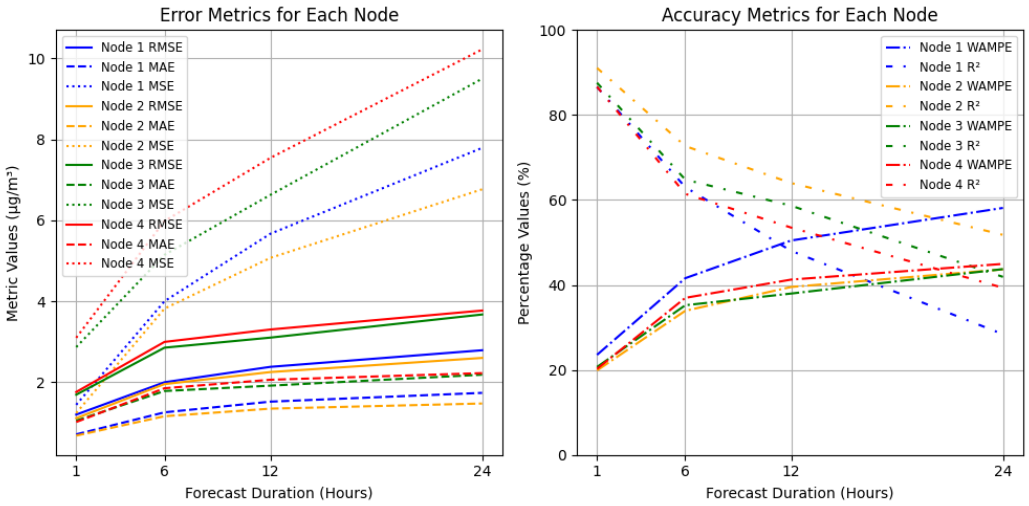


Figure 5.18: Evaluation metrics for four-node configuration to forecast long term values

In summary, the analysis demonstrates that the model is particularly effective for short-term forecasts, with error rates rising as the forecast horizon lengthens. This trend highlights the close link between prediction accuracy and forecast duration, with shorter forecasts producing more reliable results. Given that only a one-month simulated dataset was used, it is crucial to consider forecast duration when evaluating the model's overall performance.

5.5 Discussion

5.5.1 Privacy Analysis

In this section, we analyze the impact of various model characteristics—such as grid segmentation and users data distribution on user's privacy and model performance.

Specifically, we assess user privacy through PoI attacks, focusing on how different node configurations (1, 4, 9, and 16 nodes) affect privacy. In a single-node configuration, all PoI data is concentrated in one node, which presents significant privacy risks due to the aggregation of all user activities in a single location. As the number of nodes increases, such as in four-node and nine-node configurations, the distribution of PoI data becomes more balanced. However, certain nodes may still capture a disproportionate amount of PoIs, leaving residual privacy concerns. In general, increasing the number of nodes tends to distribute PoI data more evenly, enhancing privacy by reducing the sensitive information stored on any individual node. This shift highlights a trade-off between data distribution and privacy, emphasizing the need for sophisticated data management strategies as the number of nodes grows.

Our analysis further shows that as the number of nodes increases, the PoI distribution becomes more uniform, with each node covering a smaller geographic area.

This leads to reduced spatial coverage per node, which could affect the resolution of the data analysis. We also explored the effects of different grid configurations, such as 2x3 and 3x2 grids, on PoI distribution. Integrating asymmetric grid segmentation help address privacy concerns more effectively while providing deeper insights into PoI distribution, particularly in densely populated nodes.

5.5.2 Model Performance Evaluation

We evaluated the performance of our model by examining both local and global aspects across various node configurations and data segmentation schemes. Our analysis focused on four key areas: local model assessment, global model performance, scalability, and long-term predictions.

Local Model Assessment: Training on different datasets (hourly, quarterly, and minutely) consistently showed a reduction in MAE and RMSE during the early epochs, reflecting effective learning in the local models. Models trained on finer temporal intervals, such as minute-level data, converged more quickly and resulted in lower final errors after 50 epochs compared to those trained on coarser intervals, like hourly data. Despite this, validation performance varied significantly, especially with minute-level data. This variability underscores the challenges posed by diverse local data distributions in FL. Inconsistencies in error values across nodes, particularly with minute-level data, highlight the need for better aggregation strategies to manage

local data heterogeneity. Furthermore, the superior performance of certain nodes emphasizes the importance of optimizing data distribution and refining global aggregation methods to enhance overall model accuracy.

Global Model Assessment: The global model was evaluated over 10 rounds, comparing pre-trained and trainable models. The pre-trained model exhibited faster convergence and greater stability in error metrics, whereas the trainable model showed more variability, particularly when fewer nodes participated in evaluation rounds. These results indicate that pre-training enhances initial model stability and convergence. Although, limited node participation reduces data diversity, which impacts the model's generalization ability. This emphasizes the importance of ensuring adequate node contributions in global aggregation to maintain model robustness.

Scalability Assessment: We analysis of prediction metrics across distinct test sets for varying node counts (ranging from 2 to 16 nodes) using the global models that distribute to the local models. The results revealed that model accuracy improves up to 6 nodes, based on the specific dataset of 100 users (80 for training and validation and the rest 20 users for testing) used in this study. It's important to highlight that this relationship between node count and performance is closely linked to the number of users. For a larger or smaller user base, the optimal number of nodes may vary, potentially resulting in different performance outcomes and trends.

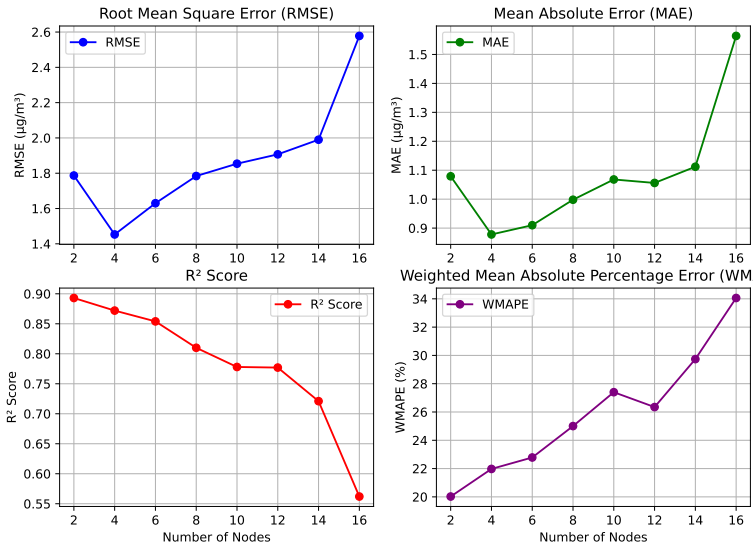


Figure 5.19: Evaluation of our model's performance with four metrics: RMSE, MAE, R^2, and WMAPE on an unseen dataset across different node configurations.

We demonstrated that privacy improves as the number of nodes increases. By plotting model performance metrics, we expect that a greater number of nodes generally enhances the model's accuracy. While the optimal number of nodes primarily depends on the user count and

spatial distribution, it also significantly affects prediction accuracy. These insights can guide the selection of node configurations to achieve an optimal trade-off between performance and complexity in forecasting tasks. Figure 5.19 visually represents the performance of different node configurations and their evaluation metrics.

The results illustrate how various performance metrics (RMSE, MAE, R^2 score, and WMAPE) behave as the number of nodes increases from 2 to 16. Overall, all metrics show an acceptable trend in the error range. Both the RMSE and MAE metrics decline from 2 to 6 nodes, indicating significant improvements in prediction accuracy with the initial increase in nodes. However, beyond 6 nodes, these metrics begin to rise, suggesting diminishing returns in performance gains. This finding indicates that configurations with 4 to 6 nodes may optimize the minimization of prediction errors in our case.

Conversely, the R^2 score steadily decreases as the number of nodes increases, implying that while errors are initially reduced, the model's ability to explain the variance in the data diminishes. This decline could be due to over-fitting in some nodes or imbalances in data distribution. The WMAPE graph further supports this observation, peaking at 4 nodes before showing only marginal improvements with additional nodes. Therefore, increasing the number of nodes beyond 6 adds complexity without significant performance gains and may reduce the model's generalization capability in our study. Figure A.4 in the appendices illustrates the correlation between the true and forecasted values for the 4-node configurations.

Long-Term Predictions: Long-Term Predictions: As discussed in the previous chapter, while the model performed well for short-term forecasts, it struggled with longer prediction horizons. This was anticipated, particularly given the challenges associated with LSTM models. As the forecast duration increased, performance declined, indicating a strong correlation between model accuracy and prediction length. This finding underscores the necessity for improved data pre-processing and modeling techniques to effectively manage complex or less predictable data. Our analysis of long-term predictions across various node configurations suggests that meticulous calibration and strategic data distribution are essential for sustaining performance during extended forecasting periods.

