Thesis for the Degree of Master of Science

# Global Air Pollutant Forecasting Using Sequential Transfer Learning: Addressing the Cold Start Problem

Department of Robot and Smart System Engineering
The Graduate School

Christopher Daniel Ash

June 2025

The Graduate School
Kyungpook National University

# Global Air Pollutant Forecasting Using Sequential Transfer Learning: Addressing the Cold Start Problem

Christopher Daniel Ash

Department of Robotics and Smart Systems Engineering
The Graduate School

Supervised by Professor Bubryur Kim

Approved as a qualified thesis of Christopher Daniel Ash
for the degree of Master of science
by the Evaluation Committee

June 2025

Chairman____________________________(인)

____________________________(인)

____________________________(인)

The Graduate School
Kyungpook National University

# Table of Contents

# Table

# Figures

# Chapter 1 Introduction

## 1.1 Background of the study

The findings from numerous international studies demonstrate that air pollution can have a wide range of severe health impacts [1]. Millions of people die each year due to these health effects, which range from respiratory illnesses such as lung cancer and asthma to cardiovascular diseases, including myocardial infarction and stroke [2]. A plethora of air pollutants, including sulfur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2), nitrogen monoxide (NO), and varying sizes of particulate matter (PM), have been identified as contributing to air pollution [3]. Hence, air pollution has been a popular topic of study throughout the years [4]. Accurately predicting air pollution can provide insights into environmental trends, which can help authorities and societies mitigate the effects of pollution [5]. In addition, it can help in developing effective preventive measures and management methods to lower air pollution [6]. Therefore, air pollution prediction plays a key role in maintaining public health.

There are several ways in which air pollution can be predicted. Three main forecasting categories include: numerical, statistical, and artificial-intelligence (AI) methods [7]. All of these methods rely on using past data to discover trends for predicting future values. Due to the ability of AI models to capture complex patterns and generate accurate forecasts, they have been extensively applied to various fields such as construction [8, 9], materials science [10], and particularly air pollution prediction [11, 12]. Within AI models, deep-learning (DL) techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), including long short-term memory (LSTM) models, and autoencoders have garnered a great deal of popularity [13, 14]. All of these methods rely on historical air pollution data, a form of time-series data. However, the cold-start problem, where limited or even no data is available, is a known challenge that is rarely addressed [15]. Typically, air pollution prediction studies rely on large amounts of regional and local data to make accurate predictions [16 – 19]. This is due to air pollution being affected by local factors such as traffic, factories, meteorological conditions, temporal, and other features [20 – 23].

The cold-start problem is therefore a significant obstacle faced in air pollution prediction studies [24 – 25]. This leaves regions without the necessary infrastructure to collect comprehensive air quality data at a disadvantage, thereby causing challenges to accurately predict air pollution and potentially posing a risk to public health in such areas.

## 1.2 Research objectives

This study seeks to explore the existence of a global pattern in air pollution data, particularly PM data, in order to address some of the challenges faced in air pollution studies. In a novel approach to air pollution prediction, an LSTM model is trained via transfer learning on data from seven major regions around the world to create a globally applicable air pollution prediction model. By incorporating these datasets, the model aims to provide globally applicable air pollution predictions, particularly benefiting regions with limited air pollution data availability. Through this study, the following contributions to the field of air pollution prediction are made:

(1) Development of an LSTM model trained via sequential transfer learning on datasets from major global regions that can potentially be used to predict air pollution anywhere in the world. This shifts the focus from region-specific models to a globally trained approach, ensuring broader applicability in air pollution forecasting.

(2) Ensuring data quality through comprehensive pre-processing and dataset characteristics analysis. Since DL models heavily rely on the training data, the quality of the data used directly affects the accuracy of the predictions. Therefore in this study, all datasets undergo thorough pre-processing and validation to maintain their natural statistical properties.

(3) Proposing a method to address the cold start problem in air pollution prediction. The developed model should not only be applicable across various regions around the world but also maintain accuracy in data-scarce conditions. To evaluate this, data scarcity is simulated on dedicated test datasets, and extensive testing is conducted to assess the model's performance and limitations.

