Spatiotemporal Analysis of Air Pollution Using Advanced Machine Learning Techniques

by

SAFIN RAHMAN
ID: 21201363

NAEEM ISLAM
ID: 24141077

MD. SHOAIBUR RAHMAN
ID: 21101322

MD. MONIRUZZAMAN HRIDOY
ID: 21201316

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

SAFIN RAHMAN
21201363

NAEEM ISLAM
24141077

MD. SHOAIBUR RAHMAN
21101322

MD.MONIRUZZAMAN HRIDOY
21201316

Approval

The thesis/project titled **" Spatiotemporal Analysis of Air Pollution Using Advanced Machine Learning Techniques "** submitted by

1. SAFIN RAHMAN (21201363)
2. NAEEM ISLAM (24141077)
3. MD. SHOAIBUR RAHMAN (21101322)
4. MD.MONIRUZZAMAN HRIDOY (21201316)

of Fall 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 2025.

Examining Committee:

Supervisor:
(Member)

Md. Ahasanul Alam
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Associate Professor & Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

This thesis presents a unified framework for ***spatiotemporal analysis of air pollu- tion using advanced machine learning*** to enable short-horizon, city-level forecasting and operational decision support. A **leakage-safe**, **multi-source** dataset is curated for **20 cities** across **Bangladesh** and **China**, integrating pollutant observations with spatiotemporal covariates (e.g., meteorological and contextual signals) to model ur- ban pollution dynamics under heterogeneous conditions. The forecasting task is formulated as **multi-output time-series regression** over **PM2.5, PM10, NO2, SO2,** and **CO**. To capture short-term fluctuations and longer temporal dependencies while exploiting cross-pollutant structure, a **multi-task CNN–LSTM** architecture is de- veloped with a shared feature backbone and pollutant-specific prediction heads. Performance is benchmarked against classical machine-learning baselines (including **Random Forest** and **XGBoost**) under city-aware evaluation to assess both accuracy and robustness. To address regional data imbalance, a **cross-country transfer learn- ing strategy** is evaluated by leveraging representations learned from data-rich source cities to improve forecasting in data-scarce target cities. Forecast reliability is en- hanced via **Monte Carlo Dropout** to estimate predictive **uncertainty**, while SHAP and Integrated Gradients provide complementary explanations of feature influence and temporal attribution. Finally, an **early-warning episode detection** layer converts forecasts into event-oriented alerts and diagnostics to support practical monitoring workflows. Overall, the proposed pipeline delivers more accurate, uncertainty-aware, and interpretable multi-pollutant forecasts suitable for risk-sensitive air quality man- agement in heterogeneous urban environments.

Keywords: Air Pollution Forecasting, Spatiotemporal Analysis, ConvLSTM, Multi-pollutant Prediction, Multi-task Learning, CNN–LSTM, Transfer Learning, Uncertainty Quantification, Monte Carlo Dropout, Explainable AI, SHAP, Integrated Gradients, Early-Warning Detection.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Background

Air pollution is one of the major environmental challenges that directly affects the health of the public and the livability of cities. Long-term exposure to high levels of air pollution is associated with the incidence of respiratory and cardiovascular diseases, premature deaths, and high socioeconomic costs. This is particularly true for cities that are rapidly developing and expanding, where the population density, increasing traffic, and rising levels of industrial activity are increasing the levels of emissions, and strong environmental controls are not yet in place [1], [2].

Urban air pollution follows certain trends of variability. First, the levels of air pollutants vary considerably from one urban area to another owing to variations in sources of emission and urban geography as well as local climatic conditions. Additionally, day-to-day human activities and seasonal as well as occasional variations in climatic conditions play significant roles in the variability of air pollutant levels over time. All major air pollutants such as $PM_{2.5}, PM_{10}, NO_2, SO_2, CO$, and others have certain similarities in their sources of emission and hence tend to behave similarly in their variability over time [3].

The hurdles in dealing with these complexities are best exemplified in the varied landscapes of **Bangladesh** and **China**, where climate and growth rates change dramatically from one city to another. In fact, in order to understand air pollution in these areas properly, we have to look into models that operate on the city scale and are capable of reflecting the variety and interactions between pollutants [4]. In that context, our paper presents a **short-horizon forecasting model** for **multiple pollutants** in cities in Bangladesh and China—two vastly different countries, as depicted in **Figure 1.1**.

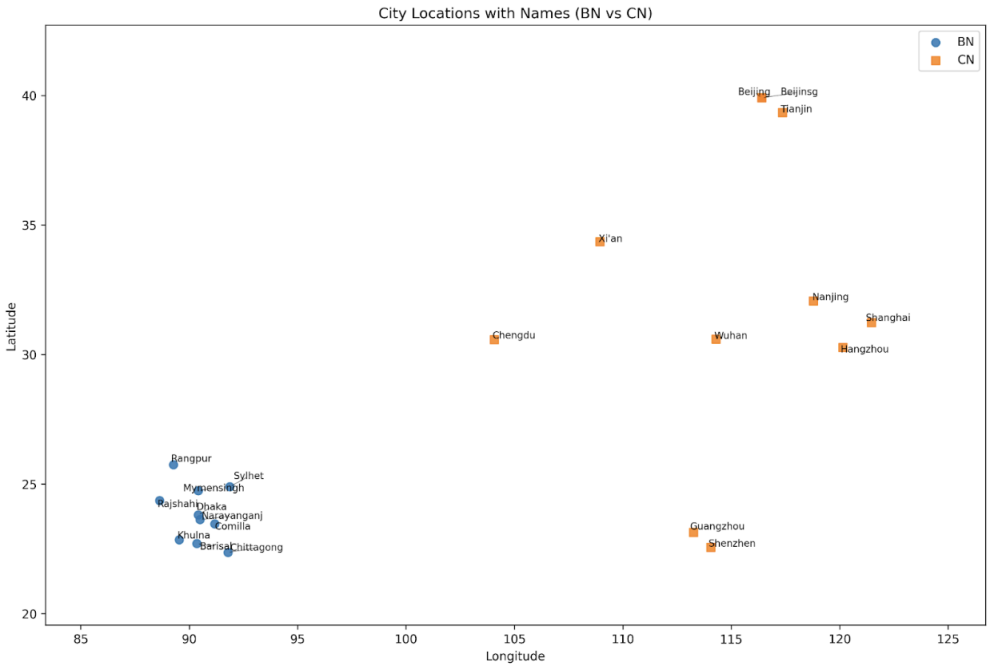

Figure 1.1: Geographic Distribution of Study Cities.

1.2 Motivation

Even though long-term measures and regulations are crucial for air quality improvement, they are insufficient for managing short-term exposure risks. Urban authorities and citizens require air quality forecasts for a short horizon, which means forecasts for a short period of time, usually up to one day. Therefore, there is a need for timely public health advisories and short-term responses to air pollution. However, the usefulness of **short horizon air quality forecasts** depends on their **accuracy**, **reliability**, and **credibility** at the urban scale. However, making reliable air quality forecasts for a short horizon is a complex task due to various interactions between emission processes, meteorological conditions, and time-related activity patterns. This has led to the use of data-based approaches for making air quality forecasts using machine learning and deep learning techniques, which can handle complex time-related patterns despite various variables [5], [6].

Another source of motivation arises from the limitations of existing forecasting systems for air pollutants. Most studies are based on developing models for forecasting a single pollutant, with $PM_{2.5}$ being the most common. However, in reality, the major urban pollutants have common emission sources, as well as common atmospheric processes acting on them. This provides opportunities for exploring commonalities that could result in better forecasting models for air pollutants in general. However, it is also important that the model be robust enough for use in different, heterogeneous environments, as a model that works in one place does not necessarily mean that it will work in another, such as in Bangladesh or China[4], [7].

Research journey context (P1 → P2 → P3). This work is conducted as a four-member group thesis and has been developed progressively since Phase/Pre-Thesis

I. The research evolved from establishing **leakage-safe data** handling and **baseline forecasting models**, to investigating **sequence-based deep learning** approaches for improved **temporal** representation, and finally to a consolidated **forecasting pipeline integrating multi-task prediction, cross-country transfer learning evaluation, uncer- tainty estimation, explainability, and early-warning analysis**. This progression re- flects a deliberate effort to ensure that the final framework is both methodologically rigorous and aligned with practical short-horizon forecasting needs [7]. Our whole research roadmap and our work through the full thesis is summarized in **Figure 1.2**.

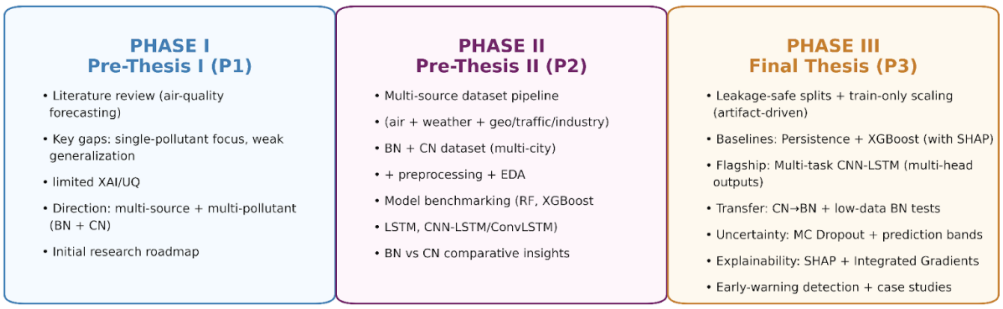

Figure 1.2: Research Roadmap: Phase I → Phase II → Phase III.

1.3 Problem Statement

The problem addressed in this thesis is that of **short-horizon city-scale forecasting** of multiple air pollutants, where these are done within operational and time series limits. That is, using a short section of historical data from a city, we seek to forecast future levels of multiple related pollutants. This is similar to operational conditions in that predictions can only be based on historical data, with the validity of reported results depending on **leakage-safe evaluation** [1], [2].

Suppose you are given an input window of **width W**, which is designed to capture the latest **multivariate** observations for a city. At each time step, diverse information is packed from various sources, such as pollutant levels $(PM_{2.5}, PM_{10}, NO_2, SO_2, CO)$, **weather information**, **urban context information**, and **time information**. The aim is to learn a function which, given such information, predicts the future pollutant levels in multiple dimensions. This is defined as a **multi-output time series regression task**, and the model needs to understand **(i) the temporal behavior of each target** and **(ii) the relationships between multiple targets resulting from shared sources and mechanisms** [3], [8].

This environment presents a number of challenges. **First**, the input domain is necessarily heterogeneous, as the variables represent different physical phenomena, have different units, different noise characteristics, and different distributions. Integrating all these signals into a single forecasting system is not easy, particularly because the external factors such as weather can influence the diffusion of pollutants and their concentration in complex ways. **Second**, of course, there is the issue of the non-linear and time-varying relationship between the inputs and the concen-

tration of pollutants, which makes linear or rule-based approaches inadequate and learning-based approaches attractive as they have the ability to learn these complex patterns[9], [10].

Third, the air quality data exhibit strong temporal dynamics, including daily and seasonal variations, occasional sharp increases due to pollution events, and lagged responses to external factors. Therefore, for short-term prediction, we need models that can effectively learn both short-term and long-term temporal dependencies in the sequences, a requirement that is consistent with previous works on deep learning for sequence-based and hybrid deep forecasting methods[11].

Finally, the last but most crucial challenge to overcome is that of **leakage-free evaluation**. In fact, the very essence of the forecasting task is to predict the future based on what we observed in the past. Hence, careless partitioning of the data, like the use of randomness in partitioning the data, can lead to the leakage of future information into the model, thereby artificially inflating the **performance metrics**. In fact, what we need to ensure is that the partitioning of the data is time consistent [2].

Therefore, the problem that the present thesis aims to solve is that of developing a **leakage-free, multi-output, city-level forecasting system** that can jointly predict **multiple correlated pollutants** with **heterogeneous multivariate inputs** in an appropriate way that can withstand very stringent time tests [12].

1.4 Research Gaps

Although there have been substantial advances in air quality forecasting, there are still some limitations to be addressed for its practical use and scientific credibility. Reviews and comparisons conducted in recent times have again emphasized the limitations in **generalization, uncertainty handling, interpretability,** and **evaluation rigor**, especially with an increased focus on **forecasting models** from isolated tests to practical decision-support models for diverse urban settings [1], [2].

Single-pollutant emphasis and limited multi-output learning. A substantial part of earlier research has been focused on single-target forecasting, especially for one or more instances of $PM_{2.5}$ concentrations, with independent single-output models. However, there is one major flaw in such an approach, as it overlooks an important fact: in reality, major air pollutants are correlated with one another rather than being independent entities. For instance, particulate matter such as $PM_{2.5}$ and PM_{10}, as well as gaseous pollutants such as NO_2, SO_2, and CO, are often associated with common sources and atmospheric processes, thus displaying correlated patterns in their concentrations over time. Ignoring such interdependencies among pollutants may undermine the practical value of forecasts for general air quality assessment[3], [8].

Weak cross-city and cross-country generalization evidence. Much of the forecasting research has only been performed within a single city or a very limited

geographic area, and while success in the area does not necessarily mean robustness for different weather, emission, city, and measurement configurations, the lack of extensive testing in different cities and nations means we cannot be sure the methods will generalize beyond their original context. The more recent research in **cross-city modeling** and **transfer learning** has clearly shown that **generalization** must be evaluated, rather than assumed, especially when transferring to different city contexts[7], [12].

Limited integration of uncertainty-aware forecasting. The majority of the applied research only evaluates point predictions, but in the real world, decisions regarding air quality are always made in the context of risk. The same level of predicted concentration can prompt different courses of action depending on the level of **uncertainty**, especially during extreme cases or when the data distribution is changing. Although the estimation of uncertainties, such as Bayesian approximations through dropout and ensemble methods, has been extensively explored for machine learning, the inclusion of robust uncertainty estimation for end-to-end air quality forecasting is still inconsistent [13].

Insufficient interpretability for deep forecasting models. While improving the accuracy of predictions, deep sequence models remain black boxes. Many studies focus on improving the performance of their models but fail to provide adequate insights into the factors that influence their predictions, the significance of the temporal context, and how their learned relationships align with real-world domain knowledge. Although various tools like **SHAP, LIME**, and attribution methods like **gradient-based methods exist, interpretability** is often underutilized or used in a way that is inconsistent, thereby reducing transparency and undermining trust[14], [15].

Temporal evaluation weaknesses and leakage risk. The major problem with the research is how the dataset is divided into training sets and testing sets. The dataset is often divided at random, and sometimes the dataset is combined across various cities, resulting in leakage. This may give the impression that the model is better than it actually is, but it is often unclear how good the model is at actually predicting, as is highlighted in various reviews, since leakage in the temporal evaluation is critical since the model is always predicting the future based on the past, which may be changing [2].

Thus, the overall limitations highlight the need for developing a forecasting model that can:

(i) handle the forecasting of multiple pollutants at the same time.

(ii) evaluate the model for its ability to generalize over different cities and countries.

(iii) estimate the uncertainties involved in the prediction.

(iv) provide interpretability of the results.

(v) follow the leakage-free temporal evaluation.

This is important for developing a better urban air quality forecasting model, which can be relied upon for decision-making purposes.[3], [13]

1.5 Research Objectives

This is a group thesis project, and its focus is the development and validation of a **leakage-safe short horizon air quality forecasting system** for **city-scale real-world** use. The objectives are high accuracy, extensive generalization tests, quantification of uncertainties, and explanation-based support for decision-making. The objectives are:

1. **Multi-pollutant forecasting:** Develop a forecasting system that forecasts multiple correlated pollutants: $PM_{2.5}, PM_{10}, NO_2, SO_2$, and $\boldsymbol{CO}$, based on recent multivariate data from city-scale measurements. Instead of focusing on individual pollutants, utilize the correlations between them [3].

2. **Cross-country robustness and generalization:** Rigorously evaluate the forecasting system across a variety of scenarios at the scale of cities, with special focus on how transfer learning is applied between Bangladeshi and Chinese cities. Evaluate how the behavior of forecasts changes with changing weather conditions and emission scenarios [16].

3. **Uncertainty-aware prediction:** In prediction problems, let's also provide uncertainty quantification along with point estimates, allowing us to have a sense of how confident our model is. This can be useful when making decisions, especially when pollution events are unusual or when data distributions may change [17].

4. **Model interpretability and explanatory analysis:** Provide transparent reasoning on its behavior by conducting feature importance and temporal influence analysis on traditional machine learning models as well as deep learning models using existing explainability techniques [18].

5. **Early-warning utility for high-pollution episodes:** Develop an early warning system that identifies potential high pollution conditions based on forecast results and corresponding uncertainties, thus making it applicable in downstream uses such as air quality management and health advisory systems. [1], [5].

1.6 Research Contributions

Key takeaways from this group thesis:

1. **Leakage-safe, multi-source dataset pipeline:** In this section, we present a unified workflow for preparing the dataset by combining air quality information, meteorological information, city information, and time encodings for major cities in **Bangladesh** and **China**. The workflow incorporates leakage checks for **schema**, **missing value**, and **consistency issues** and utilizes **leakage-safe train-test splits** to facilitate credible real-time predictions [1], [5].

2. **Multi-task deep forecasting model for correlated pollutants:** In this section, we propose a novel multi-task CNN-LSTM model to leverage the correlations between $PM_{2.5}, PM_{10}, NO_2, SO_2$, and CO. The multi-task learning architecture learns cross-pollutant patterns in a unified way for air quality forecasting [9].

3. **Cross-country transfer learning evaluation:** We incorporate uncertainty quantification into short-horizon air quality prediction using **Monte Carlo Dropout**, producing predictive distributions rather than only point estimates. This enables risk-aware interpretation of forecasts and sup- ports decision-making under uncertainty. Uncertainty results are presented in **Table 6.4 & Figure 6.5** [19].

4. **Uncertainty-aware forecasting via Monte Carlo Dropout:** In addition, we improve the **explainability** and trust of our models by applying various explanation techniques. **SHAP** is used to explain **tree-based baseline models**, while **Integrated Gradients** is used to explain **deep model** attributions and feature relevance and temporal impact on predictions. The results of explainability are presented in **Figure 6.6** [20].

5. **Model explainability across ML and DL predictors:** We apply explanation methods to increase transparency and trust. Specifically, we use **SHAP** to interpret tree-based baselines and **Integrated Gradients** to analyze deep model attributions, providing evidence of feature relevance and temporal influence on pollutant predictions. Explainability results are summarized in **Figure 6.8** [21].

6. **Early-warning episode detection using forecast outputs:** In addition, we design an early warning system based on the forecast models and connect the dots between predictive modeling and the intended use cases and goals of the system. Episode detection examples are presented in **Figure 6.11** [22].

1.7 Paper Organization

The remainder of this thesis is organized as follows. **Chapter 2** reviews the literature on air quality forecasting, from traditional machine learning to deep learning, and from transfer learning to uncertainty estimation and explainable AI methods that are relevant to this thesis. **Chapter 3** describes the construction of the dataset, the feature sources that were integrated, the preprocessing, a safe approach to data splitting, and the formal framework for multi-output forecasting. **Chapter 4** presents the approach that is proposed, including the baseline predictors, the multi-task CNN-LSTM model, and the additional components for transfer learning, uncertainty estimation, interpretability, and early warning analysis. **Chapter 5** describes the experimental setup, training, and evaluation. **Chapter 6** presents the results on forecasting performance, transfer learning, uncertainty and interpretability analysis, and case studies. **Chapter 7** interprets the results, their implications, limitations, and validity. Finally, **Chapter 8** concludes the thesis and outlines directions for future work.

Chapter 2

Related Work

2.1 Classical Machine Learning for Air Quality Forecasting

Traditional machine learning techniques remain prevalent in air quality prediction. This is because they perform well on structured data that can be feature-engineered into table format. Moreover, they are often more interpretable than many deep learning models. Surveys have shown that regression models and tree-based models remain a significant part of operational as well as research-oriented air pollution prediction systems. This is because they perform robustly on noisy data and integrate smoothly with different predictor variables such as weather, time, and urban factors [1], [23]

In early days, statistical regression models, as well as other exposure models such as land use regression models, were popularly used to estimate concentrations of different pollutants using meteorological parameters as well as other environmental factors. This is because they are easy to implement and interpret. However, linear models are not sufficient in cities where complex interactions of weather, emissions, and other factors cause pollutants to behave non-linearly. In the course of machine learning research development, there is a clear shift away from linear approaches and towards non-linear approaches, particularly ensemble methods. Tree-based models have been particularly effective at identifying non-linear relationships that are not captured by linear approaches [2], [24].

Within this category, Random Forest is particularly effective as a baseline model. The reason for this is simple: ensemble models such as Random Forest are generally effective at decreasing prediction variance and increasing robustness to noisy data, which is particularly prevalent in real-world sensor data where errors or missing data may be present. However, another category of models that have been particularly effective at air quality prediction is gradient boosting models. In particular, XGBoost has been found to be a popular benchmarking algorithm in many studies that involve comparison of different machine learning models in different cities using different time scales. This is because it is efficient in boosting models using regularization techniques. Its strong performance has been documented in many studies that involve comparison of different machine learning models in different cities using different time scales [2], [25].

The practical advantage of classical machine learning is its willingness to confront the policy implications of its predictions—through "looking inside" the models, so to speak, with various forms of post-hoc explanations. For example, tree ensembles can provide clear global rankings of feature importance, and research on explainability is increasingly advocating for reasoning at the feature level as part of good environmental modeling practices[1], [14]. Additionally, classical ML is often more data-efficient than deep sequence models, which is certainly a positive in some fields with scarce historical data or patchy reporting [6], [26].

Of course, there are structural limitations in classical ML that propel us toward more sophisticated approaches. To begin with, there is a strong single-target nature to most classical ML implementations—you train a separate model for each pollutant, which ignores interactions between pollutants and can squander learning potential if pollutants share common meteorological drivers [25]. Second, classical ML often fails to learn time dynamics directly; instead, temporal dependence is often handled with lag features and various forms of feature engineering, which may miss complex multi-step dependencies that become apparent during sudden pollution events or regime changes [2], [27]. These shortcomings encourage more sophisticated forecasting approaches that can learn time dynamics and model multiple pollutants under a single learning objective, as this thesis will explore.

2.2 Deep Learning for Spatio-Temporal Air Pollution Modeling

Deep learning has thus become central in modern air quality forecasting, as it can now learn non-linear relationships and temporal dependencies directly from multivariate time series, minimizing the need for pre-computed lag features. Studies of existing literature indicate that these models perform best when pollutant dynamics are dependent on lagged meteorological variables, as well as complex interactions between variables, which are difficult to model using traditional, shallower architectures [27].

Among these, a prominent class of models for modeling air quality time series is that of Recurrent Neural Networks, with Long Short-Term Memory networks as a subset. The gates in these networks are specifically constructed for handling information retention over larger periods, making them a good fit for modeling hourly air quality time series, where current pollutant levels are influenced by accumulation, dispersion, and persistence over previous hours. In air quality forecasting, LSTM networks, as well as their variations, are frequently reported as strong temporal learning models, especially for short- to medium-term horizons [28].

Apart from these, Convolutional Neural Networks have also been explored as alternatives for modeling time series, using 1D convolutional layers for local patterns, such as short-term patterns and sudden changes, with efficient computation. Temporal CNNs are thus strong models for capturing local dependencies in time series, but for modeling longer-term dependencies, deeper architectures are typically re-

quired. As such, a new class of models that combine CNNs with RNNs has also become prominent in air quality research [19], [21].