Data Distribution and Node Configuration: The performance of the model is significantly influenced by the distribution of data across various node configurations. As the number of nodes increases, the percentage of complete datasets per node decreases, leading to performance degradation when fewer complete time series are available. Figure 5.20 illustrates the distribution of data across different node structures; the blue bars represent the percentage of completed datasets over one month, while the green bars indicate the percentage of datasets with more than 95% completion across all nodes for each configuration. This relationship between data quality and model performance confirms that reduced data quality is associated with lower predictive

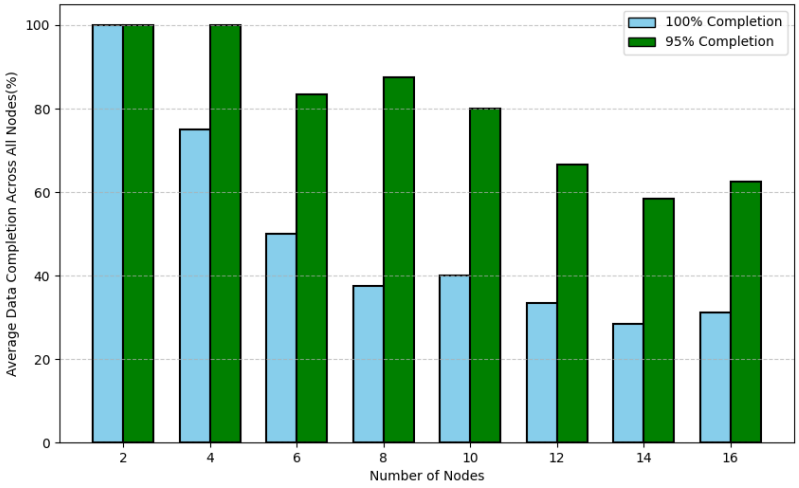


Figure 5.20: Distribution of User Data Across Different Node Configurations and the Average Percentage of Completed Time Series Data per Node Configuration.

accuracy. With a training dataset of 80 users, our study identifies the configuration of 4 to 6 nodes as the most optimal setup for local models in Dijon.

Overall, our evaluations highlight the trade-offs between model accuracy, data distribution, and user privacy. Effective performance management requires balancing node count, prediction horizon, data quality, and availability to achieve optimal results. Future improvements should focus on enhancing data distribution, optimizing node contributions, and exploring advanced modeling techniques to improve both short-term and long-term forecasting accuracy while minimizing privacy risks.

This research acknowledges several drawbacks associated with data collection, particularly in remote or less populated areas, as well as during nighttime. Gathering data in such locations can be challenging due to limited user participation and environmental factors that may affect sensor readings. Additionally, requiring users to carry sensors daily may lead to inconsistencies in data collection, as not all users might be willing or able to participate consistently.

Another critical concern is the potential risk of re-identification of users who regularly collect data. This risk underscores the importance of implementing robust privacy measures to safeguard personal information while ensuring the reliability of the data collected. Therefore, while the research provides valuable insights, it is essential to address these limitations to enhance the overall effectiveness and security of future data collection efforts.

Chapter 6

CONCLUSION

The motivation for this project arises from the growing health concerns linked to fine $PM_{2.5}$, particularly in urban areas where pollution levels are elevated due to dense human activity. $PM_{2.5}$ poses severe health risks because it can penetrate deep into the lungs and enter the bloodstream, causing chronic respiratory and cardiovascular issues. Given the pressing need to monitor and manage air quality effectively in urban settings, this research was initiated to develop a robust model capable of accurately forecasting $PM_{2.5}$ concentrations.

Urban air pollution is not only a public health risk but also a complex environmental challenge that demands precise, real-time data for effective management. While existing air quality monitoring systems provide valuable insights, they often fall short in spatial and temporal resolution, which limits their ability to capture localized pollution spikes or short-term fluctuations. Traditional monitoring stations, though effective, are expensive and offer limited coverage, underscoring the need for more scalable and cost-effective solutions. This project aims to address these gaps by leveraging advanced machine learning techniques and spatiotemporal modeling to create a more accurate and comprehensive air pollution forecasting system.

This thesis presents a detailed exploration of advanced models for air pollution prediction, focusing on spatiotemporal forecasting and data privacy considerations. The findings across the three chapters offer significant theoretical contributions and practical applications for improving environmental monitoring systems and ensuring the privacy of collected data.

6.1 Temporal Model

We initially focused on capturing the temporal characteristics of air pollutants and designed an appropriate machine learning model. The PMFORECAST model represents a significant advancement in air quality forecasting, particularly in predicting $PM_{2.5}$ concentrations. By leveraging a self-adaptive LSTM architecture enriched with temporal attention mechanisms, PMFORECAST effectively captures temporal dependencies, improving both short-term and

long-term prediction accuracy. The model's design—encompassing pre-processing, temporal attention, prediction horizon adjustment, and LSTM layers—ensures robust performance, making it a highly adaptable and dynamic solution for real-time air quality monitoring.

Empirical evaluations using data from the QAMELEO network in Dijon, France, demonstrated PMFORECAST's superior predictive accuracy, achieving 99.7% accuracy for 1-hour forecasts and 73.5% for 12-hour forecasts. Compared to other forecasting methods, including GPR, GRU, XGBoost, ARIMA, and standard LSTM models, PMFORECAST consistently outperforms them in precision, scalability, and computational efficiency. The model's ability to reconfigure efficiently for different forecast horizons further enhances its adaptability.

In addition to excelling in single-task predictions, PMFORECAST also performs exceptionally well in multi-task forecasting, predicting not only $PM_{2.5}$, but also PM_1, PM_{10}, temperature, and humidity, with correlations exceeding 98%. While the single-task model achieves marginally higher accuracy for $PM_{2.5}$ prediction, the multi-task approach proves invaluable for real-world applications requiring comprehensive air quality monitoring. Furthermore, PMFORECAST 's computational efficiency—particularly in pre-processing and its ability to provide low-latency predictions—makes it suitable for time-sensitive scenarios.

Ultimately, PMFORECAST stands out as a robust, versatile, and efficient solution for particulate matter prediction. With its scalable architecture and superior forecasting performance, PMFORECAST has significant potential for deployment in large-scale air quality monitoring systems, contributing to more informed and timely responses to air pollution challenges. However, the current model is limited to its own data and one region, without accounting for the spatial characteristics of pollutants. This limitation underscores the need to further explore spatiotemporal data to improve the model's predictive power across different regions.

6.2 Spatio-Temporal Model

We were motivated to develop a spatiotemporal model in response to the limitations observed in our first contribution. Specifically, our study introduces a model designed to accurately forecast $PM_{2.5}$ concentrations across urban regions by leveraging advanced machine learning algorithms. By integrating GCN to capture spatial dependencies and LSTM networks to model temporal patterns, the resulting GT-LSTM model offers precise predictions of air pollution levels. This dual focus on spatial and temporal patterns provides a comprehensive understanding of both temporal and spatial dependency across multiple regions of urban areas.

In chapter 4, the novel GT-LSTM model was presented for predicting $PM_{2.5}$ concentrations in suburban environments using fixed low-cost sensors and AI techniques. The model demonstrated its utility for real-time monitoring and long-term air quality planning by relying

on correlations between datasets across regions, utilizing Pearson's algorithm, instead of more complex multi-scale dependencies.

Our analysis highlights the competitive predictive capabilities of the GT-LSTM model, with high accuracy scores and a strong correlation between forecasting and observed values. By factoring in local pollution sources, neighboring influences, and historical data, the model enables proactive responses to pollution trends. The inclusion of spatiotemporal dependencies ensures not only accuracy but also relevance, capturing the dynamics of pollutant dispersion in various urban contexts.

A key strength of the GT-LSTM model is its ability to predict values even when faced with sparse observational data or in instances where actual measurements are unavailable. This capability allows the model to generate acceptable predictions without relying on multi-scale data, making it particularly valuable for regions with limited monitoring infrastructure. This flexibility highlights the model's potential for scalable deployment across diverse urban settings, enhancing both the scope and efficiency of air pollution monitoring systems.

The GT-LSTM model addresses the complexities of air pollution forecasting by capturing both spatial and temporal dependencies, ensuring high accuracy even in data-scarce environments. The chapter also highlights limitations in far long-term predictions, suggesting that future research could explore hybrid approaches to improve forecasting accuracy. Additionally, the number of monitored regions is constrained by the fixed sensor stations. To enhance data granularity, accuracy, and spatial coverage, integrating mobile sensor networks and crowdsourcing resources is recommended, which would provide more detailed insights and better resolution for air quality monitoring and forecasting.

6.3 Federated Learning Framework

In this chapter, we proposed a FL framework called FEDAIRNET that integrates mobile sensors with fixed air quality stations to achieve flexible resolution in urban areas while safeguarding user privacy. Although fixed stations offer reliable data, their dependence on the number and availability of sites can interrupt the system or affect its performance. While mobile sensors provide greater flexibility, they also raise privacy concerns for users collecting data. FL addresses this challenge by enabling model training across decentralized devices without the need to transfer raw data to a central server. This approach preserves user privacy while facilitating real-time, high-resolution air quality forecasting in urban environments. Mobile sensors deployed throughout a city can capture localized pollution data on a finer scale, contributing to a global model that delivers accurate air quality forecasts, all while mitigating the risks of re-identification and exposure of personal data within the FL system.

Due to the lack of volunteers in the Apolline project, we simulated data from 100 users over one month, creating different trajectory-based scenarios at three granular levels: hourly, quarterly, and minutely. We employed the same neural network architecture used in , utilizing a pre-trained version of the PMFORECAST model to forecast $PM_{2.5}$ concentrations. The city was divided into regions, referred to as nodes, with each node receiving a portion of user data based on the users' geographic paths. This approach enabled us to simulate the distribution of pollution data across various regions and evaluate the model's performance under diverse user mobility patterns.