(4) Comparative analysis of the effectiveness of the proposed TL method on other widely used deep learning models in air pollution forecasting. Specifically, the performance of the LSTM model is compared against the one-dimensional convolutional neural network (1D-CNN) and gated recurrent unit (GRU) models. Through this comparative analysis, not only is the effectiveness of the proposed

approach validated, but its applicability across different model architectures is also demonstrated.

In summary, this study presents an air pollution prediction model with increased generalizability and could benefit regions where data scarcity is a challenge, especially those that lack robust air quality monitoring infrastructure. The remainder of this paper is organized as follows: Chapter 2 provides an analysis of previous works and challenges faced in air pollution prediction. Chapter 3 outlines the dataset information and data pre-processing techniques used. Chapter 4 describes the applied methodology and includes the transfer learning model details and the methods used to validate the model's effectiveness. Chapter 5 evaluates the results of the study, including a comparative analysis of the effectiveness of the transfer learning method. Chapter 6 summarizes the findings of the study, discusses its limitations, and potential focuses for future research.

# Chapter 2 Review of Related Works

DL models have been widely adopted in recent air pollution prediction studies. This is due to their ability to capture representative patterns existing in the data allowing for accurate predictions [26]. Transfer learning has been shown to enhance the accuracy of air pollution predictions and is increasingly being adopted in recent studies as well [3, 27]. This section reviews advancements in DL for air pollution prediction, how transfer learning has provided benefits to this field of study, and some of the challenges that remain.

## 2.1 Overview of air pollution prediction

From the wide range of DL models that have been applied in recent air pollution forecasting studies, the models that have seen the most research attention include CNNs and RNNs, particularly LSTMs [13]. These are broad categories that encompass a large variety of models and applications. Furthermore, numerous combinations of these models, including hybrid and ensemble models, have been researched as well.

CNNs, known for their image processing and feature extraction abilities [14], are more commonly used in air pollution prediction studies alongside other models to enhance the overall predictive performance. An example of this is Guo et al. [28], where a CNN was used alongside a gated recurrent unit (GRU) to predict PM2.5 levels in indoor environments such as subways. GRUs (a type of RNN) are frequently used for their time-series prediction capability; for instance, Huang et al. [29] enhanced GRUs using additional techniques to predict hourly PM2.5 concentrations. GRUs have also been used alongside other models such as CNNs, as demonstrated by Jin et al. [30], where an ensemble model was used to predict PM2.5 levels in Beijing.

Due to their advantages, CNNs are also commonly used alongside LSTM models in air pollution prediction research; for instance, Yan et al. [31] compared the air quality forecasting performances of an LSTM, CNN, and CNN-LSTM hybrid model. Their findings indicate that the CNN model exhibited the worst performance, while the

CNN-LSTM model only slightly outperformed the LSTM model. The authors concluded that the LSTM model was the best choice for air pollution prediction. Another example where this combination was used is a study that developed a deep air quality forecasting framework using a CNN for spatial correlation extraction, and an LSTM for temporal dependency capture [23]. Due to its excellent prediction abilities, this hybrid model has been combined in various ways in numerous studies [32 – 36].

LSTM models (a type of RNN) have been extensively used in air pollution studies because of their capacity for capturing long-term temporal relationships [37]. Due to the strong predictive capabilities of LSTM models, Seng et al. [37] employed one as the sole predictor for air quality forecasting. It has also been incorporated in various hybrid models, as mentioned previously with CNN models. LSTM models have also been used with other models, such as a graph convolutional network (GCN), to forecast air pollutant concentrations with high accuracy [38, 39]. A modified version of the LSTM called LSTM Extended has been created and achieved promising results in predicting PM2.5 data [17]. Other modified or hybrid models that utilize the LSTM model for air pollution/ air quality forecasting include Xavier Reptile Switan-h-based (XRSTH)-LSTM) [40], ensemble empirical mode decomposition (EEMD) [18, 21], particle swarm optimization (PSO) [41], least-squares support vector regression (LSSVR) [42], and an autoencoder [43]. Moreover, other algorithms such as attention mechanisms have been used alongside LSTM models to improve the overall air pollution prediction performance [44].