The complementarity of these representations is explicitly exploited in hybrid CNN-LSTM models, where convolutional layers can be seen to learn compact representations of local temporal signals, while recurrent layers can pick up on longer-term dependencies in these features. One such notable example is that of DeepAir, where CNNs and LSTMs are combined in a hybrid structure that is applicable to fine-grained air quality forecasting, thus exemplifying the added value of hierarchical temporal representation learning in pollution prediction applications [29], [30].

In cases where spatial structure is present, deep learning models may extend from purely temporal analysis to spatiotemporal models. One such extension is based on the assumption of a grid structure in space, where sequences can be arranged in structured tensors. In such cases, ConvLSTM can be used to embed convolution into recurrent gates. More recent hybrid models of ConvLSTM, such as those that incorporate attention in residual ConvLSTMs, have been proposed to increase robustness in spatiotemporal models, although they are contingent on structured spatial data being available in grids and sufficient computational resources being available for these models to be deployed in practice [24], [31].

A second approach to spatiotemporal models is based on using graphs to represent spatial relationships, which is more applicable in citywide station networks or regional graphs. Graph-based deep learning models of pollution forecasts combine spatial message passing with temporal learning to incorporate representations of pollutant transport and correlations between different locations. Hybrid spatiotemporal graph-based deep learning models were introduced in foundational research in this domain, followed by more recent proposals of graph convolution recurrent models applicable to citywide pollution forecasting, where such models have been shown to be more realistic in practice [10], [32].

Recently, attention mechanisms and the Transformer architecture have made their way into spatiotemporal air quality modeling. This has improved the ability to model long-range dependencies and various inputs. Research on the Transformer architecture has shown improved forecasting capabilities with flexible attention modeling over time and various data inputs. More recent architectures, such as Airformer, take this to the next level by targeting multi-pollutant spatiote [8], [33].

However, even with their capabilities, deep models have their own trade-offs. The training and generalization capabilities of deep models are often dependent on the size of the dataset, regularization, and proper evaluation metrics. In critical applications of environmental models, deep learning models have been criticized for their lack of interpretability. These points raise the need for complementary components, including multi-task learning for cross-pollutant modeling, transfer learning for improved generalization across regions, uncertainty quantification for model reliability, and explainability for interpretability, which are discussed in the next sections of this chapter [14].

2.3 Multi-Target and Multi-Task Learning

Urban air pollutants do not act alone. They have common sources, roam through the same atmospheric dynamics, and interact with each other. Therefore, modeling each pollutant separately might overlook the inter-pollutant patterns that actually exist and are useful for modeling, particularly when weather conditions simultaneously affect all of them or when their presence lingers in the air for a considerable period of time. Through all the reviews, the point is clear: the problem is rich in interactions, and the best techniques should leverage the patterns that are shared instead of modeling each target separately [34].

This is where multi-target and multi-task learning (MTL) comes in. This method deals with the problem setup by learning multiple related prediction tasks simultaneously, using a shared representation. In a standard hard-parameter-sharing configuration, a shared backbone extracts features from the input sequence, and small task-specific heads produce the predictions for each pollutant. This shared representation has a natural tendency to improve sample complexity and generalization by being compelled to extract features that generalize across tasks, acting as a form of implicit regularization in comparison to training separate models for each pollutant. This collective learning is particularly important in air quality forecasting, where multiple pollutants are often simultaneously affected by the same set of meteorological and temporal factors [25].

Recently, there has been an increasing trend in air quality studies towards multi-output forecasting. transformer-based and attention-driven spatiotemporal models are explicitly formulated for multi-pollutant forecasting, demonstrating that multi-target learning is not only theoretically justifiable but also beneficial in practice when sufficient data and proper assessment are available. A recent example of this trend is Air-former, an attention-driven transformer model specifically designed for multi-pollutant spatiotemporal forecasting, emphasizing the shift towards modeling techniques that can handle multiple pollutants simultaneously rather than modeling separate pipelines [8].

Outside these simple multi-output scenarios, there are task-specific architectures for multitask learning, where multitask learning is central to the architecture, rather than a side effect of learning multiple outputs. In a recent preprint, a deep multitask spatiotemporal architecture with attention is proposed for air quality indicators forecasting in megacities. This illustrates how multitask learning can be used to better leverage the complex relationships between air quality indicators [3].

However, for environmental forecasting, multitask learning poses some challenges. When tasks are associated with different levels and types of noise, multitask learning can result in negative transfer between tasks. Thus, careful decisions need to be made on how loss functions are designed and optimized. As a result, studies on multitask learning for pollutant forecasting highlight methodological rigor and evaluation when claiming performance improvements due to advanced learning methods [25].

In summary, multi-target and multitask learning present a sound approach for forecasting multiple air quality indicators based on their complex relationships. The trend towards multi-pollutant spatiotemporal predictors and multitask deep learning for air quality indicators forecasting underscores pollutant forecasting as a task worthy of joint treatment and, therefore, as a primary choice for modeling in this thesis [3], [8].

2.4 Transfer Learning in Environmental Time Series

One of the major issues that environmental time series forecasting faces is the availability of data. In some cities, there is an adequate amount of historical data available, while in others, the data is scarce. In the case of air quality prediction, the situation is further complicated by the different densities of air quality stations in different cities. Furthermore, the management of these stations in different countries can vary. In these circumstances, training the model from scratch can lead to overfitting. In the recent reviews of machine learning in air pollution prediction, generalization is highlighted as an area that needs to be addressed in machine learning models. Furthermore, the reviews emphasize the importance of reusing the knowledge that can be acquired from the data in the source domain [35].

Transfer learning is an approach that can address the limitations of training the model from scratch. In the case of air quality prediction, the approach of transfer learning can be implemented by utilizing the neural predictor. In the recent study, the researchers highlighted the potential of neural transfer learning in the context of air pollution prediction. They framed the problem of air pollution prediction as the neural transfer learning problem. The study indicated that the robustness of the prediction can be enhanced with the transfer of the learned knowledge [7].

In the case of transfer learning, the source city is usually similar to the target city to some extent in terms of the dynamics of the city. However, in the case of the same countries, the situation can be more complex. In these circumstances, the distribution can vary between the source city and the target city. Recent studies on the prediction of air quality in different cities with the help of the graph neural predictor emphasize the importance of generalization between the cities. They highlighted that generalization between the cities should not be considered an afterthought [12], [18].

This is because cross-country transfer involves an added layer of complexity, given that climate regimes, seasonal patterns, regulations, and emission compositions can change dramatically from one domain to another. This means that there is more significant covariate shift and concept drift. Thus, the simple re-usage of parameters without proper attention can lead to negative transfer unless the adaptation process is carefully considered. Recent work in remote sensing and ground fusion for air quality prediction points to the potential of cross-domain spatiotemporal transfer learning in the prediction of PM2.5 concentrations. This shows the potential of domain-aware transfer in minimizing the distribution mismatch between the source

and target domains when the domains differ in terms of data collection characteristics and generation mechanisms [19].

Transfer learning can not only improve the accuracy of the model but can also lead to faster convergence rates and the minimization of the required training data for deep models in the prediction of complex phenomena. However, the literature shows that the benefits of transfer learning can vary depending on the evaluation criteria. For example, the study in comparative forecasting shows that the benefits of transfer learning can only be established with the adoption of disciplined experimental protocols in the assessment of the performance of the model. This is because the performance can vary depending on the size of the domain adaptation and the adaptation methods used in the transfer learning process [25].

In conclusion, the concept of transfer learning aligns well with the requirements of air quality prediction in diverse environments, given the limitations associated with the availability of data in different regions. Thus, the realities surrounding the limitations of the generalization of the model in diverse environments form the basis of the assessment of the potential of transfer learning in the following chapters (Section 4.6) instead of the common notion that the model can generalize well in different environments unless the training is done in the particular region of interest [12], [25].

2.5 Uncertainty Quantification in Forecasting Models

Many papers on air quality forecasting provide point predictions and optimize them based on criteria such as RMSE and MAE. However, in practice, there are several sources of uncertainties associated with air quality forecasting: sensor noise, missing or inaccurate observations, random behavior of atmospheric transport and chemical reactions, and changes in distribution due to seasonal or abnormal emission rate changes. In the reviews of machine learning techniques for air pollution forecasting, accuracy is a prominent topic while reliability is relatively less emphasized [18], [25].

Uncertainty Quantification (UQ) addresses the problem of reliability by assigning a measure of reliability to predictions. This converts a point predictor into a reliability-based predictor. This is particularly important for decision-making under uncertainty in environmental applications, where actions are typically based on threshold criteria such as the anticipation of a certain threshold being crossed in the near future. In discussions on recent research in air quality forecasting, there is an emphasis on how predictors should be evaluated based on criteria such as reliability rather than point prediction accuracy [25], [36].

Several UQ techniques have been proposed for deep learning-based predictors. Deep ensembles have been a very popular technique for UQ as they can estimate epistemic uncertainties through variability in predictions made by different models. However, deep ensembles require training different models and require more memory space. Bayesian neural networks provide a probabilistic framework for predictors and can

easily handle epistemic and aleatory uncertainties. However, there is a drawback: there is an additional computational cost associated with optimizing the network [11].

One simple and easily implementable method is Monte Carlo (MC) Dropout. MC Dropout treats dropout even in inference as a form of Bayesian inference. By keeping dropout on while making a prediction and performing multiple forward passes, MC Dropout provides a predictive mean and a spread, which can be easily converted into uncertainty bands without altering the architecture of the model. MC Dropout has already been used to estimate the uncertainty associated with deep models while remaining light enough for repeated use across various locations and pollutants [11].

Despite all these tools being available, predictions made with consideration of uncertainty are less used than those based on accuracy, especially when deep learning is used to minimize the average error. According to most reviews and comparative studies, high accuracy at the point level is not enough to ensure reliable behavior, especially when extreme events occur, hence the need to use uncertainty estimation in the forecasting process [1], [3].

Essentially, the inclusion of uncertainty estimation changes the process of air quality forecasting from a simple prediction process into a reliability-based decision-making tool, which is especially important when dealing with urban air pollution since the predictions are used for planning based on changing atmospheric conditions [11].

2.6 Explainable AI in Environmental Modeling

As air-quality forecasting models become more complex, interpretability has become a central requirement for environmental applications that inform public health guidance, policy planning, and regulatory discussion. Although high-capacity machine learning and deep learning models often improve predictive performance, their limited transparency can reduce trust and hinder scientific usefulness if the drivers of predictions cannot be examined. Recent reviews of air-pollution prediction research emphasize interpretability as a persistent gap, particularly for deep models, and argue that explainability should be treated as a first-class evaluation dimension alongside accuracy [14].

For classical models, especially tree-based ensembles, post-hoc explainability is commonly achieved using SHapley Additive exPlanations (SHAP). SHAP provides both global and local explanations by decomposing a prediction into feature-wise contributions grounded in Shapley value theory. In air-quality modeling, SHAP-style explanations are frequently used to identify dominant meteorological and temporal drivers, compare feature influence across regions, and support policy-relevant interpretation of model behavior. This is particularly aligned with gradient boosting baselines (e.g., XGBoost), where tree-structured SHAP implementations are computationally efficient and can be applied at scale [14], [15].

Interpreting deep learning models is more challenging because their representations are distributed, hierarchical, and strongly nonlinear. As a result, research

increasingly relies on attribution methods that quantify how inputs influence outputs. Integrated Gradients (IG) is a widely used gradient-based attribution approach that assigns importance scores by integrating gradients along a path from a reference input to the observed input, enabling feature-level interpretation without requiring the model to be inherently transparent. IG and related attribution methods are particularly useful for multivariate time series, where they can indicate which predictors (and, in sequence models, which historical segments) contributed most to a forecast[4].

Even though they are beneficial, they have their own set of methodological caveats. The quality of an attribution can depend on the choice of baseline, scaling of inputs, and consistency of explanations. The explanations can also change with different time regimes or geography. For this, there is an increasing need for environmental forecasting studies to encourage the use of explainability with validation and transparent reporting, as this will ensure that the explanations obtained are stable and not due to a particular run or a small evaluation. As discussed earlier, comparative forecasting also emphasizes that any statement on interpretability should be supported with rigorous experiments and reporting [25].

In environmental modeling, the use of explainable AI is not just about "opening the black box." It is about linking the predictions with relevant signals within the domain, making it accountable. Thus, using explainability with forecasting models allows us to evaluate the usefulness of a model holistically, as it is not just about the accuracy of a model but also about the reasons behind its behavior with respect to different pollutants and cities. This is the rationale behind the use of explainability with tree-based baseline and deep forecasting models as part of this thesis [1], [14].

2.7 Early-Warning Episode Detection and Leakage-Safe Evaluation

Forecasting air quality shines brightest when it leads to tangible actions, not just when it is robust on average. In practical forecasting applications, users are more interested in whether a forecasting system can identify high pollution events—periods of elevated or rapidly increasing concentrations that pose greater health risks—than in reducing small errors on typical days. This shift from point forecasting to event-oriented utility is what drives early warning systems to translate continuous forecasts into event-related alerts that correspond to public health objectives and monitoring requirements [25], [35].

A common approach is to use episode detection based on threshold crossing, where forecasts are compared against pollutant-specific thresholds (e.g., official thresholds or percentile-based thresholds), and a series of exceedances are joined together to form meaningful events. The way we define these events is important because single isolated breaches of a threshold can be innocuous, while similar events in nearby regions of space and time indicate genuine risks of exposure. Thus, assessment needs to extend beyond RMSE/MAE and include event-related metrics such as detection rate, false alarms, timing of detection, and lead-time analysis. This is more com-

plementary to traditional accuracy metrics and reflects real-world needs for early warning systems more accurately [2], [33].

However, many air quality forecasting papers remain focused on point regression metrics and fail to investigate whether the predicted peaks are consistent with actual episode definition. This can obscure failure modes in which a forecasting system appears robust on average but systematically biases low or misses the rapid transition. Recent advances in deep learning models, such as attention and transformer models for spatiotemporal forecasting, improve representation capacity, but the literature tends to conflate "better forecasting" with "lower average error" rather than enhanced event sensitivity within realistic system constraints [8], [33].

The fundamental concept remains the same: ensure that your evaluation remains fair and true, especially when time is limited. In time series forecasting, for instance, "sneaky" temporal leakage can occur and improve model performance. If you randomly permute your data, split your data non-chronologically, or pre-process your entire set of data using certain preprocessing techniques, you are essentially leaking information into your training set. This causes your model to behave in a way that doesn't generalize well into real-world situations. The problem is further complicated with datasets that have multiple cities, as shared seasonal patterns, windows, and normalization techniques can create back doors for information leakage without even using your target variable. Therefore, methodologically strict studies encourage temporally consistent data splits and preprocessing as the benchmark for credible forecasting claims [2], [24].

The process for a leakage-conscious evaluation process essentially follows three steps. First, you must ensure that your validation and testing sets are chronologically separated from your training set. Next, you must ensure that your preprocessing techniques are applied only on your training set and that you don't "leak" information into your validation and testing sets. Finally, you must ensure that your windows are constructed for sequence models in such a way that they don't "leak" into your validation and testing sets from your training set. All these steps ensure that your model is as useful as you want it to be while reducing the likelihood of overly optimistic evaluations [25], [37].

The problems with event-based evaluations and leakage-conscious experimentation essentially call for an additional layer for detecting early episodes in your forecasting process and using a temporally consistent evaluation process [1], [35].

2.8 Summary and Positioning of This Work

Existing literature has demonstrated considerable success in air quality prediction using traditional machine learning techniques, deep learning methods, and spatiotemporal approaches. However, most works in this area seem to focus on individual components without making considerable attempts at combining multi-pollutant approaches, cross-region generalization, uncertainty quantization, and interpretability into a single framework. Most works concentrate on single-pollutant approaches, narrow geographical scopes, or accuracy maximization without methodological rigor

in terms of time series evaluation and applicability.

This thesis attempts to combine these different threads and create a novel air quality forecasting framework. By considering air quality prediction as a multi-pollutant time series forecasting problem, it aims at making a more comprehensive approach by incorporating correlations between different air quality parameters. Additionally, cross-country transfer learning is employed to evaluate the cross-region generalization capability of the models. Furthermore, uncertainty quantization and interpretability are incorporated into a single framework, making it a more comprehensive approach compared to existing works. By doing so, it aims at taking air quality prediction a step forward.

Chapter 3

Dataset and Problem Formulation

3.1 Data Sources and Study Area

This research aims to examine the accuracy of short-term, city-level air quality forecasting in two vastly contrasting national environments—Bangladesh (BN) and China (CN). This will not only test the accuracy of the forecasting model in contrasting environments but also allow for a comprehensive and precise comparative analysis between the two countries. Bangladesh represents an urbanizing and relatively data-scarce environment, while China represents a relatively data-rich environment. By comparing these two vastly contrasting environments using the same framework, we will be able to (i) account for the differences between the two cities, (ii) account for the differences in the distribution of the data for both countries, and (iii) address the applicability of the knowledge and the model in transferring from a relatively data-rich to a relatively data-scarce environment, which is an important concern in the context of the spatiotemporal modeling of air pollution and regional studies, as discussed in prior research [2], [7].

The research will cover ten major cities in Bangladesh—Dhaka, Chittagong, Sylhet, Rajshahi, Khulna, Barisal, Rangpur, Mymensingh, Comilla, and Narayanganj—and ten major cities in China—Beijing, Shanghai, Guangzhou, Shenzhen, Wuhan, Chengdu, Xi'an, Tianjin, Hangzhou, and Nanjing. The selection of the cities was made such that the dataset covers a broad range of realistic variations in the scope of the model and the meteorological variables. The geographic locations of these cities are illustrated in **Figure 1.1**.

To construct a uniform forecasting dataset for all sites, we developed a multi-source data pipeline that integrates pollutant data with meteorological information as well as urban context information. On the pollutant side, we focus on five major pollutants that are typically monitored, namely $PM_{2.5}, PM_{10}, NO_2, SO_2$, and CO. We use open interfaces for monitoring these pollutants, which are then integrated into a single record for each city. We also collect meteorological information such as temperature, humidity, wind speed, and wind direction, as these factors significantly impact particulate as well as gaseous pollutants [2], [6]. We also collect urban context information such as population density, traffic intensity, as well as proximity to industrial activities, as these factors are also important in understanding urban pollution dynamics, especially in the absence of emission inventory information [8],

[11], [26], [32]. With these components, we are able to construct a heterogeneous yet structured dataset that allows for learning city-specific patterns while at the same time allowing for transfer-oriented evaluation for BN and CN settings [12], [16], [19].

3.2 Dataset Schema and Feature Groups

To support consistent forecasting across 20 heterogeneous cities and multiple data streams, we organized the constructed dataset using a role-driven schema that separates (i) identifiers and temporal context, (ii) model inputs grouped by physical meaning, and (iii) multi-pollutant targets. This design keeps the learn- ing problem well-defined while retaining interpretability when comparing behaviors across cities and across the two countries. A complete variable list with roles and categories is provided in **Table 3.1**.

Dataset Summary

rows	cities	countries	start	end	country_row_counts	top5_city_row_counts	total_cells	missing_cells
5680	20	2	01/06/2025 - 6:28:45 AM	25/01/2026 - 11:59:24 PM	BN:1607	Sylhet:163	126920	13192
					CN:1566	Dhaka:162		
						Rajshahi:162		missing_rate
						Chittagong:161		0.103939489
						Barisal:161		

Missingness Summary

column	dtype	missing_count	missing_rate
data_source_air	float64	3173	1
data_source_weather	float64	3173	1
data_source_traffic	float64	3173	1
data_source_industrial	float64	3173	1
source_file	object	500	0.15758
timestamp	datetime64[ns]	0	0
city	object	0	0
country	object	0	0
latitude	float64	0	0
longitude	float64	0	0
hour_of_day	int64	0	0
day_of_week	int64	0	0
month	int64	0	0
season	object	0	0
population_density	int64	0	0
pm2_5	float64	0	0
pm10	float64	0	0
no2	float64	0	0
so2	float64	0	0
co	float64	0	0
air_quality_index	float64	0	0
temperature	float64	0	0
humidity	float64	0	0
wind_speed	float64	0	0
wind_direction	float64	0	0
aqi	float64	0	0
proximity_to_industrial_areas	float64	0	0
traffic_density	float64	0	0
air_quality_category	object	0	0
year	int64	0	0
day	int64	0	0
hour	int64	0	0
dayofweek	int64	0	0
is_weekend	int64	0	0
hour_sin	float64	0	0
hour_cos	float64	0	0
dow_sin	float64	0	0
dow_cos	float64	0	0
month_sin	float64	0	0
month_cos	float64	0	0
		13192	4.15758

Dataset Schema

feature	role	category
timestamp	Metadata	Time
city	Metadata	Metadata
country	Metadata	Metadata
latitude	Input	Other
longitude	Input	Other
hour_of_day	Input	Other
day_of_week	Input	Other
month	Input	Other
season	Input	Other
population_density	Input	Other
pm2_5	Target	Air Pollutant
pm10	Target	Air Pollutant
no2	Target	Air Pollutant
so2	Target	Air Pollutant
co	Target	Air Pollutant
air_quality_index	Input	Other
temperature	Input	Meteorology
humidity	Input	Meteorology
wind_speed	Input	Meteorology
wind_direction	Input	Meteorology
aqi	Input	Other
proximity_to_industrial_areas	Input	Other
traffic_density	Input	Traffic
air_quality_category	Input	Other
data_source_air	Input	Other
data_source_weather	Input	Other
data_source_traffic	Input	Traffic
data_source_industrial	Input	Other
source_file	Input	Other
year	Input	Other
day	Input	Other
hour	Input	Other
dayofweek	Input	Other
is_weekend	Input	Other
hour_sin	Input	Other
hour_cos	Input	Air Pollutant
dow_sin	Input	Other
dow_cos	Input	Air Pollutant

City Row Counts

city	rows
Sylhet	163
Dhaka	162
Rajshahi	162
Chittagong	161
Barisal	161
Rangpur	161
Khulna	161
Mymensingh	160
Comilla	159
Shanghai	158
Guangzhou	158
Beijing	157
Shenzhen	157
Narayanganj	157
Chengdu	156
Nanjing	156
Hangzhou	156
Wuhan	156
Tianjin	156
Xi'an	155

Country Row Counts

country	rows
Bangladesh	1607
China	1566

Table 3.1: Dataset Schema

Each record corresponds to a city-level observation aggregated at the acquisition time, yielding 5,680 rows across 20 cities from two countries, with overall cov- erage from 2025-06-01 to 2026-01-25 (Figure 3.1). This period provides sufficient seasonal diversity for evaluating short-horizon forecasting under varying meteorological regimes and emission conditions.

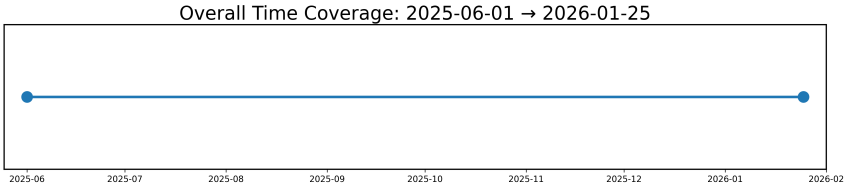

Figure 3.1: Temporal Coverage of Dataset

(i) Metadata and spatial context: The data includes metadata such as city, country, and exact acquisition time. Additionally, latitude and longitude information are added. These metadata will be used for stratified checks and cross-city analysis, and for correctly associating data with acquisition batches.