We evaluated the impact of various model characteristics, such as grid segmentation and data distribution, on both privacy and model performance. Our analysis of point-of-interest (POI) attacks using the ACCIO algorithm across different node configurations revealed significant privacy implications. Aggregating user activities in a single-node configuration posed substantial privacy risks. However, increasing the number of nodes helped distribute the POI data more diffused throughout the nodes, thereby enhancing privacy. Despite this improvement, we can face certain configurations still exhibited disproportionate PoIs capture. This is highlighting the need for advanced data management strategies. Potential approaches include using non-symmetric nodes, implementing street grid mapping, or modifying geo-data by removing the last digits to obscure the precise location of users at the scale of the nodes.

Our evaluation demonstrated effective learning across datasets with varying temporal granularity. Models trained on finer intervals, such as minute-level data, achieved lower errors but exhibited greater variability in validation performance. The RMSE and MAE metrics reached maximum values of approximately 1.1, 1.5, and 2 for local models trained on minutely, quarterly, and hourly data, respectively, indicating that finer temporal models tend to work better in this context. These results are comparable to the performance of the GT-LSTM model from fixed sites using one year of data.

The global model assessment indicated that pre-training significantly enhanced stability and convergence, requiring fewer aggregation rounds. In comparison, the trainable model performed effectively over 10 aggregation rounds, demonstrating slightly longer stability than the pre-trained model. We examined the impact of half the nodes being silent during the aggregation process. Node participation is critical for maintaining data diversity and model generalization; thus, reducing the number of participating nodes led to variability and fluctuations in convergence and performance gains in our case. However, our framework still operates effectively with random half participation in each round. This highlights the importance of active node involvement in ensuring consistent and reliable model performance.

Our scalability assessment indicated that configurations with 4 to 6 nodes were optimal for enhancing model efficiency with the available user data in our study, which included 80

users for training and 20 users for testing, along with 4 fixed stations in Dijon. We found that model functionality is closely tied to the number of users, and local models in nodes with more than 30% missing values were less productive. Therefore, controlling the sample size before aggregation is recommended.

The long-term prediction evaluation demonstrated the system's capability to forecast over extended periods, leveraging the temporal attention mechanism in PMFORECAST. However, a significant decline in model performance was observed after 12 hours, which aligns with common challenges in time-series forecasting. Techniques such as hyper-parameter optimization and increasing the training dataset could help models better capture the complexities inherent in time series data, resulting in more reliable long-term predictions.

When comparing the performance of GT-LSTM and FEDAIRNET in a 4-node configuration for long-term predictions, we found that the accuracy metrics for a 24-hour forecast horizon were fairly similar. However, the RMSE for GT-LSTM was in a lower range of error. It is important to note that the GT-LSTM model was trained on a much larger dataset, utilizing 80% of a year (approximately 9.5 months), whereas the FEDAIRNET model was trained on only 80% of one month (about 24 days).

Finally, the Flower framework enabled us to simulate different scenarios and node configurations in real-time on a host computer. This allowed us to monitor and analyze results, offering flexibility for future research and real-time monitoring solutions in urban environments.

This research paves the way for future studies on next-generation air pollution forecasting by analyzing spatiotemporal data and integrating low-cost mobile sensors (Apolline) with affordable on-ground fixed stations (Qameleo). The framework is highly adaptable, facilitating the easy incorporation of new regions and nodes. It operates in real-time and can dynamically adjust to the demographic and geographic characteristics of urban areas. Furthermore, the system is capable of transitioning between different topologies and managing updates, even when some nodes lack data or are compromised by malicious activity. Overall, our framework demonstrated effective performance across both local and global models with various node configurations, utilizing a lightweight machine learning model that is cost-effective and easily deployable on the nodes.

6.4 Future Work

Overall, the need for accurate air quality predictions in both temporal and spatial dimensions highlights the significance of this research area. Mobile sensors provide a cost-effective solution to address data scarcity issues associated with fixed stationary monitoring stations, especially in remote areas where user participation may be low. The integration of advanced techniques

such as federated learning is essential to protect user privacy. Balancing model accuracy, data distribution, and privacy considerations is crucial for effective air pollution forecasting.

The findings from this thesis contribute to a deeper understanding of how to optimize predictive models for environmental monitoring and data privacy. Future research should focus on refining these models by integrating spatial and temporal characteristics, leveraging mobile sensor networks, and exploring hybrid approaches to tackle accuracy and privacy challenges. This approach would be particularly valuable in large-scale cities with dense populations, such as Paris, where precise, real-time air quality forecasting is crucial for public health and environmental policy. The insights gained pave the way for more effective air quality management and enhanced user privacy in data-driven applications.

However, the limitations in data availability have significantly impacted model performance. While mobile sensors provide flexibility, their deployment necessitates considerable effort and the involvement of participants. Additionally, data collection may be limited during nighttime hours when users are typically asleep, resulting in collected indoor measurements or data being cut off after 21:00 due to privacy concerns. To address this issue, we recommend utilizing multi-source data, such as nighttime public transport services or taxis operating at night, to gather data across various nodes.

Conversely, increasing the number of nodes enhanced privacy protection. However, the optimal number of nodes depends not only on privacy considerations but also on the model's performance in our study. Additionally, several factors must be taken into account, including the cost of deploying edge devices per node, communication protocols, the logistics of carrying mobile sensors, calibrating those sensors, and maintaining the entire system. Furthermore, the reliability of mobile sensors may vary due to environmental conditions and calibration issues, which could impact the overall effectiveness of the air quality prediction system.

BIBLIOGRAPHY

[AAT22] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *Expert Syst. Appl.*, 199:116981, 2022.

[ABA+20] Lilian N. Araujo, Jonatas Trabuco Belotti, Thiago Antonini Alves, Yara de Souza Tadano, and Hugo Siqueira. Ensemble method based on artificial neural networks to estimate air pollution health risks. *Environ. Model. Softw.*, 123, 2020.

[ACL20] Satheesh Abimannan, Yue-Shan Chang, and Chi-Yeh Lin. Air pollution forecasting using lstm-multivariate regression model. In Ching-Hsien Hsu, Sondès Kallel, Kun-Chan Lan, and Zibin Zheng, editors, *Internet of Vehicles. Technologies and Services Toward Smart Cities*, pages 318–326, Cham, 2020. Springer International Publishing.

[ADN+21] Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Antonio Ferrara, and Fedelucio Narducci. How to put users in control of their data in federated top-n recommendation with learning to rank. In Chih-Cheng Hung, Jiman Hong, Alessio Bechini, and Eunjee Song, editors, *SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021*, pages 1359–1362. ACM, 2021.

[Age20] Environmental Protection Agency. Six pollutants, 2020. https://www3.epa.gov/airquality/urbanair/.

[Air] Air quality modeling. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/air-quality-modeling. Accessed: 2024-08-12.

[AISW24] Tania Septi Anggraini, Hitoshi Irie, Anjar Dimara Sakti, and Ketut Wikantika. Machine learning-based global air quality index development using remote sensing and ground-based stations. *Environmental Advances*, 15:100456, 2024.

[AKH+24] Christopher Andersen, Matthias Ketzel, Ole Hertel, Jesper H. Christensen, and Jørgen Brandt. The danish lagrangian model (dalm): Development of a new

local-scale high-resolution air pollution model. *Environmental Modelling & Software*, 176:106010, 2024.

[ANF+17] K. W. Appel, S. L. Napelenok, K. M. Foley, H. O. T. Pye, C. Hogrefe, D. J. Luecken, J. O. Bash, S. J. Roselle, J. E. Pleim, H. Foroutan, W. T. Hutzell, G. A. Pouliot, G. Sarwar, K. M. Fahey, B. Gantt, R. C. Gilliam, N. K. Heath, D. Kang, R. Mathur, D. B. Schwede, T. L. Spero, D. C. Wong, and J. O. Young. Description and evaluation of the community multiscale air quality (cmaq) modeling system version 5.1. *Geoscientific Model Development*, 10(4):1703–1732, 2017.

[AQ19] Shaheen Alhirmizy and Banaz Qader. Multivariate time series forecasting with lstm for madrid, spain pollution. In *2019 International Conference on Computing and Information Science and Technology and Their Applications (ICCISTA)*, pages 1–5, 2019.

[Arr96] Svante Arrhenius. On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science*, 41(251):237–276, 1896.

[BAEO20] Josue Becerra-Rico, Marco Antonio Aceves-Fernández, Karen Esquivel-Escalante, and Jesús Carlos Pedraza Ortega. Airborne particle pollution predictive model using gated recurrent unit (GRU) deep neural networks. *Earth Sci. Informatics*, 13(3):821–834, 2020.

[BFA20] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. *CoRR*, abs/2004.11791, 2020.

[BFL+16] S. Brusca, F. Famoso, R. Lanzafame, S. Mauro, A. Marino Cugno Garrano, and P. Monforte. Theoretical and experimental study of gaussian plume model in small scale system. *Energy Procedia*, 101:58–65, 2016. ATI 2016 - 71st Conference of the Italian Thermal Machines Engineering Association.

[BL79] P. W. Biggins and G. E. Likens. A regional assessment of atmospheric deposition. *Ecological Monographs*, 49(1):211–222, 1979.

[BLC18] Tien-Cuong Bui, Van-Duc Le, and Sang-Kyun Cha. A deep learning approach for air pollution forecasting in south korea lstm. *CoRR*, abs/1804.07891, 2018.