Machine-learning models, and by extension DL models, require substantial amounts of data to make accurate predictions. Transfer learning is an approach used to solve this issue by transferring patterns captured from a data-abundant source domain to a data-scarce target domain [45]. This method has also been applied to air pollution prediction, and studies have generally shown that it enhances predictive accuracy [3, 46]. Recent applications within air pollution prediction studies have shown its ability to be utilized in regions with limited data, yielding positive results [47]. An example is the study conducted by Yadav et al. [48], which applied TL using a combination of ground data and satellite imagery to estimate air pollution in data-poor regions.

## 2.2 Challenges in air pollution prediction

DL studies are often constrained by data quality, a challenge that is frequently encountered in air pollution prediction studies [6, 7, 26, 32, 49]. DL models identify subtle, complex patterns in data, which enable them to generate accurate forecasts. Therefore, prediction accuracy is highly dependent on the quality of the data [36]. Srivastava et al. [40] found that the effectiveness of the prediction system could be enhanced by ensuring the quality of the data. This further emphasizes that high-quality and reliable data is required for accurate predictions.

Due to the reliance on data for accurate predictions, the issue of generalizability also arises in air pollution studies. Air pollution is known to be influenced by multiple local factors [50], which often leads to models being trained on region-specific data. This long-standing issue has been widely documented in a plethora of studies [6, 18 - 21, 24, 28, 35, 36, 47]. When data from only one region is used to train DL models, they can potentially only be used within that region, which leads to a lack of generalizability in most air pollution prediction models [7]. This issue has persisted in recent studies. For example, Jana et al. [51] developed a short-term prediction model using a GCN, which provided promising results. However, the authors only tested the model on two datasets and acknowledged the need for further validation to ensure generalizability. Similarly, Jairi et al. [27] applied TL across different air pollutants using a multi-layer perceptron model. Meanwhile, generalizability was stated to be a limitation as the model was trained on data from only one monitoring station.

Lack of generalizability also contributes to the cold-start problem, which arises when little or no data is available to train a model. Since DL models are typically trained on region-specific data, regions in which data is scarce face significant challenges in developing air pollution prediction models [24]. Previously, researchers have explored TL [48] and meta-learning [24] to address this issue, albeit with limited success. Yadav et al. [48] utilized TL with satellite image data to estimate pollution levels but did not conduct predictive modeling. Wu et al. [24] applied meta-learning using data from nearby regional monitoring stations. However, this approach may still limit the model's applicability to other regions or countries as the learned patterns would be localized and not transferable to other geographic contexts.

In summary, the literature highlights the LSTM model as one of the most widely adopted architectures in air pollution forecasting, with CNNs and GRUs also frequently utilized. Despite advancements, the field continues to face persistent challenges, particularly those related to data quality, generalizability, and the cold-start problem. Many existing models rely on region-specific data, which limits their applicability across different geographic contexts. Moreover, efforts to address data scarcity, such as transfer learning and meta-learning, have shown promise but remain limited in scope or effectiveness. In response to these challenges, the present study proposes a globally generalizable air pollution prediction model trained via sequential transfer learning. By leveraging high-quality, pre-processed datasets from multiple world regions, this approach aims to mitigate the cold-start problem while enhancing cross-regional generalizability.

# Chapter 3 Datasets and Data Preparation

This section details all of the datasets involved in this study, from the datasets used to train the model to the testing datasets. Also, the data pre-processing techniques used to ensure the quality of the data are explained and the resulting characteristics of each data set are shown. The section then concludes with an explanation of the segmentation strategy applied to simulate data scarcity.

## 3.1 Datasets

When air pollution prediction studies are done, usually only data from one region is used. DL models are known for their ability to discover patterns in data that are difficult to discover by humans. Therefore, in a novel approach to air pollution prediction, this study gathered data from several regions around the world to determine whether an LSTM model can identify any global patterns in air pollution data. This would allow the created model to be used anywhere in the world. Therefore to explore whether such a pattern exists, the globe was divided into seven major regions and at least one dataset was selected from each region.

For this study, open-source datasets were collected from various online databases. Datasets from countries with the greatest volume of available data in each region were selected to provide the model with as much data as possible for training. Most of these datasets contained various pollutants, however only the PM data was extracted. Particulate matter is one of the major air pollutants that affect air quality. It is usually the first pollutant that is measured anywhere air quality is a concern. This is seen in the global datasets as all contain at least particulate matter measurements. Therefore, in regions where air quality is a new focus, measuring PM would be prioritized before moving on to other pollutants. Accordingly, this study focuses on predicting PM values. Hourly data was selected since air pollution is known to vary throughout the day. Being able to predict this variance would give as much information as possible to stakeholders so that it can be effectively monitored and managed.