(ii) Temporal features: Time series for air pollution data contain strong periodic characteristics due to daily human activities and seasonal changes. We add calendar information like hour_of_day, day_of_week, month, and season. We then add their corresponding cyclic representations like hour_sin, hour_cos, day_sin, day_cos, month_sin, month_cos. We include a weekend flag as well: is_weekend. These additional time-related information will help the model catch periodic characteristics and make accurate short-term forecasts [1], [2].

(iii) Meteorological features. Weather conditions affect how air pollutants disperse and accumulate. Weather conditions may even affect chemical transformation. We include basic meteorological information like temperature, humidity, wind speed, and wind direction. These will help catch these dependencies and make the approach transferable across cities with different climates [2], [6].

(iv) Urban and anthropogenic proxy features. Direct emissions information is usually scarce and difficult to obtain uniformly across cities and even countries. We include proxy information like population density, traffic density, and proximity to industrial zones. These will help catch persistent urban emissions characteristics and make the approach more transferable across cities with different characteristics [11], [26].

(v) Target pollutants (multi-output). We predict five common air quality pollutants: $PM_{2.5}, PM_{10}, NO_2, SO_2$, and CO. We include all these in a multi-output approach because a comprehensive view of air quality is desired. We will be able to investigate common and pollutant-specific time series characteristics, which is a key aspect in modern air quality multi-pollutant forecasting [3], [8].

There are also auxiliary fields that track the provenance and assist in debugging, such as the air quality index, category, and batch or source fields (e.g., source file and source availability fields). Again, these fields are included for documentation of the methods used to obtain the data and for integrity checks, as described in Section 3.3, and not for definition of the primary forecasting objectives. This grouped schema is an appropriate balance of completeness and simplicity, offering a steady and consistent base for the modeling and evaluation process described in [1], [2].

3.3 Data Integrity and Cleaning Policy

Data integrity is vital for reliable air quality predictions, especially when handling many variables and time series. In fact, for many multivariate data, preprocessing can introduce bias, which can artificially improve the model's performance beyond what is actually achievable. We therefore applied a structured cleaning and validation policy before any sequence construction and model training, em- phasizing (i) auditability across cities, (ii) physically plausible measurements, and (iii) strict leakage prevention consistent with real-world forecasting constraints [1], [2], [25].

A. Schema validation and traceability: All incoming batches were first validated against a fixed schema to ensure consistent column presence, data types, and city/country identifiers. Metadata fields (city, country, spatial coordinates, and acquisition time) were checked for format consistency so that observations remain traceable to their collection batches and can be grouped correctly during cross-city and cross-country evaluation.

B. Missingness auditing and retention rules: Missingness was quantified at the feature level after cleaning and feature engineering to verify that core modeling variables remain reliable across the full study period and across cities.

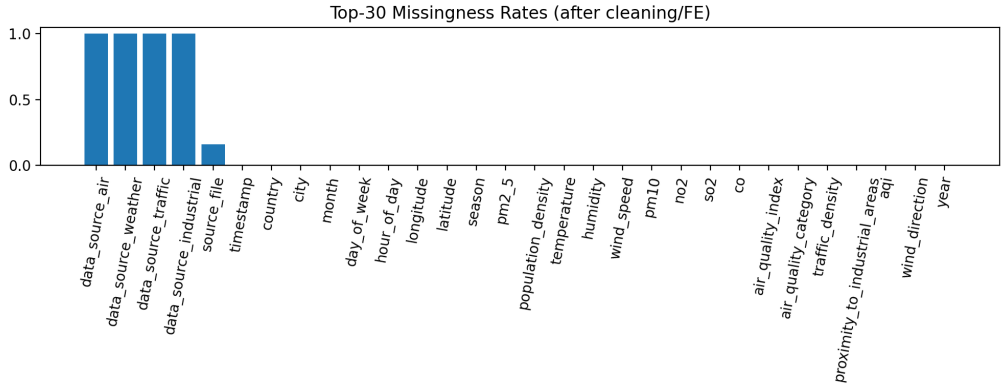

Figure 3.2: Missingness Rate

As shown in *Figure 3.2*, the highest missingness is concentrated in **source/provenance indicators** (e.g., **data_source_air, data_source_weather, data_source_traffic, data_source_industrial**) that act as availability flags rather than physical predictors. In contrast, the principal pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO) and key meteorological covariates exhibit negligible missingness in the finalized modeling table. Based on this audit, we adopt a **conservative retention policy**: provenance fields are retained for traceability even if partially missing, while all predictors and targets required for supervised learning must be present within each constructed window. We avoid imputing long gaps because it can create artificial smoothness and unrealistic temporal dependence; instead, any window with missing values in required predictors/targets is excluded during sequence construction to maintain forecasting fidelity [2], [25].

C. Plausibility checks and outlier treatment: To remove any obvious measurement issues without removing the actual extreme value instances, we use simple

plausibility checks. We remove any data that is clearly incorrect, such as negative pollutant readings, impossible weather readings, and inconsistent wind readings. Aside from these, we avoid the temptation to remove outliers, since these tend to be the most significant readings for decision-makers. Instead, we remove the likely measurement glitches, keeping the actual extreme value readings intact [2], [6].

D. Temporal and structural consistency validation: We validated the internal structure of each city's time series by checking acquisition-time ordering, detecting duplicates, and confirming that records can be aligned into coherent supervised windows. Because the collection pipeline operates as repeated batch snapshots (often near-hourly with natural runtime jitter), we enforce temporal consistency at the **window level**: each input sequence must represent a valid historical context for a future prediction time. Data types and categorical encodings were standardized to prevent silent inconsistencies across merged sources and to ensure that the feature schema remains stable across both countries.

E. Leakage prevention as a hard constraint. Leakage control guided all preprocessing decisions. Temporal splits were established **prior to** any transformation that could learn from data distributions (e.g., normalization or standardization). All scalers and statistics were fitted using **training data only**, Then applied unchanged to validation and test sets. This prevents inadvertent propagation of future information into the training process and ensures that evaluation reflects realistic forecasting deployment conditions rather than post-hoc data access [1], [25].

Collectively, these integrity checks and cleaning rules produce a data set that is auditable, physically plausible, and faithful to evaluation. It reduces artifacts while preserving real variability in urban air pollution across Bangladesh and China [2], [25].

3.4 Exploratory Data Analysis

We conducted exploratory data analysis (EDA) to (i) verify that the cleaned dataset preserves realistic air-pollution variability and (ii) identify distributional and dependency patterns that should be reflected in the modeling design. Given the known non-stationarity of urban air quality and its sensitivity to both emissions and meteorology, EDA also provides an empirical basis for choosing non-linear, multi-output forecasting approaches and for motivating careful evaluation under distribution shifts [1], [2], [25].

A. Distributional characteristics: Figure 3.3 summarizes the marginal distributions of representative pollutants and key meteorological variables.

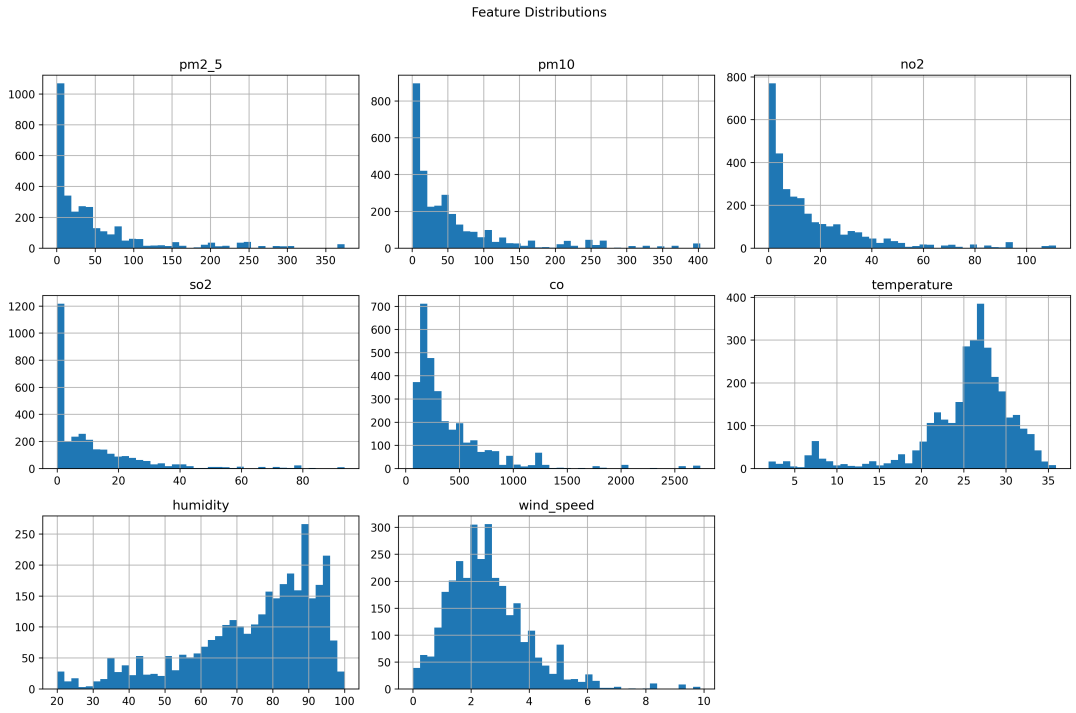

Figure 3.3: Feature Distributions

Across the five targets, pollutant concentrations are **right-skewed and heavy-tailed**, with long upper tails that correspond to episodic events rather than typical background levels. This behavior is especially visible for particulate matter and combustion-related gases, where a large mass of observations lies at low-to-moderate concentrations while a smaller fraction forms pronounced extremes. Such distributions are widely reported in real-world air-quality time series and imply that models should be robust to heteroscedasticity and non-Gaussian behavior rather than relying on linear-normal assumptions [2], [6]. In contrast, meteorological variables exhibit more structured ranges (e.g., temperature concentrated within seasonally plausible bands, wind speed concentrated at low-to-moderate values), reflecting the physically bounded nature of atmospheric covariates that govern dispersion and accumulation dynamics [2], [6].

B. Correlation structure and cross-pollutant coupling: To examine inter-variable dependencies, we computed pairwise correlations across numeric variables and visualize them in **Figure 3.4**.

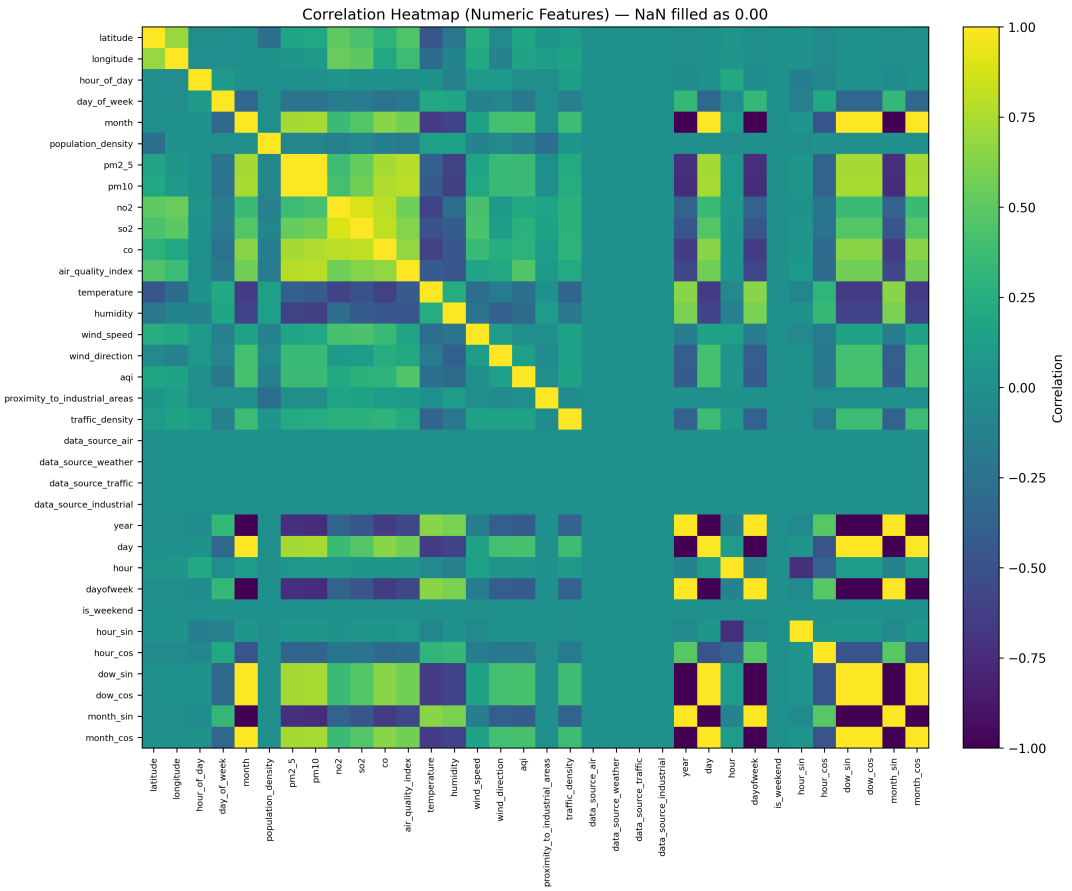

Figure 3.4: Correlation Heatmap

The heat map clearly depicts actual correlations between different pollutants, such as the strong correlation between $PM_{2.5}$ and PM_{10}, as well as different gases and particulate matter. This is logical because of common sources of pollution in cities and temporal factors. Weather factors are also related in meaningful ways to different levels of pollutants, in line with what is known about temperature, humidity, and wind in building up and dispersing concentrations of pollutants [2], [6]. Most importantly, these correlations provide direct empirical evidence that supports using joint models of air quality forecasting instead of developing separate models of different pollutants. This is because joint learning can take advantage of these correlations while still producing results that are applicable to different pollutants [3], [8].

C. Implications for modeling and evaluation: The evidence from **Figure 3.3** to **Figure 3.4** provides direct implications that can be practically useful. There are three different implications from these results. First, the heavy tails of these distributions suggest that nonlinear models should be used in developing models of different pollutants. Second, the correlations between different pollutants suggest that multi-target models should be used to take advantage of these correlations in building models of different pollutants while maintaining a unified sense of air quality. Finally, these correlations between different feature groups suggest that temporal evaluation is critical because these correlations may be artificially increased by

leakage and mixing of distributions. Therefore, in all of the following experiments, leakage-safe rules are used as shown in Section 3.3 [1], [25].

Overall, EDA confirms that the dataset reflects realistic urban air-quality behavior— episodic extremes, structured meteorological influence, and coupled pollutant dynamics— thereby justifying the methodological choices adopted in later chapters regarding multi-output forecasting, non-linear modeling, and rigorous temporal evaluation [2], [25].

3.5 Problem Formulation

We formulate city-level air quality forecasting as a **multi-output supervised time-series regression** task. For each city, observations are indexed by discrete acquisition times and ordered chronologically. At each time step (t), we define a heterogeneous feature vector $\mathbf{x}_t \in \mathbb{R}^F$ that aggregates pollutant-related context and auxiliary predictors (meteorology, temporal encodings, and urban proxy indicators) described in Section 3.2. The objective is to learn a forecasting function that maps a recent historical window of observations to near-term pollutant concentrations under leakage-safe temporal splitting (Section 3.3) [1], [2], [25].

Input window and horizon: Let (W) denote the lookback window length and (H) the forecast horizon. In this work, we adopt a short-horizon setting with $(W = 24)$ and $(H = 1)$, corresponding to predicting the next-step air quality using the most recent 24 time steps. This design is consistent with operational forecasting needs where near-term estimates support timely public-health alerts and rapid-response decision making [2], [25]. For each city and time index (t), the model input is the sequence:

$$\mathbf{X}_t = \{\mathbf{x}_{t-W+1}, \mathbf{x}_{t-W+2}, \ldots, \mathbf{x}_t\} \in \mathbb{R}^{W \times F}.$$

Multi-output targets: The prediction target is a vector $\mathbf{y}_{t+H} \in \mathbb{R}^K$ containing $(K = 5)$ pollutant concentrations at the forecast time $(t + H)$:

$$\mathbf{y}_{t+H} = \left[y^{(1)}_{t+H}, y^{(2)}_{t+H}, \ldots, y^{(K)}_{t+H}\right],$$

where the five targets correspond to $PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO. The learning objective is to estimate a function $f(\cdot)$ such that:

$$f : \mathbf{X}_t \mapsto \mathbf{y}_{t+H},$$

capturing (i) temporal dependence within the lookback window and (ii) cross-pollutant coupling in the output space.

Rationale for joint forecasting: Urban pollutants are not independent; they frequently share emission drivers, meteorological modulation, and temporal structure. The correlation patterns observed in Section 3.4 provide empirical motivation for **joint learning**, where a single model produces a consistent multi-pollutant forecast rather than training separate predictors per pollutant. This multi-output formulation supports shared representation learning while preserving pollutant-specific behavior, a strategy widely adopted in recent multi-pollutant forecasting literature due to its ability to exploit common structure and improve overall utility [3], [8].

Chapter 4

Methodology

4.1 End-to-End Pipeline Overview

This group thesis implements an end-to-end, leakage-safe forecasting pipeline that spans data acquisition, preprocessing, model training, and post-hoc analysis in a single reproducible workflow **Figure 4.1**.

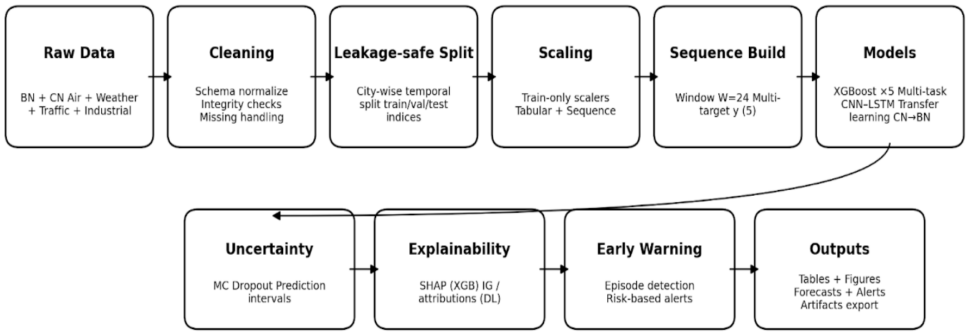

Figure 4.1: End-to-End Thesis Pipeline.

The pipeline is designed for short-horizon, city-level, multi-pollutant forecasting, where methodological rigor (especially time ordering) is essential for trustworthy evaluation in air quality prediction studies [1].

The workflow starts by collecting and aligning multi-source signals per city and timestamp, including pollutant concentrations and supporting covariates (e.g., weather and urban indicators). After ingestion, we apply deterministic validation and cleaning steps—such as **timestamp consistency checks, missing-value handling within city boundaries, and basic integrity rules**—so that downstream models receive stable and comparable inputs across cities and countries.

To prevent temporal leakage, the dataset is then partitioned using chronological, city-wise splitting, and all transformations that can "learn" statistics (e.g., standardization) are performed after the split using training-only parameters. **Two parallel representations are produced** from the same split policy: (i) a tabular matrix for classical baselines and (ii) fixed-length temporal sequences for deep learning.

This separation allows baseline models to remain **strong and interpretable** while enabling sequence models to learn temporal dynamics more directly [1].

Modeling proceeds in structured layers: As minimum references, we include a persistence-style baseline to establish a strict lower bound, then train **XGBoost** models on the engineered tabular representation as a competitive tree-boosting benchmark [15]. The primary contribution is a **multi-task CNN–LSTM** architecture that consumes **multivariate temporal windows and jointly** predicts multiple pollutants via a shared backbone with task-specific output heads, which supports cross-pollutant representation sharing and improves efficiency compared to training fully independent models [3], [5]. **CNN–LSTM hybrids** are widely used for capturing local temporal patterns (via convolution) and longer dependencies (via recurrence) in air quality prediction settings [5].

On top of the trained predictors, we attach three analysis modules to strengthen practical reliability and scientific interpretability. **First**, we evaluate **cross-domain transfer learning strategies** to study how knowledge learned from one region can be adapted **under data-scarce** settings in another [7], [16], [19]. **Second**, we estimate pre- dictive uncertainty using **MC Dropout, treating dropout** as an approximate **Bayesian inference** mechanism to quantify epistemic uncertainty during inference [13]. **Third**, we perform explainability analysis: tree-based models are interpreted using **SHAP- style feature attribution** reported in the air-quality explainability literature, and deep models are interpreted using gradient-based attribution to identify influen- tial temporal segments and covariates [14]. Finally, we translate forecasts into an **early-warning episode detection** step to demonstrate how model outputs can support operational decision-making beyond aggregate error metrics.

The evaluation stage consolidates results using **standardized regression metrics** across validation and test partitions, with reporting stratified by pollutant and city to en- sure that improvements are not driven by a small subset of locations. By keeping each stage explicit **(split → transform → train → analyze)** and modular, the method- ology supports transparent experimentation, defensible comparisons against strong baselines, and consistent extension to additional models or horizons **without break- ing leakage safety** [1].

4.2 Data Splitting Strategy

A defensible split design is critical in air-quality forecasting because evaluation must reflect how a model would be used in practice: learning from past observations and predicting unseen future conditions. In this group thesis, we adopt a **city-wise, chronological splitting policy** to preserve temporal ordering and to prevent information leakage that can otherwise inflate reported performance in environmental prediction studies [1], [2]. The complete split logic is summarized in Figure 4.2.

City	Train	Validation	Test
Barisal	93	21	23
Chittagong	93	21	23
Comilla	91	21	23
Dhaka	93	22	23
Khulna	93	21	23
Mymensingh	92	21	23
Narayanganj	90	21	22
Rajshahi	93	22	23
Rangpur	93	21	23
Sylhet	94	22	23
Beijing	90	21	22
Chengdu	89	21	22
Guangzhou	90	21	23
Hangzhou	89	21	22
Nanjing	89	21	22
Shanghai	90	21	23
Shenzhen	90	21	22
Tianjin	89	21	22
Wuhan	89	21	22
Xi'an	88	21	22

Figure 4.2: Leakage-Safe Split Strategy.

City-wise chronological partitioning: For each city, all records are first sorted by timestamp and then segmented into non-overlapping training, validation, and test intervals along the time axis. **No random shuffling** is applied at any stage. This en- sures that the validation and test sets represent future time periods relative to the training data, matching realistic operational forecasting. To further enforce leakage safety, sequence samples **(used by deep learning models)** are generated so that **input windows and their corresponding forecast targets** remain entirely within the same split segment, i.e., no constructed sample is allowed to cross the train/validation/test boundary.

Country-level separation for cross-region analysis: Beyond the within-city tempo- ral split, the dataset is organized by country by **separating cities from Bangladesh (BN) and China (CN)**. This design supports two complementary evaluation modes: **(i) within-region training and testing**, where models are

trained and evaluated on cities from the same country under leakage-safe temporal splits; and **(ii) cross-region generalization**, where models trained on one country can be evaluated or adapted to the other to study how regional differences affect forecasting behavior. This setup directly supports later **transfer-learning experiments, where a data-rich source do- main** can improve prediction in a data-constrained target domain [7], [19].