[BMH+24] Yacine Belal, Sonia Ben Mokhtar, Hamed Haddadi, Jaron Wang, and Afra Mashhadi. Survey of federated learning models for spatial-temporal mobility applications, 2024.

[Boa] California Air Resources Board. The evolution of air quality monitoring.

[Bri23] T Britannica. *Air pollution*. Encyclopædia Britannica, 2023. https://www.britannica.com/explore/savingearth/air-pollution.

[BRRM20] Sagar V Belavadi, Sreenidhi Rajagopal, Ranjani R, and Rajasekar Mohan. Air quality forecasting using lstm rnn and wireless sensor networks. *Procedia Computer Science*, 170:241–248, 2020.

[BS06] M. Bufalini and L. Sozzi. Historical trends of ozone concentration in milan, italy. *Atmospheric Environment*, 40(20):3855–3866, 2006.

[BTM+22] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. Flower: A friendly federated learning research framework, 2022.

[CCHC21] Chien-Hung Chen, Tu-Fu Chen, Shang-Ping Huang, and Ken-Hui Chang. Comparison of the radm2 and racm chemical mechanisms in o(3) simulations: effect of the photolysis rate constant. *Scientific Reports*, 11(1):5024, 2021. PMID: 33658633, PMCID: PMC7930097.

[CH20] Suzanne Crumeyrolle and Benjamin Hanoune. Le réseau de capteurs Apolline, un outil pour l'étude de la qualité de l'air et pour l'évaluation de l'exposition individuelle. In *Atelier INSU "Nouveaux capteurs environnementaux : DEFI 15"*, Banyuls Sur Mer, France, January 2020. CNRS INSU.

[Che24] Min Chen. *Understanding and Assessment of Privacy Risks in Machine Learning Systems*. PhD thesis, 2024.

[CLAL23] Muhammad Shukri Che Lah, Nureize Arbaiy, and Pei-Chun Lin. Arima-lp: A hybrid model for air pollution forecasting with uncertainty data. In *2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS)*, pages 353–356, 2023.

[CR21] Utkarsh P. Chinchole and Shital A. Raut. Federated learning for estimating air quality. In *12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021, Kharagpur, India, July 6-8, 2021*, pages 1–7. IEEE, 2021.

[CRS18] Sabine Crunaire, Nathalie Redon, and Laurent Spinelle. 1er essai national d'aptitude des micro-capteurs (eaμc) pour la surveillance de la qualité de l'air: Synthèse des résultats. https://hal.archives-ouvertes.fr/hal-04250973, 2018. LCSQA. ffhal04250973f.

[CTK+21] Prateek Chhikara, Rajkumar Tekchandani, Neeraj Kumar, Sudeep Tanwar, and Joel J. P. C. Rodrigues. Federated learning for air quality index prediction using UAV swarm networks. In *IEEE Global Communications Conference, GLOBECOM 2021, Madrid, Spain, December 7-11, 2021*, pages 1–6. IEEE, 2021.

[CWF+20] Chaochao Chen, Bingzhe Wu, Wenjin Fang, Jun Zhou, Li Wang, Yuan Qi, and Xiaolin Zheng. Practical privacy preserving POI recommendation. *CoRR*, abs/2003.02834, 2020.

[CWY+19] Yiqiang Chen, Jindong Wang, Chaohui Yu, Wen Gao, and Xin Qin. Fedhealth: A federated transfer learning framework for wearable healthcare. *CoRR*, abs/1907.09173, 2019.

[CZZ+21] Kuang Cheng, Xiangyu Zhao, Wang Zhou, Yi Cao, Shuang-Hua Yang, and Jianmeng Chen. Source term estimation with deficient sensors: Traceability and an equivalent source approach. *Process Safety and Environmental Protection*, 152:131–139, 2021.

[DAB22] Ghufran Isam Drewil and Riyadh Jabbar Al-Bahadili. Air pollution prediction using lstm deep learning and metaheuristics algorithms. *Measurement: Sensors*, 24:100546, 2022.

[DBV16] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, abs/1606.09375, 2016.

[DC08] E. Demael and B. Carissimo. Comparative evaluation of an eulerian cfd and gaussian plume models based on prairie grass dispersion experiment. *Journal of Applied Meteorology and Climatology*, 47(3):888 – 900, 2008.

[DGL+13] Sabrina De Capitani di Vimercati, Angelo Genovese, Giovanni Livraga, Vincenzo Piuri, and Fabio Scotti. Chapter 57 - privacy and security in environmental monitoring systems: Issues and solutions. In John R. Vacca, editor, *Computer and Information Security Handbook (Third Edition)*, pages 823–841. Morgan Kaufmann, Boston, third edition edition, 2013.

[DKD+22] Le Duy Dong, Tran Anh Khoa, Minh-Son Dao, Kieu-Chinh Nguyen-Ly, Hoang-Son Le, Xuan-Dao Nguyen-Thi, Thanh-Qui Pham, Van-Luong Nguyen, and Bach-Yen Nguyen-Thi. Insights into multi-model federated learning: An advanced approach for air quality index forecasting. *Algorithms*, 15(11):434, 2022.

[DP22] Sweta Dey and Sujata Pal. Federated learning-based air quality prediction for smart cities using BGRU model. In *ACM MobiCom '22: The 28th Annual*

International Conference on Mobile Computing and Networking, Sydney, NSW, Australia, October 17 - 21, 2022, pages 871–873. ACM, 2022.

[DSSY24] Narayan Babu Dhital, Ramesh Prasad Sapkota, Aleeha Sharjeel, and Hsi-Hsien Yang. Estimating potentially preventable ambient pm2.5-attributable adult deaths by improving air quality in nepal. *Atmospheric Pollution Research*, 15(8):102175, 2024.

[Dun14] B. N. et al. Duncan. Satellite observation of integrated ozone concentrations for air quality monitoring. *Atmospheric Measurement Techniques*, 7(7):3487–3511, 2014.

[EG23] Yuan Erbiao and Yang Guangfei. Sa–emd–lstm: A novel hybrid method for long-term prediction of classroom pm2.5 concentration. *Expert Systems with Applications*, 230:120670, 2023.

[FBB+20] David Fowler, Peter Brimblecombe, John Burrows, Mathew R. Heal, Peringe Grennfelt, David S. Stevenson, Alan Jowett, Eiko Nemitz, Mhairi Coyle, Xuejun Liu, Yunhua Chang, Gary W. Fuller, Mark A. Sutton, Zbigniew Klimont, Mike H. Unsworth, and Massimo Vieno. A chronology of global air quality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2183):20190314, 2020.

[FRS+20] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1):10:1–10:21, 2020.

[FZL23] Wei Fang, Runsu Zhu, and Jerry Chun-Wei Lin. An air quality prediction model based on improved vanilla lstm with multichannel input and multiroute output. *Expert Systems with Applications*, 211:118422, 2023.

[GDCM23] Lucia García-Duarte, Jenny Cifuentes, and Geovanny Marulanda. Short-term spatio-temporal forecasting of air temperatures using deep graph convolutional neural networks. *Stochastic Environmental Research and Risk Assessment*, 37(5):1649–1667, 2023.

[GHL+20] Qingchun Guo, Zhenfang He, Shanshan Li, Xinzhou Li, Jingjing Meng, Zhanfang Hou, Jiazhen Liu, and Yongjin Chen. Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol and Air Quality Research*, 20(6):1429–1439, 2020.

[GHW23] Qingchun Guo, Zhenfang He, and Zhaosheng Wang. Prediction of hourly pm2.5 and pm10 concentrations in chongqing city in china based on artificial neural network. *Aerosol and Air Quality Research*, 23(6):220448, 2023.

[GHW24] Qingchun Guo, Zhenfang He, and Zhaosheng Wang. The characteristics of air quality changes in hohhot city in china and their relationship with meteorological and socio-economic factors. *Aerosol and Air Quality Research*, 24(5):230274, 2024.

[GLC+21] Yeting Guo, Fang Liu, Zhiping Cai, Hui Zeng, Li Chen, Tongqing Zhou, and Nong Xiao. PREFER: point-of-interest recommendation with efficiency and privacy-preservation via federated edge learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1):13:1–13:25, 2021.

[GOO22] Aysenur Gilik, Arif Selcuk Ogrenci, and Atilla Ozmen. Air quality prediction using cnn+lstm-based hybrid deep learning architecture. *Environmental Science and Pollution Research*, 29(8):11920–11938, 2022.

[GSMP23] A Gangwar, S Singh, R Mishra, and S Prakash. The state-of-the-art in air pollution monitoring and forecasting systems using iot, big data and machine learning. *ArXiv*, 2023.

[GWZ+21] Liang Ge, Kunyan Wu, Yi Zeng, Feng Chang, Yaqian Wang, and Siyu Li. Multi-scale spatiotemporal graph convolution network for air quality prediction. *Appl. Intell.*, 51(6):3491–3505, 2021.

[Hay21] Jim Haywood. Chapter 30 - atmospheric aerosols and their role in climate change. In Trevor M. Letcher, editor, *Climate Change (Third Edition)*, pages 645–659. Elsevier, third edition edition, 2021.

[HBL15] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *CoRR*, abs/1506.05163, 2015.