Table 1 outlines the information on the datasets used for training the model. It details the global regions and the selected representative country.

TABLE 1. Training datasets

| Region | Representative country | Date range | Amount of data/ Time steps (hours) |
|---|---|---|---|
| North America | US | 01/01/2013 – 31/08/2023 | 93481 |
| South America | Colombia | 01/01/2016 – 31/12/2021 | 52606 |
| Europe | UK | 01/01/2019 – 01/01/2024 | 43824 |
| Africa | Uganda | 21/11/2019 – 31/12/2020 | 9760 |
| South Asia | India | 01/01/2015 – 01/07/2020 | 48192 |
| East Asia | China | 02/01/2017 – 31/12/2020 | 35040 |
| Oceania | Australia | 01/01/2020 – 01/01/2024 | 35064 |

Three datasets were also selected from varying regions for testing the trained model. These datasets, while still falling within the larger regions of the study, are from different countries from the training datasets and were not used during model training. They were only used to test the model, showing its generalizability. Table 2 outlines the information for these testing datasets.

TABLE 2. Testing datasets

| Region | Representative country | Date range | Amount of data/ Time steps (hours) |
|---|---|---|---|
| East Asia | South Korea | 01/01/2013 – 31/08/2023 | 26280 |
| North America | Mexico | 01/01/2016 – 31/12/2021 | 47592 |
| Europe | Spain | 01/01/2015 – 31/12/2021 | 61368 |

## 3.2 Data pre-processing

The quality of data is known to be a crucial factor that influences the level of accuracy of the prediction of DL models. Consequently, efforts were made to ensure high data quality. The data pre-processing process consisted of two steps: data cleaning and data transformation.

### 3.2.1 Data cleaning

Due to the nature of data collection, some data points are inevitably missing or are outliers. To make the predictions by the DL model as accurate as possible, the outlier points need to be removed, and the missing data points replaced with values that are as close as possible to what they should be. This process is called data cleaning.

In this study, removing outliers was conducted using seasonal and trend decomposition using locally estimated scatterplot smoothing (STL decomposition using Loess). This process breaks a time series ($Y_t$) down into trend ($T_t$), seasonal ($S_t$), and residual ($R_t$) components [52], shown in equation (1).

$$Y_t = T_t + S_t + R_t \tag{1}$$

The $S_t$ and $T_t$ components were estimated using Loess, and the residual components were then analyzed for deviations from the norm. This method enables the precise identification of anomalies in the dataset while preserving any inherent patterns that exist in the time series.

Once the outliers had been removed, the missing values were added via a process called data imputation. For this application, k-nearest neighbors (KNN) was found to be one of the best methods [53], therefore, this method of imputation was selected. This method imputes missing values by analyzing the k nearest data points in the feature space to estimate missing values via the weighted or unweighted average of the neighboring corresponding values. The formula for imputing , the missing data point in row i and column j is shown in equation (2).

$$x_{i,j} = \frac{\sum_{k \in N(i)} \omega_k x_{k,j}}{\sum_{k \in N(i)} \omega_k} \tag{2}$$

Where *N(i)* denotes the set of k-nearest neighbors of row *i*, $\omega_k$ represents the weight assigned to each neighbor *k*, and $x_{k,j}$ represents the feature of *j* in neighbor *k*. The imputation ensured that missing values were replaced in a manner consistent with the existing patterns in the data, thereby preserving its overall statistical properties.