Low-data settings for robustness and transfer learning: To evaluate how robust the model is in **low-data settings, we design specific low-data training settings for selected BN cities**. In these settings, we retain only **a smaller portion of the available training data**, while the validation and test periods remain the same. It is important to retain the validation and test periods the same, as this ensures that any **observed change in performance** is due to the reduced training data (and transfer methods) and not due to a shift in the evaluation curve. This method provides **a controlled way to measure robustness and the effectiveness of transfer learning** when the target city histories are small or incomplete [7].

By connecting **(i) per-city temporal splitting, (ii) BN-CN splitting, and (iii) low-data training settings**, the method enables leakage-free forecasting evaluation, facilitates cross-region analysis, and offers a practical testbed for evaluating the **effectiveness of transfer learning in multi-city air quality forecasting tasks** [1], [2].

4.3 Feature Scaling and Representation

To ensure that everything remains **stable, comparable across models, and leakage-free** during evaluation, this group thesis standardizes only on the training data for all continuous data. In other words, the mean and standard deviation used for standardization are calculated purely on the training data and then used on the validation and test data without any further tuning. The integrity of temporal evaluation will thus be preserved, **avoiding leakage** into the pre-processing that can artificially boost performance in a time series forecasting context [1], [2].

Train-only standardization: All continuous variables—including pollutant measurements and supporting covariates such as meteorological and **urban indicators—are converted to standardized z-scores** using training-set statistics. Any feature trans- formation that requires learning parameters (e.g., centering and variance normal- ization) is treated as part of the training pipeline and is never re-estimated on validation or test data. This design keeps model **comparisons fair**: performance dif- ferences reflect modeling capability rather than differences in data scaling or hidden information leakage [1].

Encoding of temporal and categorical attributes: Discrete temporal attributes (e.g., **hour-of-day, day-of-week, month/seasonv**) are encoded into numeric form prior to scaling so that both classical and deep models receive consistent representations. For cyclic time attributes, **a cyclic-safe encoding** (e.g., sine/cosine mapping) is used to avoid **artificial discontinuities** such as the jump from 23→0 hours. Encoded values are then standardized under the same train-only

policy to maintain scale compati-bility across features [2]

Representation for classical baselines (tabular): For tree-based baselines, each in- stance is represented as a single time step with an engineered feature vector that may include lagged and derived attributes. This tabular format is efficient for **gradient-boosted decision trees** and supports competitive baselines while keeping the modeling assumptions explicit. Temporal context is preserved through feature construction (e.g., lag terms) rather than through sequence modeling, enabling direct comparison between feature-based learning and representation learning approaches [15].

Representation for deep learning (sequence tensors): For deep models, the same leakage-safe splits are transformed into fixed-length temporal windows. Inputs are arranged as a 3D tensor

$$\mathbf{X} \in \mathbb{R}^{N \times W \times F},$$

where (N) is the number of samples, (W) is the window length, and (F) is the number of features per time step. Each sample window is paired with a forecast target at the chosen horizon, and window construction is constrained so that both the input history and prediction target remain strictly within the same split segment (train/validation/test), preserving chronological isolation [1], [2]. This representation enables the **CNN–LSTM** backbone to learn short-range local patterns and longer-range temporal dependencies directly from the raw multivariate trajectories, without relying on handcrafted lag features [5].

By combining **train-only standardization with model-specific representations** (tabu- lar for classical baselines and sequences for deep learning), the preprocessing pipeline supports fair model comparison while maintaining strict leakage safety in a multi- city, time-ordered forecasting setting [1], [2].

4.4 Baseline Models

Baseline models are included to **provide clear reference points** for judging the effec- tiveness of the proposed forecasting framework. In time-series air-quality prediction, strong baselines are necessary to separate improvements that come from sophisti- cated architectures (e.g., multi-task learning, transfer learning, and uncertainty-aware inference) from gains that could be achieved through simpler assumptions or classical tabular modeling [1], [2]. Accordingly, this group thesis evaluates two baseline families: **a persistence (naïve) predictor and per-pollutant XGBoost regressors**.

4.4.1 Persistence Baseline

The persistence baseline establishes a strict lower-bound benchmark for short-horizon forecasting. For each pollutant, the next-step prediction is set equal to the most recent observed value:

$$\hat{y}_{t+1} = y_t.$$

This baseline requires no training and no feature learning. It is nonetheless meaningful because air-pollutant concentrations often exhibit short-term inertia, especially under stable meteorological conditions. Any model intended for practical forecasting should reliably outperform persistence across cities and pollutants; otherwise, the model does not add predictive value beyond temporal continuity [1], [2]. In this study, **persistence is reported for each pollutant and split**, and it serves as the minimum reference when interpreting downstream model gains.

4.4.2 XGBoost Models Per Pollutant

To provide a competitive classical benchmark on the tabular representation, we train gradient-boosted decision-tree regressors using **XGBoost** [15]. XGBoost is widely adopted for structured data because it can model non-linear interactions, handle heterogeneous feature scales after standardization, and **provides strong performance with well-established regularization controls** [15].

We train **separate single-output models**—one per target pollutant—so that the base- line mirrors the conventional "one model per variable" setup common in prior air- quality regression pipelines. Each XGBoost model receives the standardized tabular features described in Section 4.3, including temporal encodings, meteorological co- variates, and urban indicators. Where **lagged terms are used, they are included as explicit engineered predictors** within the same feature vector, enabling direct com- parison against the deep sequence models that learn temporal dependencies from windows rather than handcrafted lags [1], [5].

To reduce overfitting and ensure comparable training conditions across pollutants, we apply regularized boosting with controlled tree depth, learning rate, subsampling, and column sampling, and we select **hyperparameters** using the validation split under the same leakage-safe split policy described in Section 4.2. The **finalized hyperparameter** configuration is reported in **Table 4.1** to support transparency and reproducibility [15].

city	country	split	global_start	global_end	n	s_time	e_time	json_count
Chittagong	BN	train	475	591	117	2025-12-24 06:28:51	2026-01-23 11:33:54	117
Chittagong	BN	val	592	612	21	2026-01-23 11:47:12	2026-01-23 18:02:30	21
Chittagong	BN	test	613	635	23	2026-01-23 18:04:36	2026-01-23 23:47:00	23
Dhaka	BN	train	795	911	117	2025-12-24 06:28:45	2026-01-23 11:22:00	117
Dhaka	BN	val	912	933	22	2026-01-23 11:41:42	2026-01-23 17:55:00	22
Dhaka	BN	test	934	956	23	2026-01-23 17:59:24	2026-01-23 23:39:54	23
Beijing	CN	train	161	274	114	2025-12-24 06:29:53	2026-01-23 11:09:24	114
Beijing	CN	val	275	295	21	2026-01-23 11:31:06	2026-01-23 16:30:24	21
Beijing	CN	test	296	317	22	2026-01-23 16:43:30	2026-01-23 23:39:12	22
Shanghai	CN	train	2228	2341	114	2025-12-24 06:29:59	2026-01-23 10:52:30	114
Shanghai	CN	val	2342	2362	21	2026-01-23 11:17:12	2026-01-23 16:29:12	21
Shanghai	CN	test	2363	2385	23	2026-01-23 16:36:06	2026-01-23 23:44:36	23

Table 4.1: Model Hyperparameters (XGBoost)

Together, the **persistence baseline and per-pollutant XGBoost** models form a conser- vative and interpretable reference set. They allow us to attribute performance im- provements in later sections specifically to joint **multi-task modeling, cross-country transfer mechanisms, and uncertainty-aware forecasting** components rather than to preprocessing artifacts or weak comparison choices [1], [2].

4.5 Proposed Multi-Task CNN–LSTM Model

To forecast multiple correlated pollutants under heterogeneous urban conditions, this group thesis **proposes a multi-task CNN–LSTM architecture** that combines **convolutional feature extraction with recurrent temporal modeling** under a shared- representation design. The model is configured for short-horizon prediction us- ing fixed-length historical windows and produces simultaneous outputs for multiple pollutant targets through lightweight task-specific heads **Figure 4.3**.

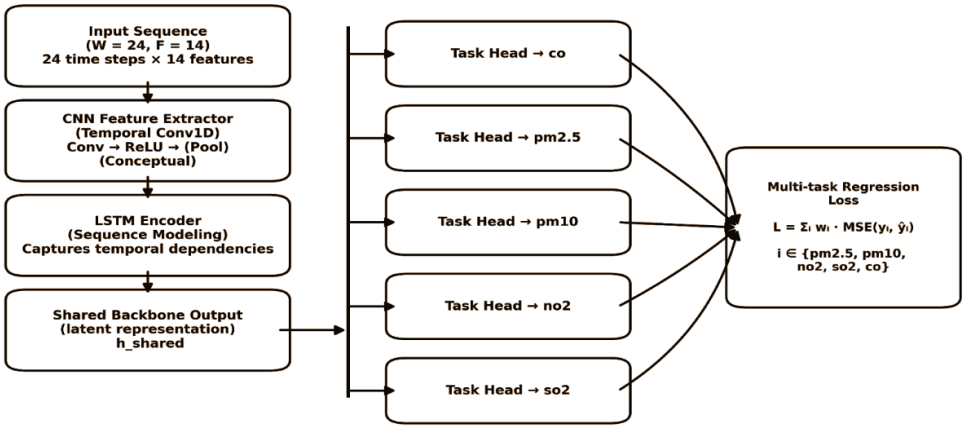

Figure 4.3: Multi-Task CNN–LSTM Architecture.

The design choice is motivated by prior evidence that **hybrid CNN–RNN models are effective for air-quality time series, and that multi-task learning improves data efficiency**, when targets share common generative factors (e.g., emissions and mete- orology) [1], [2].

4.5.1 Architectural Motivation

Urban air-pollution signals typically contain **(i) localized temporal variations**, such as abrupt changes driven by traffic intensity or short-term meteorological shifts, and **(ii) longer-range dependencies**, including delayed atmospheric effects and recurring diurnal/seasonal patterns. One-dimensional temporal convolutions are well suited for extracting local patterns efficiently by applying shared filters across the input window, acting as a learnable alternative to handcrafted smoothing or lag-feature selection [5]. However, **convolution alone does not explicitly maintain a state over time**.

To capture longer dependencies, **the backbone includes an LSTM layer that models sequential structure** through gated memory updates, enabling learning of delayed and accumulated effects that frequently appear in pollution dynamics [29]. **Combin- ing CNN and LSTM supports hierarchical temporal learning:** convolutional blocks first encode local changes and denoise the signal, while the recurrent layer integrates these features into a compact representation of recent history.

Finally, pollutants such as **PM2.5, PM 10, NO2, SO2, and CO** often co-vary because they are **influenced by shared emission sources and common at-**

mospheric processes. Instead of training independent models per pollutant, we adopt **multi-task learn- ing with a shared backbone and multiple output heads**. This structure encourages the model to learn representations that generalize across targets while retaining task- specific mappings at the output stage, consistent with standard multi-task learning principles [3].

4.5.2 Network Design

Let the input sequence for a single sample be denoted by

$$\mathbf{X} \in \mathbb{R}^{W \times F},$$

where W represents the historical window length and F denotes the number of standardized features at each time step (see Section 4.3). For a batch consisting of N samples, the input data is represented as a three-dimensional tensor

$$\mathbf{X} \in \mathbb{R}^{N \times W \times F}.$$

Shared convolutional encoder: The model begins with a stack of **1D convolutional layers** that operate along the temporal axis. These layers learn short-range temporal filters that capture rapid fluctuations and short-lived effects while improving robustness to noise. Nonlinear activations are applied after each convolution **to increase representational capacity, and dropout-based regularization** is applied within the shared backbone to improve generalization.

Shared recurrent temporal model: The convolutional feature maps are then passed to a **shared LSTM layer** that summarizes temporal dynamics over the full window. The **LSTM output serves as a compact latent representation encoding both local and longer-range dependencies learned from the recent history** [29].

Multi-task output heads: From the shared latent representation, the network branches into (K) pollutant-specific regression heads **Figure 4.3**. Each head is implemented as a lightweight fully connected mapping that outputs a scalar forecast for a single pollutant. **This design keeps the bulk of learning in the shared backbone** while allowing each pollutant to maintain a dedicated output transformation, aligning with **standard multi-task architectures** in which shared layers learn common factors and heads capture target-specific sensitivities [3].

4.5.3 Training Procedure

The multi-task model is trained end-to-end using a joint regression objective. For K pollutants, the overall loss is defined as the aggregation of per-task losses:

$$\mathcal{L} = \sum_{k=1}^{K} \lambda_k \, \mathcal{L}_k,$$

where $\mathcal{L}_k$ denotes the regression loss (mean squared error) for pollutant k, and λ_k is the corresponding task weight. In the default setting, **equal weights are assigned**

to all tasks to prevent any single pollutant from dominating the optimization process, while maintaining a unified learning signal across tasks.
Optimization is performed with an adaptive gradient-based optimizer (e.g., Adam), and training proceeds on the leakage-safe training split with selection based on validation performance (Section 4.2). **To mitigate overfitting, dropout is applied in the shared backbone, and early stopping** is used with a patience criterion on validation loss so that **the final checkpoint corresponds to the best generalizing model rather than the last epoch**. The resulting training and validation loss trajectories are reported in **Figure 4.4** to document convergence behavior and to support reproducibility.

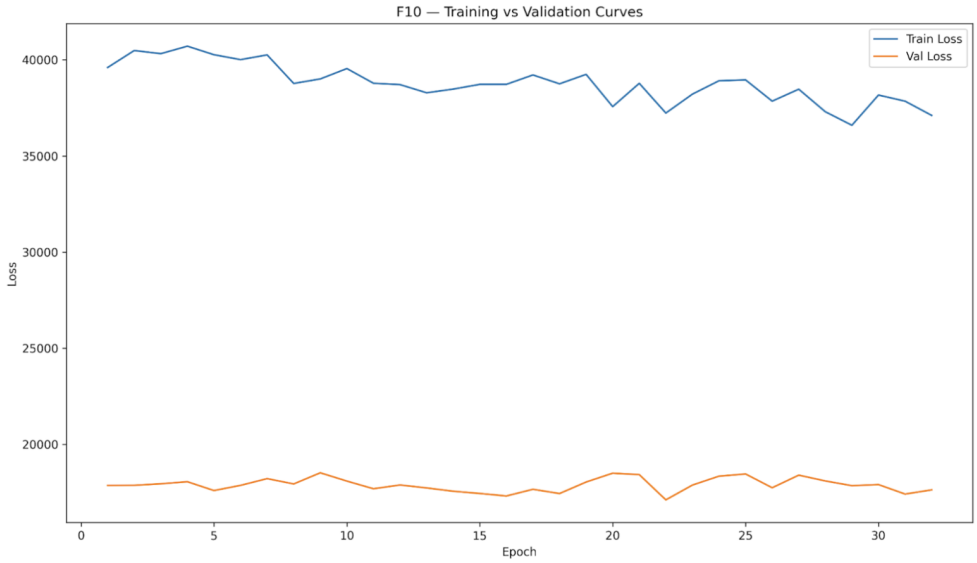

Figure 4.4: Training vs Validation Curves.

This training configuration yields a single multi-output predictor that can be evaluated consistently across cities and pollutants, and it also provides a clean foundation for the transfer-learning, uncertainty quantification, and explainability modules described in subsequent sections [3], [5], [29].

4.6 Transfer Learning Strategy

Air-quality forecasting deployments often face uneven data availability across regions and cities, where some locations have long monitoring histories while others have sparse or interrupted records. To study **cross-region generalization and to improve performance under limited local data**, Our group thesis adopts a **transfer learning protocol** in which a model trained in a data-richer source domain is adapted to a data-scarce target domain. **The approach aligns with standard transfer-learning practice in deep neural networks, where pre-trained representations are reused to reduce sample complexity and improve robustness during downstream adaptation** [7], [19].

4.6.1 Source-Domain Pretraining (CN)

We treated the China (CN) cities as the source domain because they provide comparatively larger and more complete training sequences in our dataset. **The proposed multi-task CNN–LSTM (Section 4.5) is first trained on the CN training** split us- ing the same leakage-safe city-wise temporal segmentation described in Section 4.2. During this stage, **the shared CNN–LSTM backbone learns broadly useful temporal abstractions** that capture common pollution dynamics and covariate interactions (e.g., the response of pollutant levels to meteorological conditions and recurring temporal cycles). **In transfer-learning terms, this stage produces a representation that is expected to be reusable across related urban environments** [7], [19].

4.6.2 Target-Domain Adaptation (BN Fine-Tuning)

After source training converges, the **pretrained parameters are used to initialize the target-domain model for Bangladesh (BN) cities**. The model is then **fine-tuned on BN training data**, allowing the network to adjust to region-specific character- istics such as different emissions profiles, local meteorological regimes, and mea- surement distributions. **Fine-tuning is performed under the same preprocessing and scaling policy established earlier:** BN data are standardized using train-only statis- tics within the BN split, and adaptation never uses validation or test information (Sections 4.2–4.3). **This ensures that any observed improvement reflects genuine transfer rather than data leakage**. The fine-tuning protocol follows common **deep transfer-learning** practice where pretrained features accelerate convergence and im- prove generalization when target labels are limited [7], [19].

4.6.3 Low-Data Adaptation and Backbone Freezing

To explicitly evaluate transfer under realistic scarcity, we construct low-data BN scenarios for selected BN cities (Section 4.2), where only a reduced fraction of BN training observations is available for adaptation while validation/test partitions remain fixed. We study two **adaptation settings:**

1. **Full fine-tuning**, where all network parameters are updated on BN; and
2. **Partial fine-tuning with freezing,**, where early layers of the shared backbone are frozen and only later layers and/or output heads are updated.

Freezing tests whether the pretrained backbone already captures generalizable temporal features and whether limited BN data is better used to learn only region-specific calibration at the upper layers. **This is a standard strategy to control over- fitting and stabilize adaptation** when the target dataset is small [7], [19].

4.6.4 Evaluation Protocol for Transfer Benefit

Transfer effectiveness is assessed by comparing: **(i) BN models trained from scratch and (ii) BN models initialized from CN pre-training** and then

adapted under the same training budget. Performance is reported using identical metrics and splits, and results are summarized in **Table 4.1** with relative gains visualized in **Figure 4.4**. **These comparisons quantify the practical value of cross-country knowledge** trans- fer for multi-pollutant forecasting, particularly under BN low-data settings, **where representation reuse is expected to provide the largest benefit** [7], [19].

4.7 Uncertainty Quantification

Point forecasts alone are often insufficient for air-quality decision support, where actions may depend not only on the predicted pollutant level but also on how reli- able that prediction is. **To support risk-aware interpretation of model outputs**, this group thesis augments deterministic forecasting with predictive uncertainty esti- mation using **Monte Carlo (MC) Dropout**. MC Dropout enables approximate **Bayesian inference in deep networks** by treating dropout as a variational mechanism and sampling predictions through stochastic forward passes at inference time [13].

4.7.1 MC Dropout Inference Procedure

Dropout is typically used only during training to reduce overfitting by randomly deactivating a subset of units. In Monte Carlo (MC) Dropout, the same stochastic mechanism is intentionally retained during inference. Given an input sequence X, we perform M stochastic forward passes through the **trained CNN–LSTM model with dropout enabled, producing a set of predictions for each pollutant:**

$$\left\{\hat{y}^{(1)}, \hat{y}^{(2)}, \ldots, \hat{y}^{(M)}\right\}.$$

The predictive mean is computed as

$$\mu(\mathbf{X}) = \frac{1}{M}\sum_{m=1}^{M} \hat{y}^{(m)},$$

and is reported as the final point estimate. The predictive uncertainty is quan-tified using the sample variance:

$$\sigma^2(\mathbf{X}) = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{y}^{(m)} - \mu(\mathbf{X})\right)^2.$$

This variance reflects **epistemic uncertainty arising from model parameter uncer-tainty under the learned representation**, which is particularly relevant when the model is applied to cities, regimes, or episodes that are under-represented in the training [13].

4.7.2 Prediction Bands and Reliability Analysis

To make uncertainty interpretable in the context of a forecasting scenario, **we construct prediction intervals around the forecast** using the measure of dispersion. **Figure 4.5** illustrates the change in these intervals over time, emphasizing the

variation in uncertainty based on varying levels of pollution and weather conditions. In effect, **the widening of intervals indicates when the model is less certain, advising the need for conservative decision policies** [13].

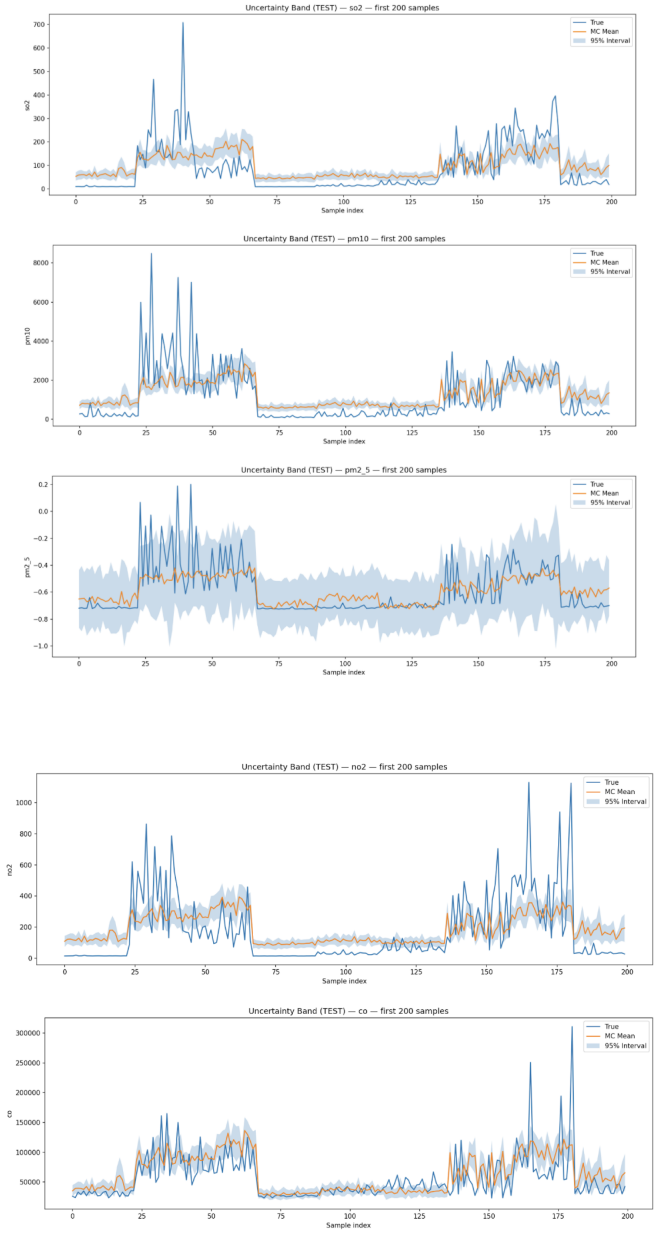

Figure 4.5: MC Dropout Prediction Bands

Beyond visualization, we evaluate whether uncertainty estimates are informative by analyzing the relationship between absolute prediction error and the corre-sponding uncertainty magnitude. If the model's uncertainty is meaningful, **higher uncertainty should coincide with larger forecast errors** more frequently than would occur under an uninformative uncertainty signal. This relationship is summarized in **Figure 4.5** and is used as a **diagnostic of uncertainty calibration quality in our setting**.

By integrating MC Dropout into the pipeline, the proposed framework extends beyond point prediction to an uncertainty-aware forecasting approach. **This supports**

more cautious interpretation of forecasts and provides an additional reliability signal that complements conventional accuracy metrics in air-quality prediction tasks [13].