[HDQ+22] Jelle Hofman, Tien Huu Do, Xuening Qin, Esther Rodrigo Bonet, Wilfried Philips, Nikos Deligiannis, and Valerio Panzica La Manna. Spatiotemporal air quality inference of low-cost sensor data: Evidence from multiple sensor testbeds. *Environmental Modelling & Software*, 149:105306, 2022.

[HK18] Chiou-Jye Huang and Ping-Huan Kuo. A deep CNN-LSTM model for particulate matter ($pm_{2.5}$) forecasting in smart cities. *Sensors*, 18(7):2220, 2018.

[HKV+19] Benjamin Hanoune, Redha Kassi, Bernard Verbeke, Eliane Assy, Laurent Clavier, Suzanne Crumeyrolle, Samuel Degrande, Xavier Le Pallec, and Romain Rouvoy. Conception and deployment of the APOLLINE sensor network for IAQ monitoring. *IOP Conference Series: Materials Science and Engineering*, 609, October 2019.

[HLC+23] Jiaan He, Xiaoyong Li, Zhenguo Chen, Wenjie Mai, Chao Zhang, Xin Wan, Xin Wang, and Mingzhi Huang. A hybrid clstm-gpr model for forecasting particulate matter (pm2.5). *Atmospheric Pollution Research*, 14(8):101832, 2023.

[HLL+21] Wei Huang, Tianrui Li, Jia Liu, Peng Xie, Shengdong Du, and Fei Teng. An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. *Information Fusion*, 75:28–40, 2021.

[HNA24] Muhammad Ayat Hidayat, Yugo Nakamura, and Yutaka Arakawa. Enhancing efficiency in privacy-preserving federated learning for healthcare: Adaptive gaussian clipping with dft aggregator. *IEEE Access*, 12:88445–88457, 2024.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[JBG19] Akshat Jain, Ashim Bhasin, and Varun Gupta. Prediction of air pollution using lstm-based recurrent neural networks. *Int. J. Comput. Intell. Stud.*, 8(4):299–308, 2019.

[KBD+15] Federico Karagulian, Claudio A. Belis, Carlos Francisco C. Dora, Annette M. Prüss-Ustün, Sophie Bonjour, Heather Adair-Rohani, and Markus Amann. Contributions to cities' ambient particulate matter (pm): A systematic review of local source contributions at global level. *Atmospheric Environment*, 120:475–483, 2015.

[Kel17] Allan Kellehear. Public health approaches to dying, death, and loss. In Stella R. Quah, editor, *International Encyclopedia of Public Health (Second Edition)*, pages 184–189. Academic Press, Oxford, second edition edition, 2017.

[KHH+20] Yeongwoo Kim, Ezeddin Al Hakim, Johan Haraldson, Henrik Eriksson, José Mairton B. da Silva Jr., and Carlo Fischione. Dynamic clustering in federated learning. *CoRR*, abs/2012.03788, 2020.

[KKB+15] Chandrasekharan Nair Kesavachandran, Ritul Kamal, Vipin Bihari, Manoj Kumar Pathak, and Amarnath Singh. Particulate matter in ambient air and its association with alterations in lung functions and respiratory health problems among outdoor exercisers in national capital region, india. *Atmospheric Pollution Research*, 6(4):618–625, 2015.

[KR24] S. Kumar and P. Rani. An ai-based liver disease prediction model based on pearson correlation feature selection method. *Biomedical and Pharmacology Journal*, 17(4), 2024.

[KRA10] Kavi Khedo, Perseedoss Rajiv, and Mungur Avinash. A wireless sensor network air pollution monitoring system. *International Journal of Wireless & Mobile Networks*, 2, 05 2010.

[KWSR19] M. Karl, S.-E. Walker, S. Solberg, and M. O. P. Ramacher. The eulerian urban dispersion model episode – part 2: Extensions to the source dispersion and photochemistry for episode–citychem v1.2 and its application to the city of hamburg. *Geoscientific Model Development*, 12(8):3357–3399, 2019.

[LAK12] D. Langmuir, D. Alexis, and S. K. Kounga. Air pollution and respiratory illness in young children. *International Journal of Occupational Medicine, Environmental Health*, 17(2):146–162, 2012.

[LAL+22] Jiangtao Li, Xingqin An, Qingyong Li, Chao Wang, Haomin Yu, Xinyuan Zhou, and Yangli ao Geng. Application of xgboost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, 276:106238, 2022.

[LBC20] Van-Duc Le, Tien-Cuong Bui, and Sang-Kyun Cha. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 55–62, 2020.

[LC12] Xi Luo and Han Cao. Evaluation of air quality using the cmaq modeling system. *Procedia Environmental Sciences*, 12:159–165, 2012.

[LCY18] Yang Liu, Tianjian Chen, and Qiang Yang. Secure federated transfer learning. *CoRR*, abs/1812.03337, 2018.

[Le23] Van-Duc Le. Spatiotemporal graph convolutional recurrent neural network model for citywide air pollution forecasting, 2023.

[LEF+15] J. Lelieveld, J. S. Evans, M. Fnais, G. Giannopoulos, and A. Pozzer. The contribution of outdoor air pollution to premature mortality in europe. *Atmospheric Environment*, 103:100–112, 2015.

[LLM+23] Yanhui Liu, Jiayin Li, Yufang Ma, Ming Zhou, Zhaofeng Tan, Limin Zeng, Keding Lu, and Yuanhang Zhang. A review of gas-phase chemical mechanisms commonly used in atmospheric chemistry modelling. *Journal of Environmental Sciences*, 123:522–534, 2023.

[LNL+20] Yi Liu, Jiangtian Nie, Xuandi Li, Syed Hassan Ahmed, Wei Yang Bryan Lim, and Chunyan Miao. Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms. *CoRR*, abs/2007.12004, 2020.

[LNL+21] Yi Liu, Jiangtian Nie, Xuandi Li, Syed Hassan Ahmed, Wei Yang Bryan Lim, and Chunyan Miao. Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms. *IEEE Internet Things J.*, 8(12):9827–9837, 2021.

[LPG+16] Natalia Liora, Anastasia Poupkou, Theodore M. Giannaros, Konstantinos E. Kakosimos, Olaf Stein, and Dimitrios Melas. Impacts of natural emission sources on particle pollution levels in europe. *Atmospheric Environment*, 137:171–185, 2016.

[LSTS19] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *CoRR*, abs/1908.07873, 2019.

[LYK+20] Yi Liu, James J. Q. Yu, Jiawen Kang, Dusit Niyato, and Shuyu Zhang. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet Things J.*, 7(8):7751–7763, 2020.

[LZ16] Phong Le and Willem H. Zuidema. Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive lstms. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 87–93. Association for Computational Linguistics, 2016.

[LZB+24] Keshuo Liu, Huanhuan Zhang, Yacong Bo, Yao Chen, Panpan Zhang, Cunrui Huang, Zengli Yu, and Zhan Gao. Ambient air pollution and children's health: An umbrella review. *Atmospheric Pollution Research*, 15(6):102108, 2024.

[LZL+19] Yee Leung, Yu Zhou, Ka-Yu Lam, Tung Fung, Kwan-Yau Cheung, Taehong Kim, and Hanmin Jung. Integration of air pollution data collected by mobile sensors and ground-based stations to derive a spatiotemporal air pollution profile of a city. *International Journal of Geographical Information Science*, 33(11):2218–2240, 2019.

[Mac04] Michael C. MacCracken. The discovery of global warming. *Eos, Transactions American Geophysical Union*, 85(28):270–270, 2004.

[MAD89] S.E. Metcalfe, D.H.F. Atkins, and R.G. Derwent. Acid deposition modelling and the interpretation of the united kingdom secondary precipitation network data. *Atmospheric Environment (1967)*, 23(9):2033–2052, 1989.

[Mar] Julia Martinez. Great smog of london. https://www.britannica.com/event/Great-Smog-of-London.

[MCB+93] Paulette Middleton, Julius S. Chang, Mark Beauharnois, Linda Hash, and Francis S. Binkowski. The role of nitrogen oxides in oxidant production as predicted by the regional acid deposition model (radm). *Water, Air, and Soil Pollution*, 67(1):133–159, 1993.

[MM04] M. J. Molina and L. T. Molina. Atmospheric chemistry: environmental influence on the formation of tropospheric ozone. *Annual Review of Environment and Resources*, 29(1):415–449, 2004.

[MMPJY22] Axel Gedeon Mengara Mengara, Eunyoung Park, Jinho Jang, and Younghwan Yoo. Attention-based distributed deep learning model for air quality forecasting. *Sustainability*, 14(6), 2022.

[MMRyA16] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

[MNS+23] Nadège Martiny, Marilleau Nicolas, Marion Sarah, Diallo-Dudek Julita, Lola Canovas, Alexandre Bisquerra, Mario Rega, and Thomas Thévenin. Quality of air module for environmental learning engineering and observation network (qameleondijon) : un réseau dense de mesures de qualité de l'air à dijon», climatologie. 20(4), 2023.

[MPP+21] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[MSM21] Afra Mashhadi, Joshua Sterner, and Jeffrey Murray. Deep embedded clustering of urban communities using federated learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[MSSB20] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: A review." frontiers in public health. *Front Public Health*, 8(14), 2020.

[MT23] Subhojit Mandal and Mainak Thakur. A city-based pm2.5 forecasting framework using spatially attentive cluster-based graph neural network model. *Journal of Cleaner Production*, 405:137036, 2023.