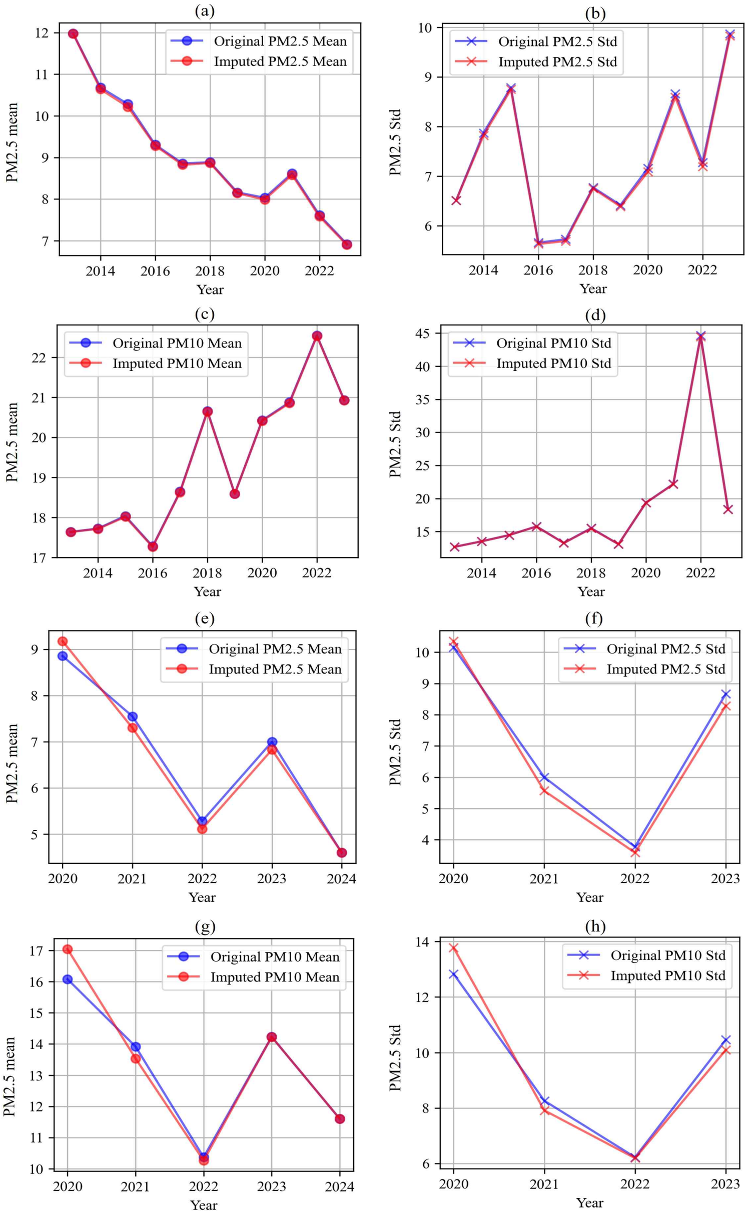


Figure 1. Dataset characteristics comparison before and after imputation: (a) US Dataset particulate matter < 2.5 µm (PM2.5) mean comparison, (b) US Dataset PM2.5 Std comparison, (c) US Dataset particulate matter < 10 µm (PM10) mean comparison, (d) US Dataset PM10 Std comparison, (e) Australia Dataset PM2.5 mean comparison, (f) Australia Dataset PM2.5 Std comparison, (g) Australia Dataset PM10 mean comparison and, (h) Australia Dataset PM10 Std comparison.

Figure 1 shows a comparison of the change in characteristics, mean and standard deviation (Std), of some of the datasets before and after imputation, while Table 3 reports the complete results of the data transformation. In particular, the US (Figure 1 (a-d)) and Australia (Figure 1 (e-h)) datasets show only minor differences between the original and imputed values, indicating that the imputation process preserved the statistical properties of the data. As reported in Table 3, the mean and standard deviation changes were below 2.5% across all datasets. This minimal deviation confirms that the data cleaning was effective in preserving the underlying data characteristics while addressing missing values and outliers.

TABLE 3. Data cleaning results summary

| Country | Change in PM2.5 mean (%) | Change in PM2.5 Std (%) | Change in PM10 mean (%) | Change in PM10 Std (%) |
|---|---|---|---|---|
| US | 0.428 | 0.578 | 0.061 | 0.167 |
| Colombia | 0.732 | 1.579 | 0.465 | 2.281 |
| UK | 0.056 | 0.238 | 0.080 | 0.321 |
| Uganda | 1.196 | 1.164 | 1.382 | 0.972 |
| China | 0.449 | 1.143 | 0.314 | 0.799 |
| India | 0.460 | 0.149 | 0.839 | 0.513 |
| Australia | 0.430 | 1.747 | 1.310 | 2.486 |
| South Korea | 0.071 | 0.571 | 0.112 | 0.572 |
| Mexico | 2.185 | 0.982 | 0.800 | 2.330 |
| Spain | 0.127 | 1.638 | 0.081 | 1.346 |