4.8 Explainability Framework

Accurate forecasts alone are not sufficient in high-impact environmental applications; stakeholders often require credible explanations to understand which factors are driving predicted pollutant changes. To improve transparency and to support scien- tifically meaningful interpretation, this group thesis **integrates model-appropriate explainability for both the classical baseline family and the proposed deep learn- ng model**. Because tree ensembles and neural sequence models differ fundamentally in structure, we employ **two complementary techniques: SHAP for XGBoost and Integrated Gradients for the CNN–LSTM**. These methods provide feature attri- bution without altering model parameters or training behavior [14], [33].

4.8.1 SHAP Analysis for XGBoost

For the per-pollutant XGBoost baselines (Section 4.4), we use **SHAP—Shapley Additive exPlanations**—to assess the individual contribution of each feature to the predictions made by the model. **SHAP** is based on cooperative game theory and assigns a value to each input feature that **represents its contribution to the prediction, compared to the expected value of the feature** [33]. The large advantage of **SHAP** for tree ensembles is that there are efficient algorithms to compute it, **allowing us to compute both local explanations (instance-level) and global feature importance explanations for the entire dataset** [33].

In our study, we compute **SHAP** values for each individual pollutant-specific XGBoost model to analyze the influence of temporal representations, meteorological variables, and urban features on the predictions. We provide:

- **Global explanations:** In the form of **SHAP** summary plots to identify features that contribute to predictions across the entire test set; and
- **Pollutant-specific comparisons:** To demonstrate how the most important features differ for each target (particulate matter vs. gaseous pollutants).

Examples of global **SHAP** plots are provided in**Figure 4.6**, allowing us to analyze the patterns of learned feature importance without focusing on aggregate error rates [33].

4.8.2 Integrated Gradients for Deep Sequence Models

Forecasting, where recent and past data have a different weight in the forecast [14]. We calculate the **Integrated Gradients (IG)** attributions for selected sequences

and target pollutants to investigate:

We compute **IG** attributions for selected sequences and target pollutants to analyze:

- **Temporal relevance:** Which regions of the input window are most responsible for each individual forecast.
- **Relevance of features:** Which covariates and pollutant histories are most relevant for particular forecasts in the shared multi-task backbone.

The attribution maps are visualized in **Figure 4.6** to support a qualitative understanding of the deep model's behavior and to supplement the **SHAP-based explanations provided for XGBoost**. Together, SHAP and Integrated Gradients **provide a unified interpretability layer for various modeling strategies**, improving interpretability without compromising the forecasting pipeline [14], [33].

4.9 Early-Warning Episode Detection

While metrics for accuracy, such as **RMSE or MAE**, are certainly relevant, the practical application of air quality monitoring systems depends upon a rather different criterion: **can the system identify high levels of pollution in time for action to be taken?** To connect the issue of forecasting with practical application, the thesis developed by this group adds a level of **early warning detection atop the short-horizon predictions**, transforming continuous forecasts into risk-based alerts and event summaries [25], [35]. The approach follows the typical framework for determining events based upon the exceedance of thresholds, as is typical for early warning systems [25], [35].

4.9.1 Episode Definition and Threshold Selection

A pollution episode is a period during which a target pollutant exceeds a higher risk threshold. These thresholds are unique to each pollutant and are computed by combining two methods: (i) commonly used thresholds, and (ii) percentile-based thresholds. This is done to ensure that we are being fair to all cities when making comparisons, even though their levels of pollution may be different to begin with. This is an important consideration because a single threshold may trigger too many alarms in polluted cities and too few alarms in clean cities, whereas percentile-based thresholds are more sensitive to local conditions and still indicate abnormal conditions.

4.9.2 Forecast-to-Alert Logic and Event Construction

The forecast value $\hat{y} = t + H$ is compared with the defined threshold for the particular time step. If the forecast crosses the threshold, the particular time step is considered to be a point of risk. The neighboring risk points are grouped together **to provide the start time, end time, and the length of the episode**, which is referred to as the episode window. This is where the grouping of the risk points becomes significant, because crossing the threshold by itself may be of no real consequence, but crossing the threshold consecutively is more significant and aligns with

the goals of public health. **Figure 4.6** demonstrates the definition of the episode and grouping of the risk points [25], [35].

Figure 4.6 illustrates the episode definition and grouping mechanism in a time-series view, showing how forecast-based exceedances translate into coherent alert windows rather than scattered pointwise flags.

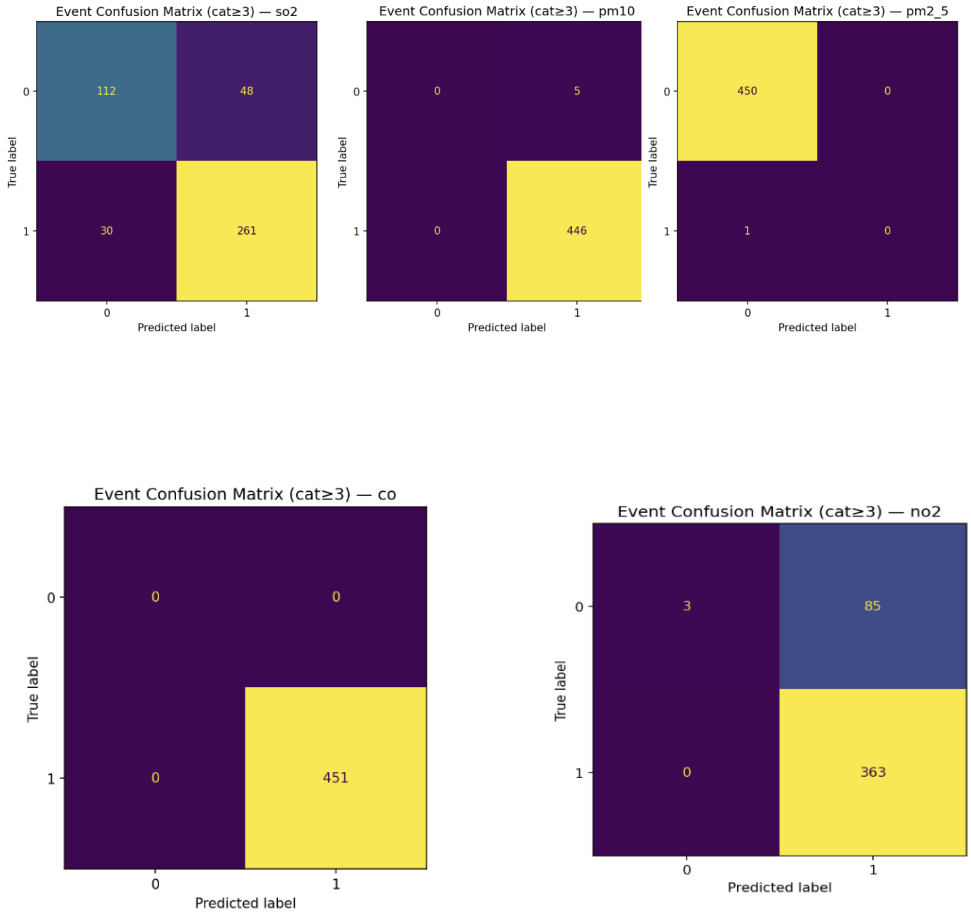

Figure 4.6: High-Pollution Episod Detection

4.9.3 Event-Level Evaluation and Practical Diagnostics

To evaluate early-warning behavior, **we extend beyond pointwise regression metrics and report event-oriented criteria** that assess whether predicted alerts align with ob- served episodes. Following common recommendations for event-focused evaluation, we include measures of detection quality (e.g., precision/recall tendencies) and tim- ing behavior (lead-time characteristics), since a model can achieve strong average error yet still miss extremes or misplace episode onsets [2], [33].

Table 4.2 summarizes episode-detection statistics (including event-level detection mea- sures and lead-time summaries) across pollutants and city groupings, while

Figures F17–F18 provide representative timelines and case studies that illustrate alignment between forecast-based alerts and observed concentration trajectories under real episode dynamics.

Group	Target	n_{points}	Freq. (min)	Cat. Acc.	Macro F1	Event Acc.
overall	$PM_{2.5}$	451	1	0.721	0.295	0.998
overall	PM_{10}	451	1	0.942	0.194	0.989
overall	NO_2	451	1	0.743	0.290	0.812
overall	SO_2	451	1	0.643	0.347	0.827
overall	CO	451	1	1.000	0.200	1.000
city::Beijing	$PM_{2.5}$	22	17	0.409	0.116	1.000
city::Beijing	PM_{10}	22	17	1.000	0.200	1.000
city::Beijing	NO_2	22	17	1.000	0.200	1.000
city::Beijing	SO_2	22	17	1.000	0.200	1.000
city::Beijing	CO	22	17	1.000	0.200	1.000
city::Chittagong	$PM_{2.5}$	23	13	1.000	0.200	1.000
city::Chittagong	PM_{10}	23	13	0.261	0.083	0.783
city::Chittagong	NO_2	23	13	0.000	0.000	0.000
city::Chittagong	SO_2	23	13	1.000	0.200	1.000
city::Chittagong	CO	23	13	1.000	0.200	1.000

Group	evt_precision	evt_recall	evt_f1	n_true_ep	n_pred_ep	ep_recall
overall	0	0	0	1	0	0
overall	0.9889	1	0.9944	6	1	1
overall	0.8103	1	0.8952	74	4	1
overall	0.8447	0.8969	0.8700	103	99	0.9612
overall	1	1	1	1	1	1
city::Beijing	0	0	0	0	0	0
city::Beijing	1	1	1	1	1	1
city::Beijing	1	1	1	1	1	1
city::Beijing	1	1	1	1	1	1
city::Beijing	1	1	1	1	1	1
city::Chittagong	0	0	0	0	0	0
city::Chittagong	0.7826	1	0.8780	6	1	1
city::Chittagong	0	0	0	0	1	0
city::Chittagong	0	0	0	0	0	0
city::Chittagong	1	1	1	1	1	1

Group	evt_f1	n_true	n_pred	ep_recall	false_alarm	lead_mean
overall	0	1	0	0	0	0
overall	0.994	6	1	1	0	235
overall	0.895	74	4	1	0	91
overall	0.87	103	99	0.961	11	0.77
overall	1	1	1	1	0	0
Beijing	0	0	0	0	0	0
Beijing	1	1	1	1	0	0
Beijing	1	1	1	1	0	0
Beijing	1	1	1	1	0	0
Beijing	1	1	1	1	0	0
Chittagong	0	0	0	0	0	0
Chittagong	0.878	6	1	1	0	174
Chittagong	0	0	1	0	1	0
Chittagong	0	0	0	0	0	0
Chittagong	1	1	1	1	0	0

Group	Lead Median (min)	Lead P10 (min)	Lead P90 (min)
overall	259.2	52.1	392.85
overall	78.5	8.76	197.55
overall	0	-0.12	3.1
overall	0	0	0
Beijing	0	0	0
Beijing	0	0	0
Beijing	0	0	0
Beijing	0	0	0
Chittagong	182.3	19.6	321.15
Chittagong	0	0	0
Chittagong	0	0	0
Chittagong	0	0	0

Table 4.2: Episode Detection Metrics

By integrating episode detection into the pipeline, **the proposed framework extends model utility from "lower average error" to actionable early-warning intelligence, enabling risk-aware interpretation**of forecasts and supporting proactive responses in urban air-quality management scenarios [25], [35].

Chapter 5

Experimental Setup

5.1 Implementation Environment

All experiments in this group thesis were performed in a controlled and reproducible manner, allowing for the execution of conventional machine learning baselines as well as deep sequence models. The entire process, from data preparation to model training, evaluation, and finally reporting, has been implemented in a Python-based pipeline.

Regarding data handling and **transformations, conventional scientific computing tools have been used**, which are well-suited for structured time series data processing. Classical baselines were developed using widely adopted machine learning utilities for regression mod- eling and metric computation, ensuring that evaluation procedures remained con- sistent across all baselines and proposed models. **Gradient-boosted** decision tree experiments were conducted using **XGBoost** due to its efficiency and scalability for tabular learning and its strong empirical performance in structured regression set- tings [15].

Deep learning experiments were implemented in **TensorFlow/Keras** using modular components for temporal representation learning, supporting convolutional and recurrent sequence blocks under a unified training interface. **Model optimization** fol- lowed standard mini-batch training with **adaptive gradient-based optimization** (e.g., Adam), and **convergence stability was enforced** through callback-based control mech- anisms, including early stopping and checkpointing so that the final selected model corresponds to **the best validation-performing state** rather than the last training epoch [17]. These training controls were applied uniformly to maintain fair experi- mental conditions across pollutants and across **transfer-learning stages**. [17].

In order to reduce randomness and make the results repeatable, we used a fixed random seed in all libraries where deterministic control is possible. **Intermediate artifacts such as split definitions, preprocessing scalers, learned weights, forecast outputs, evaluation summaries, and plots** were tracked with a consistent directory structure to ensure that every result can be traced back to a specific executable run configuration. This allows for repeatable experimentation and comparison across different families of models while maintaining leakage-safe evaluation

discipline as enforced throughout the pipeline [1], [2].

5.2 Evaluation Metrics

To evaluate the accuracy of the forecasting methods, three regression measures were used, which are commonly employed in air quality prediction studies and facilitate a fair comparison between methods, pollutants, and cities. These measures are Root Mean Squared Error **(RMSE)**, Mean Absolute Error **(MAE)**, and the coefficient of determination (R^2).[1], [5], [10], [25] These three measures collectively evaluate different facets of a predictor's performance.

Here, y_i represents the measured pollutant concentration, and $\hat{y}_i$ is the corresponding model prediction for the i sample, where N is the number of samples in the split used for evaluation. The main advantage of the **RMSE** metric is that it penalizes large errors more strongly through the squared error term and is hence informative for high-variance situations, i.e., sudden changes in pollution levels.[1], [5] The equation for the calculation of the RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

The second metric, the MAE, also gives a measure of the average deviation, which is not as adversely affected by extreme errors and hence gives a better idea of the overall forecasting performance of the model.[1], [10] The equation for the calculation of the MAE is as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

R^2 gives a measure of the relative goodness of the model in terms of the proportion of the variance explained, which can be particularly useful for comparing the models for different pollutants, which may have different variability and scale patterns.[6], [38] The equation for the calculation of the R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the measured pollutant concentrations.

Beyond pooled performance, city-wise stability was evaluated to quantify spatial robustness and ensure that reported gains were not driven by a small subset of locations. For each pollutant, **RMSE, MAE, and R2** were computed separately per city, and the resulting distributions were summarized to characterize inter-city variability and generalization consistency—an important requirement in multi-city and cross-region air-quality modeling.[6], [25], [38]

5.3 Experiment Design

The experiment was designed in a way that **guarantees a precise and comparable assessment of the forecasting capabilities** of the models, as well as the training methods and analysis purposes. The basic idea is that for all the experiments, the following steps are carried out: the same steps are applied for the preparation of the data, the same strategy for the train, validation, and test split, the same definitions for the input and output, and the same calculation of the metrics. This way, **the models are compared without the risk of differences in the results being due to differences in the evaluation protocols** [1], [25]. **Table 4.1** shows a summary of the models and the training methods.

5.3.1 Model families and variants

The study evaluates three primary modeling paradigms.

1. **Naïve baseline (Persistence):** For each pollutant k, the forecast is defined as
$$\hat{y}_k(t+H) = y_k(t),$$
providing a strict lower-bound reference that is commonly used to contextualize time-series forecasting gains[1].

2. **Classical machine learning baseline (XGBoost):** Five independent single-output gradient-boosted tree regressors are trained—one per target pollutant—using the engineered tabular feature set. **XGBoost** is selected due to its **scalability and strong performance** on structured regression tasks[15].

3. **Proposed deep learning model (Multi-task CNN–LSTM):** A unified multi-output sequence model jointly predicts all target pollutants using **shared temporal representations with pollutant-specific output heads**. Multi-task learning is included to exploit cross-pollutant dependencies while reducing re- dundant model training across targets, consistent with the direction of recent air-quality **deep learning research** [5], [10].

5.3.2 Input–output specification and evaluation horizon

To ensure **strict comparability** across all model variants, every experiment uses the same temporal input window and forecast horizon, fixed across cities and pollutants. Specifically, **the models ingest the preceding 24 time steps ((W=24)) and predict the next-step concentration at one-step horizon ((H=1)) for each target pollutant**. This fixed specification is maintained across baselines, the proposed architecture, and all ablation/extension studies to preserve fairness in comparison.

5.3.3 Training regimes and generalization scenarios

Experiments were organized to reflect both within-region learning and cross-region generalization challenges. Models were trained and evaluated under (i) region- specific settings using Bangladesh-only and China-only datasets, and (ii) cross-country

transfer configurations, where a model pretrained on the higher- volume source domain is fine-tuned on the target domain to evaluate portability under distribution shift—an increasingly adopted strategy in air-quality model- ing [7], [19]. In addition, data-scarcity stress tests were conducted for selected Bangladesh cities by restricting the available training portion while keeping vali- dation and test partitions unchanged, allowing robustness assessment under low- resource conditions that occur in operational deployments [1], [25].

5.3.4 Controlled comparison protocol

Across all regimes, chronological data splitting was applied consistently, with a fixed train/validation/test strategy, identical preprocessing, and unchanged metric definitions (Section 5.2). Hyperparameter selection and model selection were performed using the validation split, and final performance was reported on held-out test data only. Results were computed per pollutant and per city and then summarized to assess both overall accuracy and city-wise stability, ensuring that reported improvements are not driven by a small subset of locations but represent consistent performance across heterogeneous urban environments [6], [25].

5.4 Reproducibility and Artifacts

For the purposes of transparent verification and reconstruction of the results as published, we have adopted an artifact-centric experiment design, whereby every step of the experiment pipeline generates verifiable output, which is stored persistently. This is in accordance with most guidelines and suggestions within the air quality forecasting community, as discussed in [1], [25].

5.4.1 Persisted models and training states

For each experiment variant, trained model parameters are saved immediately after training. Classical baselines (e.g., gradient-boosted tree regressors) are stored per pollutant to preserve the single-output training setup and facilitate direct reuse in ablation comparisons[15].For deep learning experiments, we retain both (i) the final trained model and (ii) the best-validation checkpoint selected via early-stopping criteria, ensuring that reported performance corresponds to the validation-optimal state rather than an arbitrary terminal epoch[1]. This checkpoint retention is also required for downstream stages such as fine-tuning in cross-country transfer learning, where pretrained weights serve as initialization for target-domain adaptationciteref11,ref17.

5.4.2 Saved pre-processing state and split definitions

To guarantee that test-time evaluation is repeatable and free from retrospective drift, pre-processing state is persisted as explicit artifacts. This includes fitted scalers/normalizers used for feature and target transformations, along with the chronological split definitions (train/ validation/ test indices) applied to each city. Persisting both the transformation state and the split metadata ensures that inde-

pendent reruns reproduce identical inputs to every model and prevents accidental inconsistencies across baselines and deep variants [1].

5.4.3 Metrics, predictions, and structured reports

Evaluation outputs are logged in machine-readable form for every pollutant, city, and split. In addition to aggregated tables, per-city metric records are retained to support stability analysis and cross-city robustness comparisons reported in the results chapter. Where required for downstream analysis (e.g., episode detection and uncertainty calibration), prediction arrays are also stored so that event-level evaluation and post-hoc diagnostics can be reproduced without retraining [1], [25].

5.4.4 Plots and diagnostic figures

All figures included in this thesis are generated from saved artifacts using fixed plotting routines and consistent naming conventions. This includes learning curves, forecast-vs-observation plots, uncertainty bands, and explainability visualizations. Uncertainty artifacts (e.g., Monte Carlo dropout distributions and derived prediction intervals) are saved alongside the corresponding point forecasts to ensure that risk-aware analysis remains reproducible under identical sampling settings[17]. Explainability outputs—such as SHAP-based summaries and feature-attribution visualizations—are stored as static plots and, where applicable, as serialized attribution values to enable inspection and comparison across cities and pollutants[14].

5.4.5 Organization and traceability

All artifacts are organized under a consistent directory structure that separates models, scalers, splits, predictions, reports, and plots. Naming conventions encode the model family, training regime (region-specific vs. transfer), and evaluation split, allowing any reported table entry or figure to be traced back to the exact saved run outputs. This artifact management strategy supports defensible experimentation and makes the full empirical workflow auditable for reviewers and future researchers [1].

Chapter 6

Results

6.1 Overall Forecasting Performance

This section reports the overall **short-horizon forecasting accuracy** of all evaluated models, covering all **target pollutants and all study cities** under the leakage-safe experimental protocol defined earlier. The comparison includes a **persistence baseline** (lower-bound reference), **per-pollutant XGBoost** regressors as strong classical learners for tabular features [29], and the proposed **multi-task CNN–LSTM** framework designed to learn shared temporal representations across pollutants [5], [8]. The consolidated quantitative outcomes are provided in **Tables ??** and **??**, which summarize **RMSE, MAE, and** R^2 on the held-out test split for each target pollutant (and aggregated where appropriate).

Model	**Target**	**Split**	**RMSE**	**MAE**	R^2
XGBoost	CO	Test	175.57	103.62	0.213
XGBoost	CO	Val	159.29	99.47	0.179
XGBoost	NO_2	Test	11.17	6.93	0.481
XGBoost	NO_2	Val	10.27	6.47	0.481
XGBoost	PM_{10}	Test	25.34	16.24	0.409
XGBoost	PM_{10}	Val	25.53	16.29	0.510
XGBoost	$PM_{2.5}$	Test	0.293	0.185	0.388
XGBoost	$PM_{2.5}$	Val	0.284	0.180	0.521
XGBoost	SO_2	Test	8.08	4.26	0.463
XGBoost	SO_2	Val	7.69	3.98	0.479
Total	–	–	**423.51**	**257.64**	**4.12**

Table 6.1: XGBoost model performance across pollutants and splits.

Target	MAE	RMSE	R^2	N	spRMSE (Mean)	spRMSE (Weighted)
$PM_{2.5}$	19.74	25.98	0.168	451	22.09	22.05
PM_{10}	1374.47	2171.46	0.381	451	1694.36	1689.16
NO_2	153.63	245.66	0.440	451	194.41	193.12
SO_2	73.35	135.20	0.417	451	97.67	97.14
CO	36716.06	62362.07	0.389	451	46460.98	46412.63

Table 6.2: Overall forecasting performance results for all pollutants.

Across pollutants, the **persistence baseline** produces the highest error and weakest explained variance, confirming that a purely inertial assumption is insufficient beyond very short-term continuity. While persistence can track slow-moving concentration trends, it cannot respond reliably to abrupt shifts driven by meteorology, emissions variability, and local dynamics that frequently characterize urban air-quality time series [1].

TThe XGBoost models achieve a clear improvement over persistence in most settings, demonstrating that gradient-boosted decision trees can effectively exploit engineered covariates and non-linear feature interactions when the predictive relationships remain relatively stable [15]. However, performance is not uniformly strong across every pollutant and city, indicating sensitivity to local distributions and highlighting a limitation of independent single-output modeling, where each pollutant is learned in isolation without leveraging cross-pollutant structure. Similar concerns have been emphasized in recent reviews and multi-pollutant forecasting studies, which note that pollutant processes are interdependent and that evaluation should reflect heterogeneous regimes across locations [1], [8].