[MV23] Dhanalakshmi M and Radha V. Novel regression and least square support vector machine learning technique for air pollution forecasting. *CoRR*, abs/2306.07301, 2023.

[MWC17] Peter Mueller, John Watson, and Judith Chow. A historical perspective on air quality measurements from the career of dr. peter k. mueller, 2017 critical review discussion addendum. *Journal of the Air & Waste Management Association*, 67:S1–S10, 10 2017.

[NDS+21] Bassirou Ngom, Moussa Diallo, Madoune Robert Seyc, Mamadou Simina Drame, Christophe Cambier, and Nicolas Marilleau. Pm10 data assimilation on real-time agent-based simulation using machine learning models: case of dakar urban air pollution study. In *2021 IEEE/ACM 25th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, pages 1–4, 2021.

[NFZM19] Lazrak Noussair, Jesualdo Tomás Fernández-Breis, Jihad Zahir, and Hajar Mousannif. Towards distributed learning in internet of things. air quality monitoring use case. In J. Christian Attiogbé, Flavio Ferrarotti, and Sofian Maabout, editors, *New Trends in Model and Data Engineering - MEDI 2019 International Workshops, DETECT, DSSGA, TRIDENT, Toulouse, France, October 28-31, 2019, Proceedings*, volume 1085 of *Communications in Computer and Information Science*, pages 154–159. Springer, 2019.

[NKT+12] Uarporn Nopmongcol, Bonyoung Koo, Edward Tai, Jaegun Jung, Piti Piyachaturawat, Chris Emery, Greg Yarwood, Guido Pirovano, Christina Mitsakou, and George Kallos. Modeling europe with camx for the air quality model evaluation international initiative (aqmeii). *Atmospheric Environment*, 53:177–185, 2012. AQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models - Phase 1.

[Noh21] Seol-Hyun Noh. Analysis of gradient vanishing of rnns and performance comparison. *Inf.*, 12(11):442, 2021.

[NZ21] Do-Van Nguyen and Koji Zettsu. Spatially-distributed federated learning of convolutional recurrent neural networks for air pollution prediction. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3601–3608, 2021.

[OA] National Oceanic and Atmospheric Administration. A brief history of pollution. https://oceanservice.noaa.gov/education/tutorial_pollution/02history.html.

[Org21] World Health Organization. Ambient (outdoor) air pollution and health, 2021. https://www.who.int/news-room/fact-sheets/detail/ambient-%28outdoor%29-air-quality-and-health.

[Org23] World Health Organization. Ambient (outdoor) air pollution, 2023.

[PCK+23] Chung Park, Taekyoon Choi, Taesan Kim, Mincheol Cho, Junui Hong, Minsung Choi, and Jaegul Choo. Fedgeo: Privacy-preserving user next location prediction

with federated learning. SIGSPATIAL '23, New York, NY, USA, 2023. Association for Computing Machinery.

[PCP+21] Karisma Trinanda Putra, Hsing-Chung Chen, Prayitno, Marek R. Ogiela, Chao-Lung Chou, Chien-Erh Weng, and Zon-Yin Shae. Federated compressed learning edge computing framework with ensuring data privacy for pm2.5 prediction in smart city sensing applications. *Sensors*, 21(13), 2021.

[PDSE23] Vasileios Perifanis, George Drosatos, Giorgos Stamatelatos, and Pavlos S. Efraimidis. Fedpoirec: Privacy-preserving federated poi recommendation with social influence. *Inf. Sci.*, 623:767–790, 2023.

[Pet14] K. A. Peterson. *Air Quality Monitoring and Forecasting*. Elsevier, 2014.

[PFH24] Daniel A. Potts, Emma J.S. Ferranti, and Joshua D. Vande Hey. Investigating the barriers and pathways to implementing satellite data into air quality monitoring, regulation and policy design in the united kingdom. *Environmental Science & Policy*, 151:103621, 2024.

[PMB+18] Vincent Primault, Mohamed Maouche, Antoine Boutet, Sonia Ben Mokhtar, Sara Bouchenak, and Lionel Brunie. Accio: How to make location privacy experimentation open and easy. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 896–906. IEEE, 2018.

[PMR+20] Unjin Pak, Jun Ma, Unsok Ryu, Kwangchol Ryom, U. Juhyok, Kyongsok Pak, and Chanil Pak. Deep learning-based pm2.5 prediction considering the spatiotemporal correlations: A case study of beijing, china. *Science of The Total Environment*, 699:133561, 2020.

[PRGDS+24] Ignacio-Iker Prado-Rujas, Antonio García-Dopico, Emilio Serrano, M. Luisa Córdoba, and María S. Pérez. A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on deep learning. *Engineering Applications of Artificial Intelligence*, 127:107271, 2024.

[PT23] Vinoth Panneerselvam and Revathi Thiagarajan. Acbigru-dao: Attention convolutional bidirectional gated recurrent unit-based dynamic arithmetic optimization for air quality prediction. *Environmental Science and Pollution Research*, 30(37):86804–86820, 2023.

[QLY+23] Junlong Qian, Duanyang Liu, Shuqi Yan, Muning Cheng, Rongwei Liao, Shengjie Niu, Wenlian Yan, Shuyao Zha, Lulu Wang, and Xiaoxiao Chen. Fog scavenging of particulate matters in air pollution events: Observation and simulation in the yangtze river delta, china. *Science of The Total Environment*, 876:162728, 2023.

[QYZ+19] Dongming Qin, Jian Yu, Guojian Zou, Ruihan Yong, Qin Zhao, and Bo Zhang. A novel combined prediction scheme based on cnn and lstm for urban pm2.5 concentration. *IEEE Access*, 7:20050–20059, 2019.

[RCAM+24] Maryam Rahmani, Suzanne Crumeyrolle, Nadége Allegri-Martiny, Amir Taherkordi, and Romain Rouvoy. Pmforecast: leveraging temporal lstm to deliver in situ air quality predictions. *Environmental Science and Pollution Research*, 2024.

[RCH+18] N. Redon, S. Crunaire, B. Herbin, L. Spinelle, and C. Marchand. French joint intercomparison exercises for air quality sensors (eaμc): Results and assessment. In *International Symposium on Individual air pollution sensors: Innovation or Revolution*, 2018.

[RGL+20] Aleksey A. Romanov, Boris A. Gusev, Egor V. Leonenko, Anastasia N. Tamarovskaya, Alexander S. Vasiliev, Nikolai E. Zaytcev, and Ilia K. Philippov. Graz lagrangian model (gral) for pollutants tracking and estimating sources partial contributions to atmospheric pollution in highly urbanized areas. *Atmosphere*, 11(12), 2020.

[RRBPZ19] A. Susana Ramirez, Steven Ramondt, Karina Van Bogart, and Raquel Perez-Zuniga. Public awareness of air pollution and health threats: Challenges and opportunities for communication strategies to improve environmental health literacy. *Journal of Health Communication*, 24(1):75–83, 2019. PMID: 30730281.

[RSB+11] Regina Rückerl, Alexandra Schneider, Susanne Breitner, Josef Cyrys, and Annette Peters. Health effects of particulate air pollution: A review of epidemiological evidence. *Inhalation Toxicology*, 23(10):555–592, 2011. PMID: 21864219.

[Rud03] William F. Ruddiman. The anthropogenic greenhouse era began thousands of years ago. *Climatic Change*, 61(3):261–293, 2003.

[SBW+22] Yuxi Sun, Peter Brimblecombe, Peng Wei, Yusen Duan, Jun Pan, Qizhen Liu, Qingyan Fu, Zhiguang Peng, Shuhong Xu, Ying Wang, and Zhi Ning. High resolution on-road air pollution using a large taxi-based mobile sensor network. *Sensors*, 22(16):6005, 2022.

[SCH+15] M. Schaap, C. Cuvelier, C. Hendriks, B. Bessagnet, J.M. Baldasano, A. Colette, P. Thunis, D. Karam, H. Fagerli, A. Graff, R. Kranenburg, A. Nyiri, M.T. Pay, L. Rouïl, M. Schulz, D. Simpson, R. Stern, E. Terrenoire, and P. Wind. Performance of european chemistry transport models as function of horizontal resolution. *Atmospheric Environment*, 112:90–105, 2015.

[SG22] G. Pius Agbulu S. Gunasekar, G. Joselin Retna Kumar. Air quality predictions in urban areas using hybrid arima and metaheuristic lstm. *Computer Systems Science and Engineering*, 43(3):1271–1284, 2022.

[SKD23] Harshit Srivastava and Santos Kumar Das. Air pollution prediction system using xrsth-lstm algorithm. *Environmental Science and Pollution Research*, 30(60):125313–125327, 2023.

[SKR24] Mohamed Shakir, U. Kumaran, and N. Rakesh. An approach towards forecasting time series air pollution data using lstm-based auto-encoders. *J. Internet Serv. Inf. Secur.*, 14(2):32–46, 2024.

[SP16] John H. Seinfeld and Spyros N. Pandis. *Atmospheric chemistry and physics: From air pollution to climate change*. John Wiley & Sons, 2016.

[SS17] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications*. Springer Texts in Statistics. Springer Cham, Springer International Publishing AG, 4 edition, 2017.

[SSS+24] Guanqun Sun, Han Shu, Feihe Shao, Teeradaj Racharak, Weikun Kong, Yizhi Pan, Jingjing Dong, Shuang Wang, Le-Minh Nguyen, and Junyi Xin. Fkd-med: Privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation. *IEEE Access*, 12:33687–33704, 2024.