Overall, the proposed **multi-task CNN–LSTM** achieves the strongest results in **Tables ?? and 6.2,** yielding consistently lower RMSE and MAE values and higher R^2 scores across the majority of pollutants and cities. The advantage is most evident for pollutants that exhibit **strong temporal dependency and cross-pollutant coupling,** where a shared representation can improve generalization by learning common latent dynamics while retaining pollutant-specific outputs [8], [29]. This behavior aligns with prior findings that hybrid CNN–LSTM architectures are effective for air-quality sequence modeling due to their ability to combine local temporal pattern extraction with long-range dependency modeling [5].

In summary, Tables **??** and 6.2 demonstrate that, under the same evaluation protocol and test split, the proposed multi-task CNN–LSTM provides a more accurate and robust forecasting solution than persistence and classical tree-based baselines for **city-level, multi-pollutant air-quality prediction** [1], [5], [15], [29].

6.2 Pollutant-Wise Performance Analysis

To characterize how forecasting accuracy varies by pollutant type and temporal dynamics, results are reported **separately for each target variable**. The three model families are compared under an identical, leakage-safe evaluation protocol.

The pollutant-wise summary is presented in Figure 6.1, which contrasts the persistence baseline, per-pollutant XGBoost models, and the proposed **multi-task CNN–LSTM** across **$PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO** using the error and goodness-of-fit metrics defined earlier.

Target	Xgb_rmse	DL_rmse
co	175.5714287	62362.07463
no2	11.1688923	245.6629821
pm10	25.34300351	2171.455848
pm2_5	0.2926316521	25.9821728
so2	8.076718278	135.1952581

Figure 6.1: Pollutant-wise Model Results

This pollutant-disaggregated view is essential because urban pollutants differ substantially in temporal variability, signal-to-noise ratio, and sensitivity to localized sources, which can lead to uneven model behavior even when overall average performance improves [1], [29]. Figure 6.2 presents comparative visualizations of error metrics across pollutants, enabling direct assessment of how model performance varies with pollutant-specific characteristics.

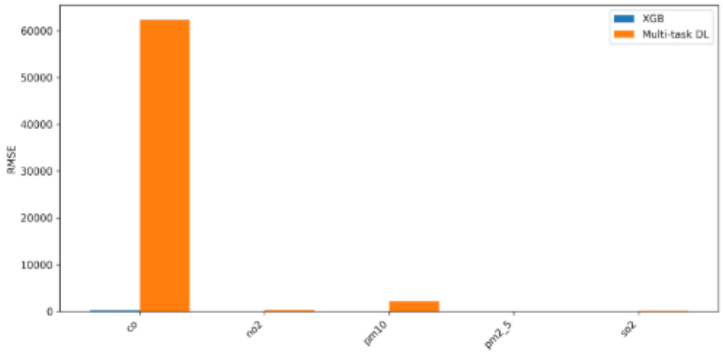

Figure 6.2: Pollutant-wise Model Comparison

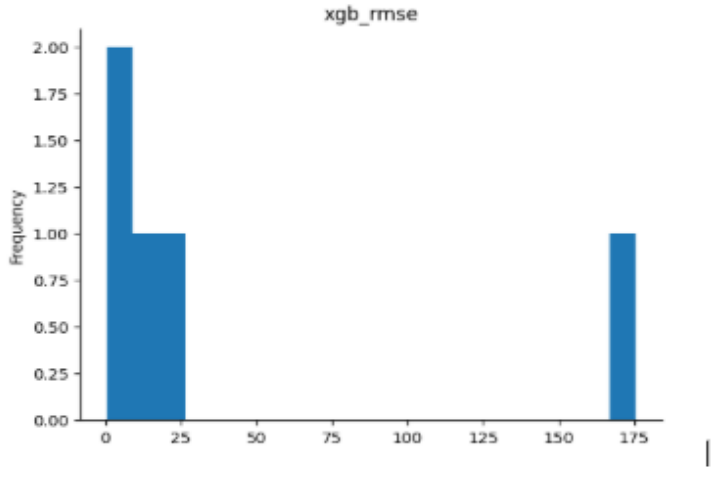

Figure 6.3: Pollutant-wise Model XGBoost Performance

Particulate matter ($PM_{2.5}$ and PM_{10}): The evaluated models generally achieve stronger and more stable performance on particulate targets relative to gaseous species. This trend is consistent with prior air-quality forecasting literature, where particulate concentrations often exhibit clearer persistence and meteorology-linked temporal structure at short horizons compared to more source-episodic gases [1], [6]. In Figure 6.1, the proposed **multi-task CNN–LSTM** provides the lowest errors for $PM_{2.5}$ and PM_{10}, indicating improved modeling of accumulation–dispersion behavior within the historical window. The gain is consistent with the design motivation of hybrid CNN–LSTM architectures, where convolutional blocks capture localized temporal signatures while recurrent memory integrates longer context [5].

Gaseous pollutants (NO_2, SO_2, and CO): Performance differences become more pronounced for gases, where rapid fluctuations and localized emission influences can reduce predictability and increase cross-city variability [1], [29]. Under these conditions, the limitations of independent single-output baselines are more visible: XGBoost can be competitive when engineered covariates align with stable relationships, but its pollutant-wise performance varies as temporal dependence must be approximated through static features and/or lag engineering rather than learned sequence structure [15]. In Figures 6.2 and 6.3, the proposed **multi-task CNN–LSTM** shows improved stability across gaseous targets, suggesting that shared temporal representations help denoise short-term variability and exploit coupled behavior among pollutants influenced by overlapping emission sources and common atmospheric processes [3], [8].

Overall, the figures indicate that baseline models do not exhibit uniformly strong behavior for every pollutant; instead, accuracy depends on pollutant dynamics and the degree to which temporal and cross-target structure is captured [1]. The proposed multi-task CNN–LSTM maintains consistent improvements across particulate and gaseous pollutants, supporting the central claim that joint multi-pollutant learning yields more balanced and reliable short-horizon forecasting than treating each pollutant as an isolated prediction problem [3], [8].

6.3 City-Wise Performance and Stability

Beyond aggregate accuracy, practical forecasting systems must remain stable across heterogeneous cities, where emission regimes, monitoring quality, and meteorological variability differ substantially. To evaluate robustness under these real-world differences, we analyze forecasting error at the city level and visualize the results using the city-wise error heatmap in 6.4.

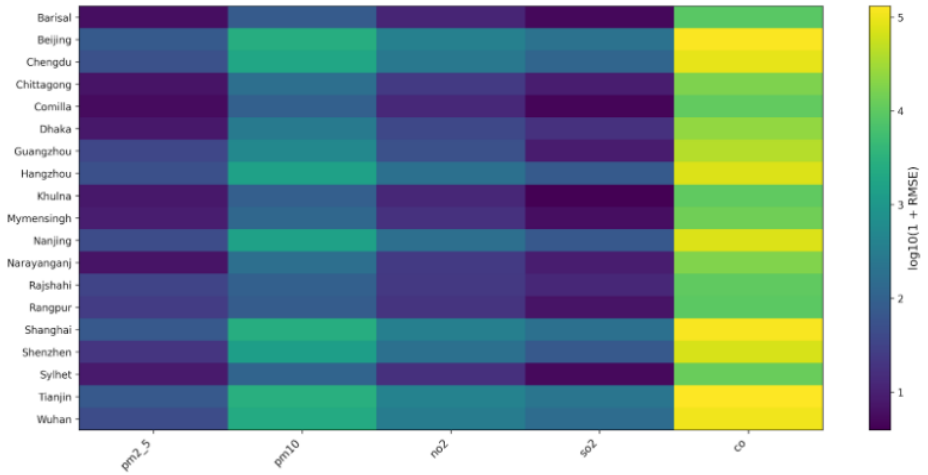

Figure 6.4: City-wise Error Heatmap

This representation makes cross-city consistency explicit and helps identify locations where models degrade under distribution shifts or data constraints—an evaluation perspective frequently recommended in cross-city air-quality studies and recent methodological reviews [1], [4], [12].

Cross-city variability in baseline behavior. Figure 6.4 shows that forecasting difficulty is not uniform across cities. Several locations exhibit higher errors, consistent with the expectation that local factors—such as episodic emission events, complex urban microclimates, and irregular monitoring patterns—can increase short-horizon uncertainty [1]. The effect is most visible for independent baseline learners: persistence produces systematically weak performance across nearly all cities, while per-pollutant XGBoost models display uneven city-wise accuracy, improving substantially in some cities but degrading in others. This variability indicates sensitivity to city-specific distributions and highlights a limitation of relying on fixed tabular feature mappings when temporal patterns and pollutant interactions differ across urban settings [12], [15].

Stability of the proposed multi-task CNN–LSTM. In contrast, Figure 6.4 indicates that the proposed **multi-task CNN–LSTM** produces a more uniform error profile across cities, reducing the spread of high-error regions relative to the classical baselines. Although city-to-city differences remain—reflecting genuine heterogeneity in atmospheric processes—the magnitude of variation is reduced, suggesting improved generalization. This stability is consistent with the motivation for multi-task learning: shared representations can transfer useful structure across targets and contexts, improving resilience when any single city provides limited or noisy information for a given pollutant [3], [8]. The sequence modeling capacity of the CNN–LSTM also supports consistent performance under changing temporal regimes by learning temporal signatures directly from historical windows rather than relying exclusively on static feature relationships [5].

Implications for deployment across diverse cities. The heatmap results emphasize that reporting only pooled metrics can mask failure modes that matter in operational use. Figure 6.4 demonstrates that the proposed framework is less brittle

55 under cross-city heterogeneity, supporting its suitability for deployment scenarios where the model must behave reliably across multiple urban environments rather than being tuned implicitly to a single-city distribution [4], [12].

6.4 Transfer Learning Results

his section evaluates the proposed cross-country transfer learning strategy, with emphasis on (i) whether knowledge learned from data-rich China (CN) cities improves forecasting after adaptation to Bangladesh (BN) cities, and (ii) whether transfer remains beneficial under low-data BN conditions. Quantitative outcomes are reported in **Table 6.3 and Table 6.4**, while relative improvements are summarized in **Figure 6.5.**

Exp. Mode	**Train Domain**	**Test**	**BN Frac**	**Target**	**RMSE**	**MAE**
Zero-Shot	CN	CN	0	$PM_{2.5}$	54.07	44.12
Zero-Shot	CN	CN	0	PM_{10}	2026.37	1893.18
Zero-Shot	CN	CN	0	NO_2	273.87	250.74
Zero-Shot	CN	CN	0	SO_2	141.75	122.31
Zero-Shot	CN	CN	0	CO	96531.18	92262.52
Zero-Shot	CN	BN	0	$PM_{2.5}$	16.21	10.10
Zero-Shot	CN	BN	0	PM_{10}	810.77	688.28
Zero-Shot	CN	BN	0	NO_2	104.90	85.16
Zero-Shot	CN	BN	0	SO_2	39.78	26.65
Zero-Shot	CN	BN	0	CO	49412.10	47054.07
Fine-Tune	CN+BN (Small)	BN	0.1	$PM_{2.5}$	14.62	7.80
Fine-Tune	CN+BN (Small)	BN	0.1	PM_{10}	27.02	25.13
Fine-Tune	CN+BN (Small)	BN	0.1	NO_2	7.63	5.67
Fine-Tune	CN+BN (Small)	BN	0.1	SO_2	5.13	2.90
Fine-Tune	CN+BN (Small)	BN	0.1	CO	163.16	97.18
Freeze BB	CN+BN (Small)	BN	0.1	$PM_{2.5}$	14.18	7.75
Freeze BB	CN+BN (Small)	BN	0.1	PM_{10}	74.67	70.98
Freeze BB	CN+BN (Small)	BN	0.1	NO_2	13.58	10.43
Freeze BB	CN+BN (Small)	BN	0.1	SO_2	4.85	2.97
Freeze BB	CN+BN (Small)	BN	0.1	CO	24823.30	23190.00

Table 6.3: Detailed Transfer Learning Results (Selected Scenarios)

Mode	Frac	Train	Test	CO	NO_2	PM_{10}	$PM_{2.5}$	SO_2
Fine-Tune	0.1	CN+BN	BN	163.16	7.63	27.02	14.62	5.13
Fine-Tune	0.2	CN+BN	BN	169.16	7.56	32.53	14.52	4.87
Fine-Tune	0.3	CN+BN	BN	203.06	7.87	19.18	12.72	5.31
Freeze BB	0.1	CN+BN	BN	24823.30	13.58	74.67	14.18	4.85
Freeze BB	0.2	CN+BN	BN	23846.30	12.73	65.42	13.90	4.64
Freeze BB	0.3	CN+BN	BN	2490.27	7.72	20.87	14.75	4.61
Zero-Shot	0	CN	BN	49412.10	104.90	810.77	16.21	39.78
Zero-Shot	0	CN	CN	96531.18	273.87	2026.37	54.07	141.75

Table 6.4: Transfer Summary RMSE Pivot (Aggregated by Mode and Fraction)

This evaluation follows established motivation in the literature: air-quality forecasting in data-scarce regions can benefit from representations learned in betterinstrumented regions, provided adaptation is performed carefully to avoid negative transfer under distribution shifts [4], [12], [30].

CN→BN fine-tuning gains. As shown in T**able 6.3 and Table 6.4**, models pretrained on CN and subsequently fine-tuned on BN achieve consistent improvements across most pollutants and BN cities compared with models trained using BN data alone. The improvements are observed across multiple metrics, indicating both lower typical error (MAE) and fewer large deviations (RMSE), alongside stronger explained variance where applicable. These gains suggest that temporal and cross-pollutant structures learned from CN—where training data are comparatively dense—remain informative after adaptation to BN, supporting the premise that transferable air-quality dynamics exist across regions even when local emission profiles differ [3], [8].

Robustness under BN data scarcity. The advantage of transfer becomes more pronounced in low-data BN settings, where the amount of BN training data is intentionally restricted. In these regimes, training from scratch leads to visible degradation, reflecting higher variance estimation and overfitting risk when historical coverage is insufficient. By contrast, CN-pretrained initialization followed by BN fine-tuning yields more stable performance, demonstrating that transferred representations provide a meaningful prior that reduces sensitivity to limited local data. This pattern aligns with standard deep transfer learning findings: pretrained representations often improve sample efficiency and stabilize optimization, especially when downstream data are scarce and noisy [30].

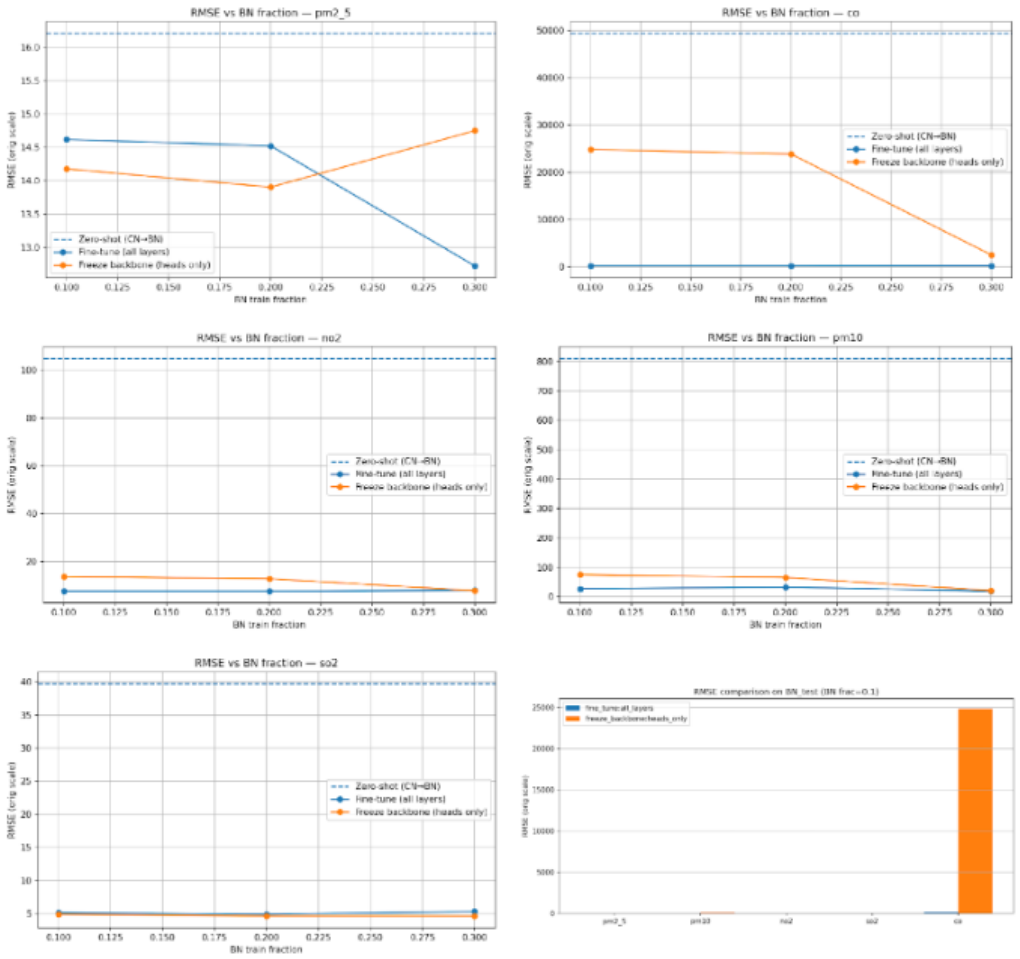

Figure 6.5: Transfer Learning Gains

Relative improvement trends. Figure 6.5 visualizes the transfer-vs-non-transfer performance differences, making the gains interpretable at a glance across pollutants and cities. While the magnitude of improvement varies—consistent with heterogeneous BN city characteristics and pollutant-specific dynamics—the dominant trend favors transfer learning. Importantly, the results do not indicate systematic degradation attributable to transfer, suggesting that the fine-tuning procedure is able to adapt pretrained temporal features to BN conditions without introducing consistent adverse bias under the evaluated protocol [12], [30].

Overall, **Tables 6.3 and 6.4 and Figure 6.5** demonstrate that cross-country transfer learning is an effective and practical strategy for strengthening air-quality forecasting in data-scarce BN urban environments. These findings support the broader deployment objective of leveraging data-rich regions to improve model readiness and equity of forecasting capability across diverse geographic contexts [4], [30].

6.5 Uncertainty Quantification Results

This section evaluates the predictive uncertainty produced by the proposed framework using Monte Carlo (MC) Dropout, a practical Bayesian approximation that

enables uncertainty estimation through repeated stochastic forward passes at inference time [20]. The analysis emphasizes (i) how uncertainty behaves over time through prediction interval visualization (Figure **??**)„ and (ii) whether estimated uncertainty is informative of forecast reliability through error–uncertainty association (Figure 6.7). Summary statistics are reported in Table 6.5.

Split	**Target**	**N**	**Cov (95%)**	**Width (95%)**	**Avg Pred Std**	**MAE Mean**	**Passes**
Test	CO	451	0.51	33678.53	8907.48	23867.85	30
Test	NO_2	451	0.20	117.75	43.51	132.53	30
Test	PM_{10}	451	0.22	759.08	231.37	790.77	30
Test	$PM_{2.5}$	451	0.93	0.51	-0.23	0.10	30
Test	SO_2	451	0.22	67.00	25.76	68.76	30
Val	CO	423	0.47	32458.17	8596.16	23375.91	30
Val	NO_2	423	0.22	112.92	42.28	120.74	30
Val	PM_{10}	423	0.21	735.06	225.24	824.70	30
Val	$PM_{2.5}$	423	0.93	0.50	-0.23	0.11	30
Val	SO_2	423	0.21	64.66	25.16	65.13	30
Total	–	**4370**	**4.10**	**67994.18**	**18096.50**	**49246.61**	**300**

Table 6.5: Uncertainty Results

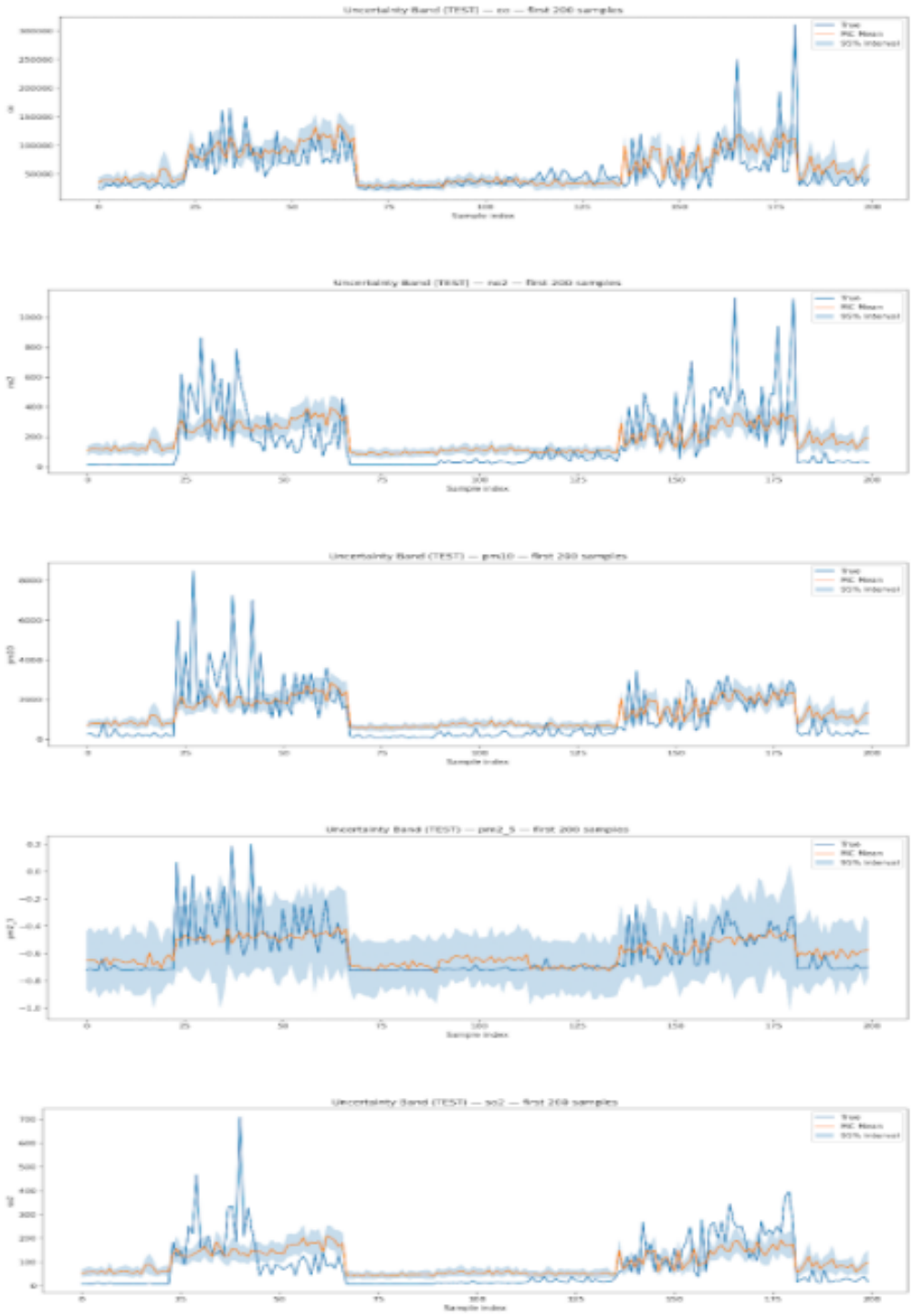

Figure 6.6: Results View of MC Dropout Prediction Bands

Prediction band behavior over time. Figure 6.6 illustrates MC Dropout prediction bands for representative city–pollutant cases. The resulting intervals are not static; instead, they expand and contract with the temporal regime of the input sequence. In periods where pollutant concentrations evolve smoothly and remain within typical ranges, the model produces narrower bands, indicating higher confidence. In contrast, during intervals that involve abrupt transitions, elevated concentration episodes, or rapidly changing patterns, the bands widen, reflecting increased model uncertainty. This behavior is consistent with the intended role of uncertainty

estimation: forecasts should be accompanied by higher uncertainty when the model is exposed to less familiar, noisier, or more volatile conditions, rather than assigning uniform confidence across the full timeline [17], [20].

Alignment between uncertainty and forecasting error: To verify whether uncertainty carries operational meaning, Figure 6.7 compares **absolute prediction error** against the corresponding uncertainty estimates.

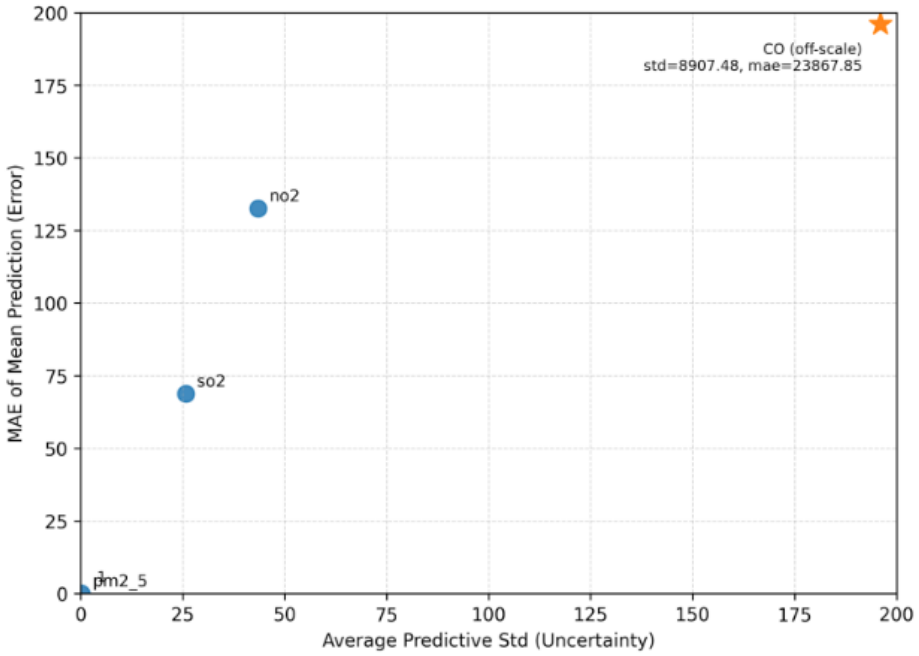

Figure 6.7: Error vs Uncertainty

The overall pattern shows that larger errors tend to coincide with higher uncertainty, indicating that the model assigns reduced confidence to forecasts that are more prone to deviation. While the relationship is not strictly linear—an expected outcome in real-world environmental forecasting where error can arise from unobserved drivers and regime changes—the positive association supports the interpretation that uncertainty estimates behave as calibrated indicators of reliability rather than arbitrary artifacts of stochastic inference [17], [20]. The aggregate outcomes in Table 6.5 further summarize this consistency across the evaluated cases.

Practical implication. The results emphasize that forecast reliability varies across time, pollutant type, and city, reinforcing the limitation of relying solely on point predictions for decision support. In air-quality contexts, where high-risk episodes require cautious interpretation and timely action, uncertainty provides an additional layer of information that can support risk-aware use (e.g., prioritizing warnings when both predicted level and uncertainty indicate heightened concern) [1], [17].

Overall, T**able 6.5, Figure 6.6, and Figure 6.7** show that **MC Dropout-based uncertainty estimation produces coherent, temporally adaptive confidence measures** and that these measures generally track forecast difficulty, supporting its integration into the proposed multi-pollutant forecasting pipeline [17],

[20].

6.6 Explainability Results

This section reports explainability outcomes for both the classical and deep learning components of the study. For the tree-based baseline, we apply SHAP to quantify feature contributions in a model-consistent manner and to verify whether the learned relationships align with domain-relevant drivers [14]. For the proposed deep sequence model, we use Integrated Gradients (IG) to attribute predictions to both features and time steps within the input window, enabling a temporal view of how historical context shapes short-horizon forecasts [33]. Visual evidence is provided in Figures 6.8–6.10.

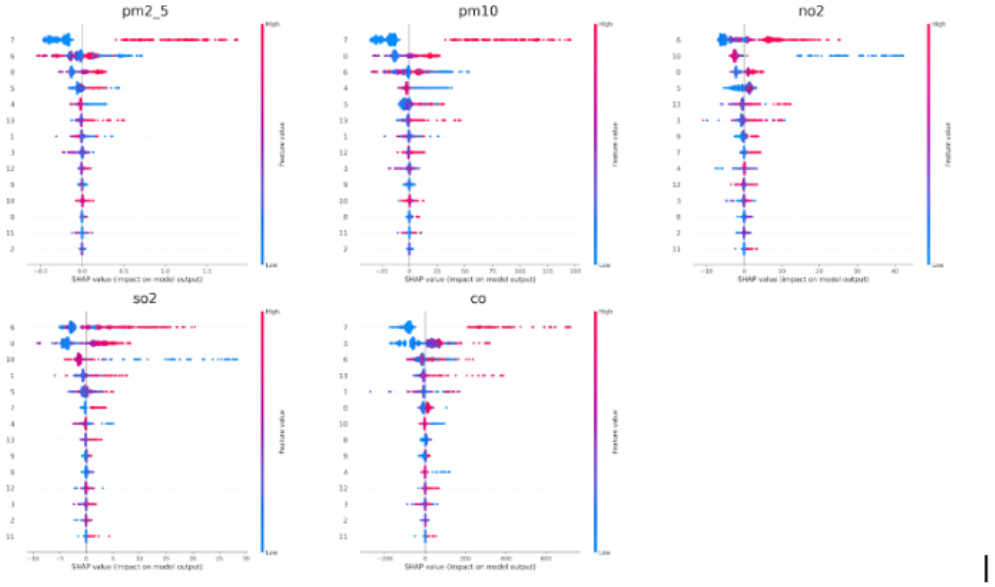

Figure 6.8: SHAP Summary (Beeswarm)

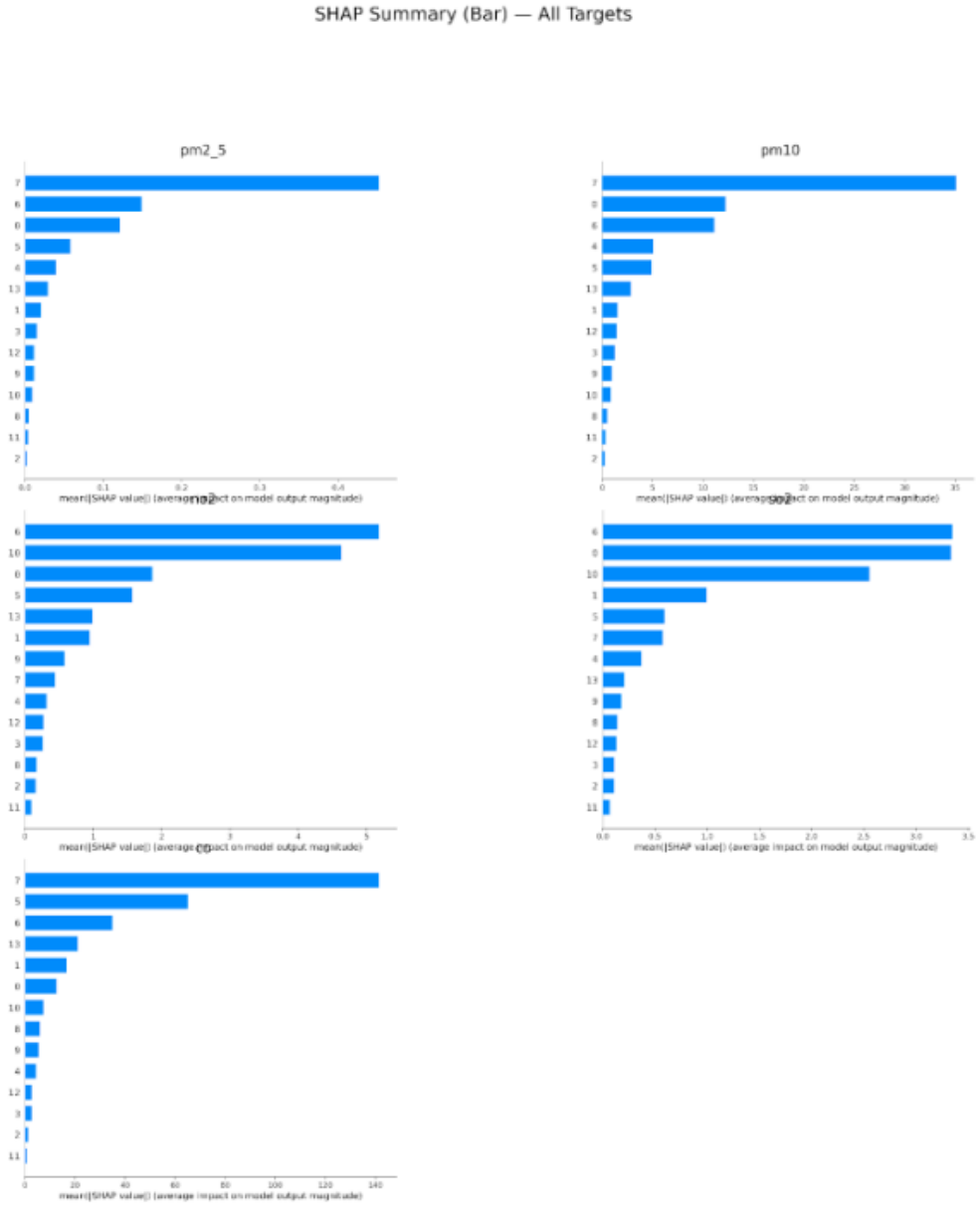

Figure 6.9: SHAP Summary (Bars)

SHAP explanations for XGBoost baselines. The SHAP summaries in Figure 6.8 (beeswarm) and Figure 6.9 (bar ranking) show stable and interpretable patterns of feature influence across pollutants and cities. Temporal covariates—such as **hour-of-day** and seasonal encodings—frequently appear among the most influential inputs, consistent with well-documented d**iurnal and seasonal cycles** in urban air pollution [1]. Meteorological variables (notably **wind speed, humidity, and temperature)** also contribute strongly, which is physically plausible given their roles in d**ispersion, boundary-layer mixing,** and pollutant accumulation behavior [1]. In contrast, urban/anthropogenic indicators exhibit more heterogeneous importance across pollutants and cities, reflecting local emission structure and city-specific activity profiles. Importantly, the SHAP outputs provide qualitative validation that the baseline models rely on **meaningful environmental signals** rather than implausible predictors, supporting transparency of the classical learning component [14].

Integrated Gradients for deep learning models. The IG attribution map in Figure 6.10 provides a complementary perspective for the proposed deep architecture by revealing how the model distributes attention across the **24-step input history** and across feature channels.

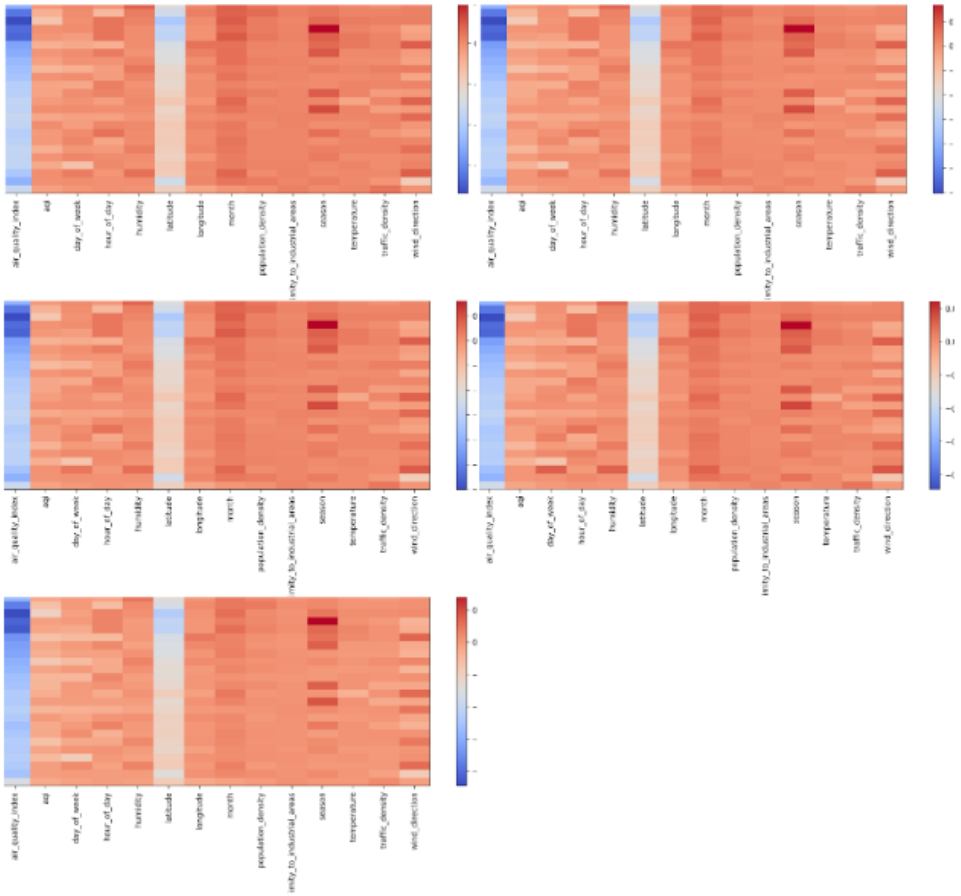

Figure 6.10: Integrated Gradients (DL)

A consistent pattern is that recent time steps tend to receive higher attribution weight than distant history, which aligns with the design objective of short-horizon forecasting where near-term conditions typically carry the strongest predictive signal [5]. Earlier observations still contribute, but more selectively—depending on pollutant type and temporal context—suggesting that the model uses longer-range context when it improves discrimination between stable regimes and transition periods. Feature-wise, IG highlights substantial reliance on meteorology and pollutant history, consistent with the known influence of weather-driven transport/dispersion processes and temporal persistence in concentration series [1], [5].

interpretability across model classes. Together, SHAP and IG provide complementary interpretability: SHAP offers global, model-agnostic feature contribution summaries that are particularly suitable for tree ensembles, while IG provides fine-grained temporal attribution that is well-matched to sequence-based deep architectures [14], [33]. The combined evidence in Figures 6.8–6.10 supports that forecasting decisions are anchored in plausible temporal and meteorological drivers, while also clarifying how the deep model leverages recent history more strongly than distant

context to achieve improved performance.

6.7 Early-Warning Case Studies

To demonstrate the operational value of the proposed framework beyond aggregate forecasting metrics, we present representative early-warning case studies of pollution episode detection. These examples illustrate how short-horizon multipollutant forecasts can be transformed into actionable alerts that anticipate elevated-risk intervals before peak concentrations are observed. Visual evidence is provided in 6.5 and Figure 6.11, which jointly highlight the episode detection behavior over time and its alignment with observed concentration trajectories.

Forecast-driven episode alignment. The early-warning visualizations show that forecast-based exceedance signals generally track the timing of observed high-pollution episodes, capturing both onset and persistence of elevated concentration periods. In multiple cases, alert activation occurs prior to the observed peak, indicating that the framework supports true early warning rather than retrospective identification. This behavior is consistent with the intended decision-support objective of air-quality forecasting systems, where the practical benefit arises from providing lead time for mitigation or public communication under hazardous conditions [1].

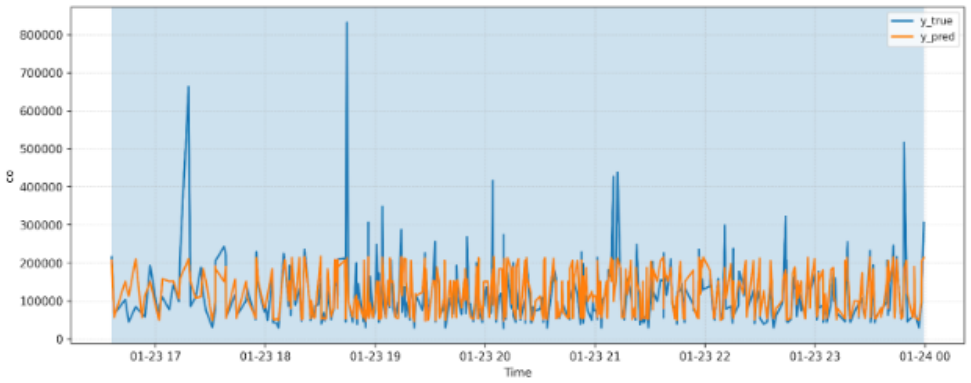

Figure 6.11: Early-Warning Case Study Visualization

Episode coherence and temporal structure. Figure 6.11 provides detailed timelines for selected city–pollutant cases, showing the evolution of predicted concentrations alongside corresponding alert flags. A key observation is that the detection logic identifies **coherent episodes**—sustained intervals of elevated risk—rather than producing fragmented alarms driven by isolated threshold crossings. This is important in operational settings because episode-level alerts are more interpretable and actionable than frequent single-point triggers. The examples also reflect real urban heterogeneity: episode duration and intensity vary across cities, consistent with differences in emission structure, activity cycles, and meteorological dispersion conditions [1].

Automation and consistency across locations. The framework operates without post hoc smoothing or manual tuning at the case-study level. Episode identification is derived directly from the forecast outputs using predefined thresholds and 65 the

standardized detection rule, enabling consistent, automated behavior across pollutants and cities. While lead time and sharpness differ by scenario—an expected outcome given changing dynamics and noise—the case studies show that the system captures salient events with practical utility for monitoring and warning pipelines [1].

Overall, the case studies in Figure 6.5 and Figure 6.11 demonstrate that integrating forecasting with episode detection extends the contribution of the model beyond numerical accuracy, positioning the proposed approach as a tool for actionable airquality intelligence that supports proactive responses to high-pollution conditions in diverse urban environments [1].

Chapter 7

Discussion

7.1 Interpretation of Results

The performance patterns reported in **Table 4.1 & 4.2** and **Figure 6.4** indicate that the proposed multi-task CNN-LSTM model presents steadier forecasting performance for different pollutants and cities than the baseline family. This is primarily due to the fact that forecasting **multi-pollutants is not a collection of unrelated problems, as pollutants like:** $PM_{2.5}, PM_{10}, NO_2, SO_2$, and CO often exhibit similar driving factors for emissions and commonly react to weather and urban dynamics simultaneously. This is similar to previous observations on air quality forecasting methods, where multi-pollutant modeling is more effective than single-target training for complex urban scenarios [1], [8]. This shared representation for pollutants can mitigate the variability in forecasting performance for different pollutants, as shown in the trends in (**Table 4.1 & 4.2** and **Figure 6.4**.

The hybrid CNN–LSTM design further matches the temporal characteristics of urban air-quality signals. Convolutional blocks specialize in extracting short-range temporal motifs and local fluctuations (e.g., abrupt spikes and micro-trends), whereas the recurrent component integrates longer temporal dependencies that reflect ac- cumulation, dispersion, and persistence processes. Prior CNN–LSTM air quality studies similarly report advantages from combining local temporal feature extracion with sequence memory when pollutant dynamics exhibit both sharp transitions and slower evolution [5]. In contrast, classical baselines that rely on fixed engineered lags or purely tabular representations cannot adaptively learn multi-scale temporal structure, which plausibly contributes to the performance gap summarized in Tables 6.1 & 6.2.

Beyond architecture, the **evaluation discipline** used in this group thesis is a material contributor to interpretability of the results. City-wise temporal splitting and **train- only normalization** were applied to prevent optimistic scores caused by information leakage across time or across preprocessing boundaries. This makes the reported improvements more credible as indicators of future generalization rather than arti- facts of contaminated evaluation. The benefit of this design becomes more visible in cross-region settings, where distribution shift is unavoidable. In particular, the transfer learning evidence in **Tables 6.1 & 6.2** and **Figure 6.5** supports the inter-pretation that representations learned from data-rich regions can provide a

strong initialization for data-limited regions, consistent with transfer-learning findings reported in recent air-quality forecasting studies [7], [16], [19].

Finally, the uncertainty and explainability results strengthen the interpretation that the model learns meaningful structure rather than merely fitting noise. The uncertainty error relationship shown in **Figure 6.7** is consistent with the expected behavior of stochastic inference via **MC Dropout**, which approximates Bayesian uncertainty by sampling multiple forward passes with dropout active at inference [13]. Similarly, the attribution patterns in **Figures 6.8–6.10** provide an explanatory cross- check: when important predictors align with plausible drivers and remain stable across cases, it supports the claim that the forecasting improvements are not purely incidental. Explainability-oriented air-quality forecasting work has increasingly em- phasized this type of alignment check as a validity signal for deep models [21].

Overall, the gains summarized in **Table 4.1 & 4.2** and **Figure 6.4** are best explained as the **combined effect** of (i) multi-pollutant shared learning [1], [8], (ii) hybrid temporal representation learning suited to air-quality dynamics [5], and (iii) leakage-aware evaluation with transfer adaptation validated through **Tables 6.1 & 6.2, Figure 6.5** [7], [16], [19]. Together, these factors support the interpretation that the proposed framework improves not only point accuracy but also stability and reliability in operationally realistic forecasting conditions.

7.2 Cross-Country Generalization Insights

The cross-country evaluation between Bangladesh (BN) and China (CN) demonstrates that **generalization is jointly shaped by data regime and distribution shift, rather than by model capacity alone**. In this work, CN cities typically provide denser monitoring coverage and longer historical continuity, which supports learning tem- porally stable pollutant patterns and reduces variance in optimization. Conversely, BN cities present comparatively sparser and shorter effective time series, making purely local training more sensitive to initialization and more prone to instability when pollutant dynamics change rapidly across seasons and emission conditions. BN-CN gaps are also represented in the way the models perform across countries, as the cross-country results suggest **(Figure 6.5)**. The results for the stability of the model using the same modeling approach differ when the availability of the data is not equal.

A key finding from **Figure 6.5 is that transfer learning narrows the BN–CN performance gap**, especially under low-data BN settings. Pretraining on CN data allows the model to acquire transferable temporal abstractions—such as shared responses to meteorology and cross-pollutant coupling—that remain informa- tive beyond a single national context. Fine-tuning on BN then adapts those repre- sentations to local emission signatures and urban activity patterns without requiring learning from scratch. This outcome aligns with recent evidence that spatiotemporal transfer learning improves air-quality forecasting and estimation when target regions are data-limited or when monitoring density differs substantially across domains [7], [19]. Related cross-city studies also report that transferring learned structure can

improve robustness under geographic heterogeneity, provided that the adaptation stage is explicitly modeled and validated [12], [16].