[Sto97] et al. Stockwell, W. R. Regional atmospheric chemistry mechanism for eastern north america (racm-ea). *Journal of Geophysical Research: Atmospheres*, 102:25847–25864, 1997.

[TSC+10] Claudio Terzano, F Stefano, V Conti, E Graziani, and Angelo Petroianni. Air pollution ultrafine particles: Toxicity beyond the lung. *European review for medical and pharmacological sciences*, 14:809–21, 2010.

[Tur70] D. Bruce Turner. *Workbook of atmospheric dispersion estimates*. US Department of Health, Education, and Welfare, Public Health Service, Environmental Health Service, 1970. https://www.keysolutionsinc.com/references/Turner%20Workbook.pdf.

[TVC23] Theodore H. Tulchinsky, Elena A. Varavikova, and Matan J. Cohen. A history of public health. In Theodore H. Tulchinsky, Elena A. Varavikova, and Matan J. Cohen, editors, *The New Public Health (Fourth Edition)*, pages 1–54. Academic Press, San Diego, fourth edition edition, 2023.

[TW22a] Zihan Tu and Zhe Wu. Predicting beijing air quality using bayesian optimized cnn-rnn hybrid model. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pages 581–587, 2022.

[TW22b] Zihan Tu and Zhe Wu. Predicting beijing air quality using bayesian optimized CNN-RNN hybrid model. In *Asia Conference on Algorithms, Computing and Machine Learning, CACML 2011, Hangzhou, China, March 25-27, 2022*, pages 581–587. IEEE, 2022.

[TZC18] Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. Air pollution forecasting using rnn with lstm. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 1074–1079, 2018.

[USS+24] Manjari Upreti, Purabi Saikia, Shilky, Preet Lal, and Amit Kumar. Chapter 2 - major challenges in the urbanizing world and role of earth observations for livable cities. In Amit Kumar, Prashant K. Srivastava, Purabi Saikia, and Rajesh Kumar Mall, editors, *Earth Observation in Urban Monitoring*, Earth Observation, pages 23–52. Elsevier, 2024.

[USSJ23] Mitra Unik, Imas Sukaesih Sitanggang, Lailan Syaufina, and I Nengah Surati Jaya. Pm2.5 estimation using machine learning models and satellite data: A literature review. *International Journal of Advanced Computer Science and Applications*, 14(5), 2023.

[Val14] Damiano Vallero. Source apportionment of airborne particulate matter by principal component analysis and absolute principal component scores. *Atmospheric Environment*, 84:881–890, 2014.

[VAMD18] Ishan Verma, Rahul Ahuja, Hardik Meisheri, and Lipika Dey. Air pollutant severity prediction using bi-directional lstm network. In *Air Pollutant Severity Prediction Using Bi-Directional LSTM Network*, pages 651–654, 12 2018.

[VBT16] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized collaborative learning of personalized models over networks. *CoRR*, abs/1610.05202, 2016.

[vK83] N.D. van Egmond and H. Kesseboom. Mesoscale air pollution dispersion models—i. eulerian grid model. *Atmospheric Environment (1967)*, 17(2):257–265, 1983.

[VMC21] Andrey Vlasenko, Volker Matthias, and Ulrich Callies. Simulation of chemical transport model estimates by means of a neural network using meteorological data. *Atmospheric Environment*, 254:118236, 2021.

[Wil16] Granville Tunnicliffe Wilson. , by , , and , . published by , , pp. 712. isbn:. *Journal of Time Series Analysis*, 37(5):709–711, 2016.

[Wil19] Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*. Elsevier, fourth edition edition, 2019.

[WQH24] Yang Wu, Chonghui Qian, and Hengjun Huang. Enhanced air quality prediction using a coupled dvmd informer-cnn-lstm model optimized with dung beetle algorithm. *Entropy*, 26(7), 2024.

[XZL21] Xinghan Xu, Chengkun Zhang, and Yi Liang. Review of satellite-driven statistical models pm2.5 concentration estimation with comprehensive information. *Atmospheric Environment*, 256:118302, 2021.

[YDL+24a] Z. You, X. Dong, X. Liu, S. Gao, Y. Wang, and Y. Shen. Location privacy preservation crowdsensing with federated reinforcement learning. *IEEE Transactions on Dependable and Secure Computing*, (01):1–18, may 2024.

[YDL+24b] Z. You, X. Dong, X. Liu, S. Gao, Y. Wang, and Y. Shen. Location privacy preservation crowdsensing with federated reinforcement learning. *IEEE Transactions on Dependable and Secure Computing*, (01):1–18, may 2024.

[YHZC21] Hongwei Yang, Hui He, Weizhe Zhang, and Xiaochun Cao. Fedsteg: A federated transfer learning framework for secure image steganalysis. *IEEE Transactions on Network Science and Engineering*, 8(2):1084–1094, 2021.

[YWC+24] Yujie Yang, Zhige Wang, Chunxiang Cao, Min Xu, Xinwei Yang, Kaimin Wang, Heyi Guo, Xiaotong Gao, Jingbo Li, and Zhou Shi. Estimation of pm2.5 concentration across china based on multi-source remote sensing data and machine learning methods. *Remote Sensing*, 16(3), 2024.

[Zan90] Paolo Zannetti. *Gaussian Models*, pages 141–183. Springer US, Boston, MA, 1990.

[ZH18] Talat Zaree and Ali Honarvar. Improvement of air pollution prediction in a smart city and its correlation with weather conditions using metrological big data. *Turkish Journal of Electrical Engineering and Computer Sciences*, 26(3):1302 – 1313, 01 2018.

[ZH19] Huilin Zhu and Jinglu Hu. Air quality forecasting using svr with quasi-linear kernel. In *2019 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5, 2019.

[Zha16] L. et al. Zhang. A global aerosol forecast system using wrf/chem: Chemical integration and forecast evaluation. *Geoscientific Model Development*, 9(4):2097–2129, 2016.

[ZHLL22] Qi Zhang, Yang Han, Victor O. K. Li, and Jacqueline C. K. Lam. Deep-air: A hybrid cnn-lstm framework for fine-grained air pollution estimation and forecast in metropolitan cities. *IEEE Access*, 10:55818–55841, 2022.

[ZL22] Jiaxuan Zhang and Shunyong Li. Air quality index forecast in beijing based on cnn-lstm multi-model. *Chemosphere*, 308:136180, 2022.

[ZLSS23] Jiakun Zhao, Hailun Luo, Weiguang Sang, and Kun Sun. Spatiotemporal semantic network for ENSO forecasting over long time horizon. *Appl. Intell.*, 53(6):6464–6480, 2023.

[ZRY+22] Bo Zhang, Yi Rong, Ruihan Yong, Dongming Qin, Maozhen Li, Guojian Zou, and Jianguo Pan. Deep learning for air pollutant concentration prediction: A review. *Atmospheric Environment*, 290:119347, 2022.

[ZWL20] B Zhang, B Wu, and J Liu. Pm2.5 pollution-related health effects and willingness to pay for improved air quality: Evidence from china's prefecture-level cities. *Journal of Clean*, 273:122876, 2020.

[ZXL+23] C. Zhan, M. Xie, H. Lu, B. Liu, Z. Wu, T. Wang, B. Zhuang, M. Li, and S. Li. Impacts of urbanization on air quality and the related health risks in a city with complex terrain. *Atmospheric Chemistry and Physics*, 23(1):771–788, 2023.

[ZZDL21] Hongye Zhou, Feng Zhang, Zhenhong Du, and Renyi Liu. Forecasting pm2.5 using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability. *Environmental Pollution*, 273:116473, 2021.

[ZZX+23] Bin Zhou, Sanbao Zhang, Ruibin Xue, Jiayi Li, and Shanshan Wang. A review of space-air-ground integrated remote sensing techniques for atmospheric monitoring. *Journal of Environmental Sciences*, 123:3–14, 2023.

[ZZZ+22] Zhen Zhang, Shiqing Zhang, Xiaoming Zhao, Linjian Chen, and Jun Yao. Temporal difference-based graph transformer networks for air quality pm2.5 prediction: A case study in china. *Frontiers in Environmental Science*, 10, 2022.

Appendix A

APPENDICES

A.1 Appendix

This appendix provides the algorithms used to simulate user data based on individual trajectories.

A.2 Appendix

This section contains the training and validation loss, MAE, and errors measured by RMSE across 50 epochs for 12 federated learning nodes, using datasets at hourly, quarterly, and minute-level granularity.

A.3 Appendix

The final appendix includes scatter plots comparing the true values and forecasted values for the 4-node configurations.