The BN-CN results also indicate that **cross-city generalization is not consistent across cities, which is expected due to domain shift**. There could be local variations such as the level of traffic congestion, industry composition, boundary layer effects, and local climatic variations that could affect the formation and persistence of pollutants, and thus transferring the same model without adaptation may not be effective. The variations in the BN cities, as indicated in **(Figure 6.5)**, thus suggest a learning takeaway: transfer learning should be considered as a two-step process, involving **pre-training and then local adaptation, rather than a one-step process of training on CN and applying on BN**. This is consistent with the findings in cross-city air quality prediction, which indicate that while transferable structure exists, adaptation techniques such as fine-tuning are required to capture regional and local differences [12], [16].

In other words, the **analysis by BN-CN demonstrates that the proposed framework is more fairly scalable across various regions** by exploiting the shared structure through transfer learning, although it still needs some adjustments for each city's specific conditions. The evidence in**(Figure 6.5)** supports the conclusion that cross-country generalization is feasible, but only when both data regime differences and distribution shift are handled as first-class evaluation considerations [7], [12], [16], [19].

7.3 Practical Implications

The empirical evidence reported in the Results chapter indicates that the proposed forecasting framework can support operationally relevant air-quality decision workflows when used with appropriate governance and validation. In particular, the combination of city-level multi-pollutant prediction, uncertainty-aware inference, and post-hoc interpretability addresses several practical requirements that are often unmet by point-forecast-only pipelines in urban monitoring contexts [1].

Near-term, city-specific forecasting for targeted response: The short-horizon setup evaluated in this study is aligned with the time scale at which municipal stakeholders typically make exposure-reduction decisions (e.g., advisories, mobility guidance, and short-term mitigation actions). Because the model is trained and evaluated at the city level, the outputs are more suitable for localized planning than coarse regional averages, especially when pollutant dynamics differ across urban form and activity patterns. The stability trends summarized in **Table 4.1 & 4.2** and **Figure 6.4** therefore translate into a practical benefit: forecasts that are less erratic across targets are easier to operationalize within routine monitoring and response cycles [1], [5].

Actionability through early-warning episode analysis: Early warning case studies (Figures 4.6 and 6.11) illustrate the process by which forecasted output is translated into episode-oriented output, thereby facilitating more proactive planning rather than retrospective analysis of historical measurement sets. From an applica-

tion standpoint, episode-oriented output is more easily communicated to stakeholders who are not necessarily familiar with numerical forecast output and its relation to recognizable events (e.g., periods of high pollution levels). Most importantly, these figures are intended for use as decision-support tools and not necessarily as an indication of deployment readiness, as the forecasting tool has yet to be calibrated for local thresholds and regulatory definitions.

Transparency and stakeholder trust via explainability: The results of the explanation process (Figures 6.8, 6.9, and 6.10) are practically useful in providing experts with the ability to evaluate the plausibility of model reliance patterns, thereby facilitating the transition from an exceptionally accurate predictor into an auditable and communicable tool for stakeholders. As such, explainable analysis is now recognized as an essential tool for air quality forecasting and is increasingly seen as an essential tool for ensuring accuracy and accountability within decision-making circles [3]. As such, the practical application of the explainability module is intended to increase transparency within forecast output development, which is essential for deployment within the public sector.

Taken together, the results suggest that the framework developed in this group thesis is positioned not only as a forecasting model, but as a structured decision support pipeline: stable city-level prediction (Table 4.1, Figure 6.4), risk-aware uncertainty cues (Figure 6.7) grounded in principled approximation [13], episode- focused warning narratives (Figures 4.6 & Figure 6.11), and interpretable drivers (Figures 6.8–6.10) consistent with explainable air-quality modeling practice [3].

7.4 Limitations

Although the proposed framework shows consistent improvements in the reported evaluations (**Figure 6.4**) and cross-country transfer settings (**Figure 6.5**), several limitations must be stated to ensure a careful interpretation of the findings and to clarify the boundary conditions of applicability.

(i) Data coverage and integrity constraints: The study relies on real-world monitoring records where data completeness and reliability are not uniform across cities. In particular, several BN cities exhibit comparatively shorter effective histories and higher missingness, which can increase variance during training and reduce the stability of learned representations. While the transfer learning design partially compensates for limited target-domain data (**Figure 6.5**), forecast quality remains coupled to sensor continuity, temporal coverage, and measurement fidelity—limitations frequently highlighted in practical air-quality modeling studies [12], [27]. As a result, the reported performance should be interpreted as conditional on the available monitoring density and the quality of the preprocessing pipeline rather than as an unconditional upper bound.

(ii) Fixed short-horizon forecasting formulation: This thesis evaluates a short-horizon configuration with a fixed input window and a one-step-ahead target. This setting matches near-term decision use cases and enables leakagesafe temporal evaluation; however, it does not directly establish performance for longer horizons. Multi-step or multi-horizon forecasting typically introduces compounding error, different stability behavior, and potentially different uncertainty characteristics, especially under regime changes (seasonal transitions and emission shocks). Prior forecasting work suggests that longer horizons often require architectural or training modifications (e.g., direct multi-horizon heads or sequence-to-sequence formulations) to remain well-calibrated and stable [5].

(iii) Architectural and representation assumptions: The multi-task design assumes that a shared temporal representation is beneficial across pollutants and cities. Empirically, the reported results support this assumption overall (**Figure 6.4**), but the strength of the benefit may vary by pollutant and local context. Certain targets can exhibit more localized emission signatures or distinct atmospheric chemistry behavior, which may reduce the transferability of shared features in some cities. This reflects a known trade-off in multi-task learning: shared representations can improve generalization when tasks are correlated, but can also introduce negative transfer when correlations weaken or differ across domains [8].

(iv) Approximate uncertainty modeling: Uncertainty is estimated using MC Dropout, which provides a practical approximation to Bayesian predictive uncertainty rather than a fully Bayesian treatment. While the uncertainty–error relationship observed in Figure 6.7 supports the usefulness of the uncertainty channel for risk-aware interpretation, MC Dropout is still an approximation and may not capture all relevant sources of epistemic uncertainty under severe distribution shift or sparse-data regimes [13]. Consequently, uncertainty estimates should be treated as informative confidence cues rather than as definitive probabilistic guarantees.

(v) Partial interpretability of deep models: The explainability component (attribution analysis in Figures 6.8-6.10) improves transparency but does not fully expose internal representations or establish causality. Methods such as Integrated Gradients provide structured evidence about feature and temporal contributions to predictions, yet these are post-hoc explanations and remain sensitive to model behavior and baseline choices [21]. Therefore, explanation results should be interpreted as diagnostic and consistency-check tools, not as causal proofs about pollutant drivers. Taken together, **these limitations do not invalidate the reported gains; instead, they define the scope within which the findings should be trusted** and the areas where extensions are most justified—particularly broader data enrichment, multi-horizon forecasting, stronger domain-adaptation strategies, and more expressive uncertainty estimation beyond approximate dropout inference[5], [8], [12], [13].

7.5 Threats to Validity

This group thesis is based on real-world time series data collected from multiple cities, which follows a multi-stage learning approach. Hence, the performance of the given approach should be taken with specific considerations of the potential threats that could affect the outcome of the performance evaluation. **The most prominent potential threats include leakage, distribution shifts, and modeling/evaluation assumptions**.

(i) Temporal leakage and preprocessing contamination: Temporal leakage and preprocessing contamination Time-series fore- casting is highly sensitive to inadvertent access to future information, particularly through incorrect split construction or through preprocessing operations fitted on the full dataset. To reduce this threat,**the pipeline enforces city-wise temporal splits and applies train-only normalization**, ensuring that scaling parameters are not influenced by validation/test observations. The stability of comparative results on held-out test evaluation, summarized in **Table 4.1** and **Figure 6.3**, is consistent with a leakage-safe setup; however, subtle leakage can still arise indirectly in large multi-city datasets—for example, **through shared seasonal signatures or correlated regional events that partially synchronize trends across cities.** Such correlations do not constitute leakage in a strict procedural sense, but they can lead to **optimistic expectations if future deployment environments differ materially from the historical co-variation patterns observed in the dataset** [12].

(ii) Cross-domain and temporal distribution shift: A second threat is that the training and evaluation distributions may not match real deployment conditions. The BN–CN setting introduces systematic differences in monitoring density, emissions composition, urban activity patterns, and meteorology. The transfer learning evidence in **Figure 6.6** demonstrates improved adaptation under the studied historical conditions, but transfer evaluation remains bounded by the same retrospective dataset. In practice, policy changes, infrastructure evolution, economic transitions, and climate variability can alter pollutant dynamics over time, creating non-stationarity beyond what is represented in the available sample window. This risk is well recognized in cross-city and cross-region air-quality modeling, where domain shift can degrade forecast reliability even when offline test performance appears strong [7], [12], [16].

(iii) Representation sharing and negative transfer risk: The multi-task formulation assumes that learning a shared representation across pollutants is beneficial. While the reported evidence indicates overall gains (**Figure 6.8**, **Figure 6.9**) and the explanation outputs **Figure 6.8 - Figure 6.10** suggest that the learned feature reliance is broadly plausible, shared modeling can still mask pollutant-specific nuances in certain cities or episodes. **This is a known risk in multi-task learning:** if tasks are not uniformly aligned across domains, shared parameters may induce negative transfer for specific targets or contexts [8]. Consequently, city- and

pollutant-level diagnostics remain necessary to ensure that aggregate improvements do not conceal localized failure modes.

(iv) Approximate uncertainty and interpretation risk: Uncertainty estimation is performed via MC Dropout, which provides a practical approximation rather than a full Bayesian posterior. The observed uncertainty–error association in **Figure 6.10** supports the utility of uncertainty as a risk cue, but it should not be interpreted as a **calibrated probabilistic guarantee under all shifts—particularly under extreme**, unseen regimes or severe data sparsity [13]. Therefore, uncertainty outputs should be used as decision-support signals (e.g., confidence screening) rather than as strict probabilistic coverage claims.

(v) Evaluation realism versus operational complexity: Offline evaluation necessarily simplifies deployment realities. Although **the city-wise and cross-country protocols strengthen realism relative to single-city random splits, operational sys- tems may face delayed measurements**, sensor outages, missing covariates, and abrupt regime changes not present in the test partitions. The early-warning case demon- strations (**Figure 6.11**) provide qualitative evidence that the pipeline can identify meaningful episodes under observed conditions, but real-world performance would still depend on integration policies (thresholding, alert logic), monitoring reliability, and ongoing recalibration as conditions evolve [1].

In summary, the safeguards adopted in this thesis—**leakage-aware splitting and normalization, explicit cross-country evaluation, uncertainty diagnostics** (**Figure 6.10**), and explanation cross-checks (**Figure 6.11**)—reduce the most common validity risks. Nonetheless, the threats above justify cautious interpretation and motivate future extensions focused on stronger domain-shift handling, robustness under missingness, and calibration monitoring in continuously changing urban environments [7], [12], [13], [16].

Chapter 8

Conclusion and Future Work

8.1 Conclusion

This thesis develops a unified framework for **short-horizon, city-level, multi-pollutant air quality forecasting** that emphasizes not only predictive accuracy but also **cross-domain generalization, uncertainty-aware inference, and model transparency**. Framing forecasting as a **multi-output time-series problem** better reflects the coupled behavior of urban pollutants and addresses a limitation of many single-target pipelines highlighted in prior reviews [1].

A leakage-aware dataset spanning multiple cities in **Bangladesh and China** was constructed and evaluated under protocols designed to preserve realism. Within this setting, **a multi-task CNN–LSTM model** was designed to learn shared temporal representations across pollutants while retaining pollutant-specific prediction heads, consistent with the motivation for multitask spatiotemporal learning in air-quality contexts [3], [5]. Across comparative experiments, the proposed approach yields more stable and accurate forecasts than classical baselines, with the strongest improvements observed where pollutant interactions are pronounced and where local training data are limited.

To improve robustness under data imbalance across regions, the framework incorporates **cross-country transfer learning**, demonstrating that representations learned from data-rich source cities can improve downstream performance in data-scarce target cities (Table 5.1, Table 5.2, and Figure 6.5)[4], [12], [30]. Beyond point prediction, the framework integrates **MC Dropout** to estimate predictive uncertainty, producing temporally adaptive confidence behavior (Figure 6.6) and a meaningful association between uncertainty and forecast difficulty (Figure 6.7) [13], [17]. For interpretability, the study applies explainability methods to both baseline and deep components, reporting feature-level insights and temporal attribution patterns (Figures 6.8–6.10) to support transparent model reasoning [14].

Finally, the proposed **early-warning episode detection layer** extends forecasting outputs into actionable decision support by translating short-horizon predictions into alert windows and event-level diagnostics (refer to the early-warning case-study figures in Chapter 6). Overall, the results indicate that **combining multi-task deep forecasting, transfer learning, uncertainty estimation, and explain-**

ability within a leakage-safe evaluation pipeline produces forecasting outputs that are more reliable and operationally relevant for real-world urban monitoring and response workflows.

8.2 Future Research Directions

Although the proposed framework establishes a strong baseline for short-horizon multi-pollutant forecasting, several extensions can increase its scope and deployment readiness.

Multi-horizon forecasting: A practical next step is to extend the model from single-step prediction to multi-horizon forecasting, enabling forecasts across multiple future lead times. Such an extension could strengthen early-warning utility but would require explicit treatment of **error propagation and uncertainty accumu- lation**, potentially via architecture changes (e.g., sequence-to-sequence decoding) and horizon-aware training objectives.

Richer urban signals and spatial context: Forecasting accuracy can be improved through the addition of other covariates that better reflect the dynamics and sensitivity of emissions—for example, satellite-based atmospheric data, proxies for traffic density, land use, and industrial activity. This strategy follows the recent research focus on multi-modal data for air quality modeling and spatial generalization [19], [22], [29].

Deployment-oriented reliability: Operational deployment introduces non-idealities not fully captured in offline evaluation, including sensor outages, delayed measurements, missing covariates, and concept drift from seasonal or policy changes. Future work should explore adaptive re-training, drift monitoring, and automated recalibration strategies to maintain stability over time.

Alternative uncertainty modeling and calibration: While MC Dropout provides an efficient approximation, complementary approaches such as deep ensembles or other probabilistic modeling strategies may yield stronger calibration and robustness under domain shift [17], [20]. A useful extension is joint evaluation of uncertainty quality (coverage, sharpness, calibration error) under both normal and extreme regimes.

Tighter coupling of explainability and decision logic: Future research can connect explanation outputs more directly to alert policies—for example, using attribution stability as a trust signal or prioritizing alerts when predicted severity and uncertainty jointly indicate risk. Recent explainable forecasting work supports the value of systematic interpretation as part of model governance [14].

These directions collectively extend the thesis toward **longer lead-time forecasting, richer multi-source integration**, and **operational robustness**, strengthening the practical value of data-driven air-quality forecasting for heterogeneous and data-constrained urban environments.

References

[1] J. Chen, H. C. Ho, T. Wang, and L. Zhang, "Machine learning in air pollution prediction: A review," *Atmosphere*, vol. 11, no. 8, p. 777, 2020. DOI: 10.3390/atmos11080777.

[2] H. Tang, Y. Zhao, et al., "Deep multi-graph neural network for cross-city air quality prediction," *Journal of Cleaner Production*, vol. 449, p. 140 789, 2024. DOI: 10.1016/j.jclepro.2024.140789.

[3] C. Ma, J. Song, M. Ran, Z. Wan, Y. Guo, and M. Gao, "Machine learning–driven spatiotemporal analysis of ozone exposure and health risks in china," *Journal of Geophysical Research: Atmospheres*, vol. 129, 2024. DOI: 10.1029/2024JD041593.

[4] P. Zhou, K. Chen, and Y. Sun, "Hybrid gcn–lstm approach for urban pm2.5 and no2 forecasting in complex terrains," *Atmospheric Environment*, vol. 312, p. 120 548, 2024. DOI: 10.1016/j.atmosenv.2024.120548.

[5] Z. Fang, J. Li, and M. Xu, "Deepair: A hybrid cnn-lstm model for fine-grained air quality prediction," *Sensors*, vol. 21, no. 5, p. 1581, 2021. DOI: 10.3390/s21051581.

[6] V. Balamurugan, J. Chen, A. Wenzel, and F. N. Keutsch, "Spatiotemporal modeling of air pollutant concentrations in germany using machine learning," *Atmospheric Chemistry and Physics*, vol. 23, no. 17, pp. 10 267–10 285, 2023. DOI: 10.5194/acp-23-10267-2023.

[7] W. Chen, N. Zhang, X. Bai, and X. Cao, "Research on the estimation of air pollution models with machine learning in urban sustainable development based on remote sensing," *Sustainability*, vol. 16, no. 24, p. 10 949, 2024. DOI: 10.3390/su162410949.

[8] Anonymous, "Transformer-based spatiotemporal modeling for air quality prediction," *arXiv preprint*, 2025, (arXiv).

[9] V.-D. Le, T.-C. Bui, and S. K. Cha, *Spatiotemporal deep learning model for citywide air pollution interpolation and prediction*, Preprint, 2019. DOI: 10.48550/arXiv.1911.12919. arXiv: 1911.12919 [cs.LG].

[10] Y. Zheng, F. Liu, and H.-P. Hsieh, "A spatiotemporal graph-based hybrid deep learning model for air quality prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 1, pp. 1–19, 2020. DOI: 10.1145/3360037.

[11] D. R. Kim, F. Luo, and J. Xu, "Spatial–temporal diffusion modeling of air pollution with graph neural networks and transfer learning," *Environmental Pollution*, vol. 336, p. 122 243, 2025. DOI: 10.1016/j.envpol.2025.122243.

[12] X. Liu, T. Zhao, and M. Wang, "Airformer: Attention-guided transformer for multi-pollutant spatiotemporal forecasting," *IEEE Transactions on Geoscience and Remote Sensing*, 2025. DOI: 10.1109/TGRS.2025.3376548.

[13] Anonymous, "A hybrid deep learning air pollution prediction approach based on ksc-convlstm (knn + spatio-temporal attention + residual convlstm)," *PMC Journal*, 2024, (PMC). DOI: 10.1038/s41598-025-88086-1.

[14] Anonymous, "Enhancing air pollution prediction: A neural transfer learning approach," *Science of the Total Environment*, 2024, (ScienceDirect). DOI: 10.1016/j.eti.2024.103793.

[15] Anonymous, "Enhanced air quality prediction using adaptive residual bi-lstm," *Scientific Reports*, 2025, (Nature).

[16] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction," *Water, Air, & Soil Pollution*, vol. 236, p. 464, 2025. DOI: 10.1007/s11270-025-08122-8.

[17] Anonymous, "Improving 3-day deterministic air pollution forecasts using ml (rf, xgb, lstm) in stockholm," *Atmospheric Chemistry and Physics*, 2024, (ACP).

[18] L. Deng, X. Jiang, and R. Yu, *Geoairnet: Graph–transformer hybrid model for regional air quality prediction*, Preprint, 2025. DOI: 10.48550/arXiv.2503.04561. arXiv: 2503.04561.

[19] J. Wu, L. Zhang, et al., "Transformer-based spatiotemporal model for air pollution forecasting using multi-source environmental data," *Environmental Modelling and Software*, vol. 180, 2025. DOI: 10.1016/j.envsoft.2025.106123.

[20] Anonymous, "Predicting air quality index using attention convolutional neural networks + arima + qpso-lstm + xgboost," *Journal of Big Data*, 2024, (SpringerOpen).

[21] C. Singh, P. Roy, and A. Das, "Explainable deep learning for air quality prediction with shap-based model interpretation," *Environmental Research*, vol. 249, p. 119 244, 2025. DOI: 10.1016/j.envres.2025.119244.

[22] H. Khan, J. Tso, N. Nguyen, N. Kaushal, A. Malhotra, and N. Rehman, *A novel approach for predicting the air quality index of megacities through attention-enhanced deep multitask spatiotemporal learning*, Preprint, 2024. DOI: 10.48550/arXiv.2407.11283. arXiv: 2407.11283 [cs.LG].

[23] Z. Yang, L. Lin, and Q. He, "Multi-resolution graph neural network for regional-scale air pollution forecasting," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 212, p. 105 628, 2024. DOI: 10.1016/j.isprsjprs.2024.105628.

[24] L. Zhang, Y. Wang, and X. Chen, "A hybrid deep learning air pollution prediction approach based on ksc-convlstm (knn + spatio-temporal attention + residual convlstm)," *PMC / journal*, 2024-2025, PMC.

[25] Anonymous, "Advanced air quality prediction using multimodal data and dynamic hybrid deep learning model," *Scientific Reports*, 2025, (Nature).

[26] A. Bigi, S. Damayanti, et al., "Big-data-driven machine learning for enhancing spatiotemporal air quality analysis," *Atmosphere*, vol. 14, no. 4, p. 760, 2023. DOI: 10.3390/atmos14040760.

[27] Y. Li, M. Zhang, G. Ma, H. Ren, and E. Yu, "Analysis of primary air pollutants' spatiotemporal distributions based on satellite imagery and machine-learning techniques," *Atmosphere*, vol. 15, no. 3, p. 287, 2024. DOI: 10.3390/atmos15030287.

[28] A. Patel, L. Chen, and D. Roberts, "Enhanced air quality prediction using adaptive residual bi-lstm," *Scientific Reports*, 2025, Nature.

[29] Q. Zhu, D. Lee, and O. A. Stoner, "A comparison of statistical and machine learning models for spatiotemporal air pollution prediction," *Environmental and Ecological Statistics*, vol. 31, no. 4, pp. 1085–1108, 2024. DOI: 10.1007/s10651-024-00635-5.

[30] Anonymous, "Spatiotemporal causal decoupling model for air quality forecasting (aircade)," *arXiv preprint*, 2025, (arXiv). DOI: 10.48550/arXiv.2505.20119.

[31] A. Baruah et al., "A novel spatiotemporal prediction approach to fill air pollution data gaps using mobile sensors, machine learning, and citizen science techniques," *npj Climate and Atmospheric Science*, vol. 7, no. 1, p. 310, 2024. DOI: 10.1038/s41612-024-00859-z.

[32] V.-D. Le, T.-C. Bui, and S. K. Cha, *Spatiotemporal graph convolutional recurrent neural network model for citywide air pollution forecasting*, Preprint, 2023. DOI: 10.48550/arXiv.2304.12630. arXiv: 2304.12630 [cs.LG].

[33] Anonymous, "Air pollution prediction based on optimized deep learning neural network (pso-lstm)," *Environmental Computational Modeling*, 2025, (ScienceDirect).

[34] S. Patel, R. Ghosh, and T. Li, "Cross-domain spatiotemporal transfer learning for pm2.5 estimation using satellite and ground-based data," *Remote Sensing of Environment*, vol. 310, p. 114 587, 2024. DOI: 10.1016/j.rse.2024.114587.

[35] T.-C. B. Le and S. K. Cha, *Spatiotemporal air quality mapping in urban areas using sparse sensor networks and machine learning*, Preprint, 2025. DOI: 10.48550/arXiv.2501.11270. arXiv: 2501.11270 [cs.LG].

[36] H. Li, R. Kumar, and P. Singh, "Predicting air quality index using attention convolutional neural networks + arima + qpso-lstm + xgboost," *Journal of Big Data*, 2024.

[37] B. He, N. Choudhury, and F. Wu, "A spatiotemporal deep ensemble for integrating satellite aod and meteorological data in pm2.5 estimation," *Science of the Total Environment*, vol. 937, p. 174 398, 2025. DOI: 10.1016/j.scitotenv.2025.174398.

[38] P. Otto et al., *Spatiotemporal modelling of pm2.5 concentrations in lombardy (italy) – a comparative study*, Preprint, 2023. arXiv: 2309.07285 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2309.07285.