Algorithm 2 User Movement and Data Simulation

```
function ParseGPX(file_path)
    Open and parse GPX file, extract coordinates
    return List of coordinates
end function
function GenerateTimestamps(coordinates, start_time, interval, end_time)
    Initialize empty list timestamps
    for each coordinate in coordinates do
        if start_time ≥ end_time then
            break
        end if
        Append (start_time, coordinate) to timestamps
        start_time += interval
    end for
    return timestamps, start_time
end function
function CreateDayData(morning_gpx, evening_gpx, stay_hours, date, last_coord, start_hour)
    interval ← 15 seconds
    start_hour ← start_hour or 7:00
    start_time ← date + start_hour
    end_time ← date + 21:00
    Initialize empty list timestamps
    if last_coord then
        while start_time < start_hour do
            Append (start_time, last_coord) to timestamps
            start_time += interval
        end while
    end if
    for each route, time in [(morning_gpx, start_time), (evening_gpx, end_time)] do
        route_coords ← ParseGPX(route)
        timestamps, end_time ← GenerateTimestamps(route_coords, time, interval, end_time)
        Append timestamps to timestamps
    end for
    return timestamps as DataFrame, evening_gpx[-1]
end function
```

Algorithm 3 Creating Comprehensive User Datasets

```
function ProcessUserData(user_file, reference_points, ground_truth_map, output_file)
    Read user data into DataFrame
    Initialize list for combined data
    for each row in user data do
        Extract latitude, longitude, and timestamp
        Use Haversine functions to find the closest reference point
        Read and find closest values from ground truth
        Append combined data
    end for
    Create DataFrame from combined data
    Create CSV file from the DataFrame
end function
Define reference points
Load ground truth file
Process user data and generate output file
Check for missing values in output file
```

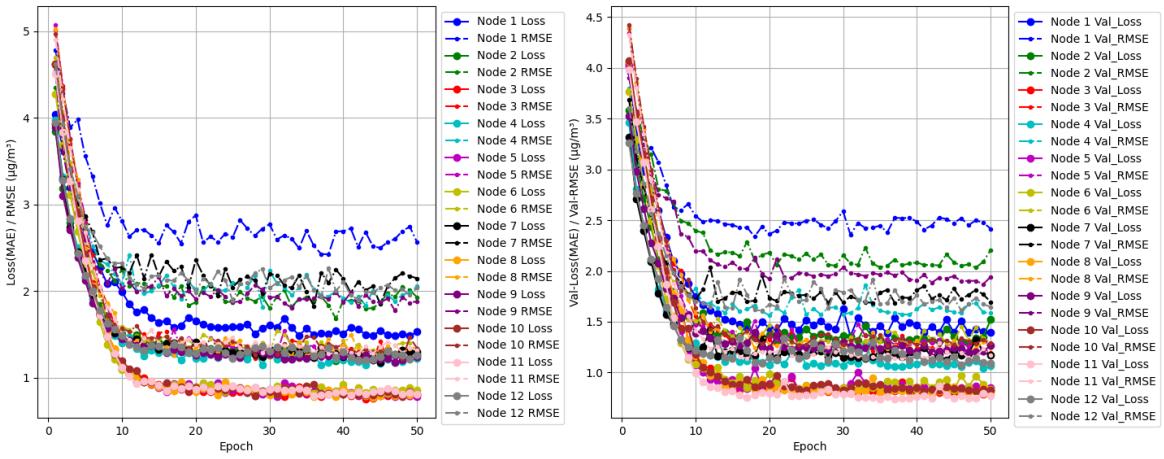


Figure A.1: Training and validation loss (measured by MAE) and error (represented by RMSE) across 50 epochs for 12 federated learning nodes using hourly datasets. The left panel illustrates the convergence of training loss over time, while the right panel highlights the variability in validation performance among the four node configurations.

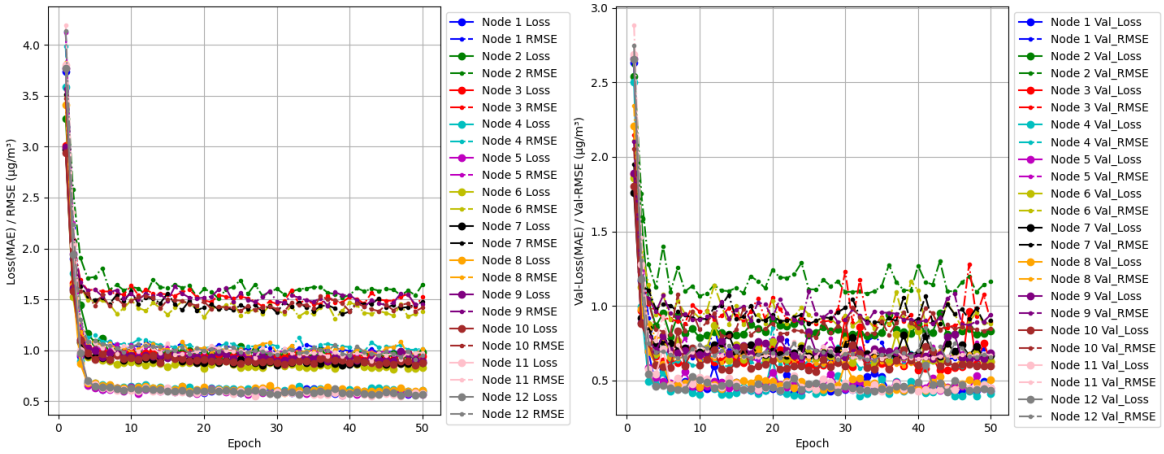


Figure A.2: Comparison of loss (measured by MAE) and error (represented by RMSE) for local models across 12 nodes, trained with quarterly datasets. The left panel depicts the consistent reduction in loss over time during training, while the right panel emphasizes the variability in validation performance among the four nodes.

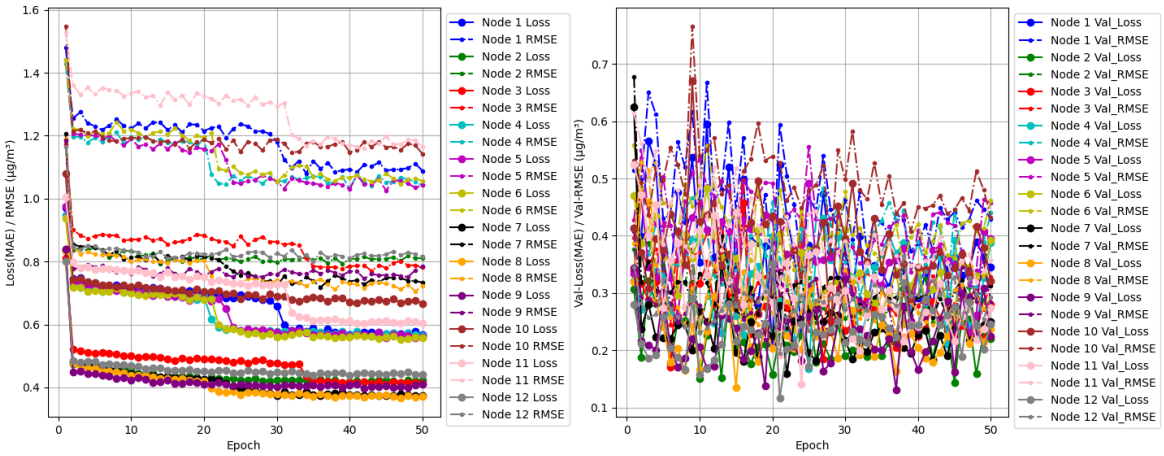


Figure A.3: Comparison of loss (measured by MAE) and error (represented by RMSE) for local models across 12 nodes, trained with minute-level datasets. The left panel illustrates the consistent reduction in loss over time during training, while the right panel highlights the variability in validation performance across the nodes.

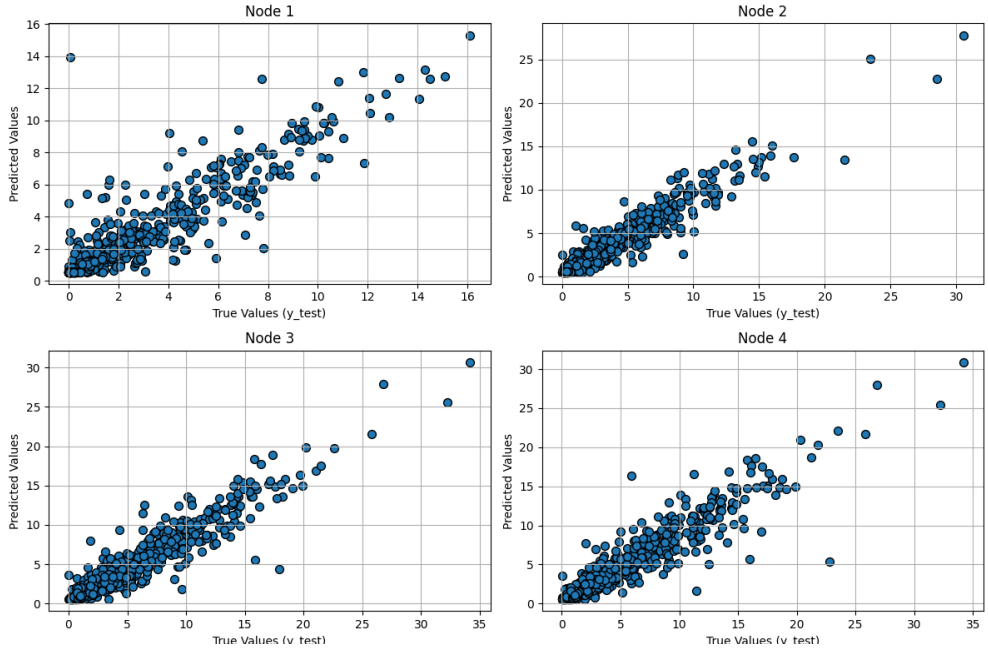


Figure A.4: Scatter plot illustrating the correlation between true values and forecasted values for the 4-node configurations.