### 3.2.2 Data transformation

To ensure proper training of the LSTM model, data transformation was conducted on the cleaned data. Within this step, filtering using a Butterworth low-pass filter and normalization were carried out. First, the data was sent through a Butterworth low-pass filter to apply smoothing by removing high-frequency noise, such as missed outliers, while retaining the low-frequency components encompassing meaningful data trends. This filter was selected for its ability to minimize any distortions within the regular data points and only focus on attenuating higher frequencies usually associated with noise or variations irrelevant to the overall data. This step assisted in ensuring that the data was of high quality. The mathematical transfer function for the Butterworth filter is expressed as equation (3).

$$H(\omega)=\frac{1}{\sqrt{1+\left(\frac{\omega}{\omega_c}\right)^{2n}}} \tag{3}$$

Where $\omega$ represents the signal frequency obtained from the input data, $\omega_c$ represents the cut-off frequency used to determine which values are noise, and $n$ is the filter order that affects the sharpness of the cut-off. The output of the Butterworth filter was clipped to 0 to ensure that the data maintained its realism, since PM data cannot be negative.

The following step was data normalization. This scales the data into a uniform range between 0 and 1 to eliminate the effects that differing magnitudes and units can have on model training. This process ensures that the LSTM model treats all input features with equal importance, thereby preventing larger values and ranges from skewing the learning process. The normalization process, mathematically represented in equation (4) is:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where $x$ is the original value and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the dataset, respectively.

By combining filtering and normalization, the data transformation step improved the data’s usability and quality. This ensured that the LSTM model could focus on determining the intrinsic patterns inherent in the data without being affected by noise or scale inconsistencies. These processes contributed significantly to the model's robustness, accuracy, and generalizability with new datasets.

## 3.3 Data segmentation

In this study, the data segmentation process was important for effective learning of the LSTM model. The pre-processed data was partitioned into three subsets: training, validation, and testing. The training subset was used by the LSTM model to learn the temporal patterns and relationships within the data. The validation subset was used during the training process for hyperparameter tuning and to help prevent overfitting. The testing subset was used to ensure that the model could correctly identify the patterns inherent in unseen data.

Each of the seven training datasets was segmented according to the following data split: 60% for training, 20% for validation, and 20% for testing. This data split enabled proper training while still providing enough data for validation and testing so that the training process could accurately capture patterns from each dataset.

### 3.3.1 Data scarcity simulations

The three testing datasets were segmented to evaluate the model's performance, specifically under cold-start conditions. In Wu et al [24], data-scarce conditions were roughly described as having days to weeks of data. This is supported by the observation that most air pollution studies use datasets spanning several years [5]. Similarly, the datasets used in this study span a minimum of three years (Tables 1 and 2), except for the African dataset, which originates from a region with limited infrastructure to measure air pollution [48]. Consequently, the span of 3.5 days (0.5 weeks) to 3 months (90 days) was selected as a suitable range to realistically simulate varying degrees of data scarcity.

For model evaluation, each testing dataset was segmented according to the selected data scarcity periods. Each data scarce simulated subset was divided using an 80:20 ratio for training and validation. The remaining portions of data served as the testing subset to assess the proposed model's predictive abilities. Table 4 summarizes the data splits utilized in this study. Given that each dataset varies in size, each data scarcity period produces a slightly different data split. This extensive level of data segmentation was carried out to ensure that the proposed model is viable in cold-start scenarios. By incorporating this segmentation strategy, this study ensures that the model is robust, adaptable, and capable of accurate air pollution predictions in data-scarce environments.

TABLE 4. Data-scarcity simulation dataset segmentation

| Dataset | Data scarcity period | Percentage of dataset used for training and validation (%) | Percentage of dataset used for Testing (%) | Size of testing dataset |
|---|---|---|---|---|
| South Korea | 0.5 weeks (84 hours) | 0.3196 | 99.6804 | 26196 |
| | 1 week (168 hours) | 0.6393 | 99.3607 | 26112 |
| | 2 weeks (336 hours) | 1.2785 | 98.7215 | 25944 |
| | 1 month/ 30 days (720 hours) | 2.7397 | 97.2603 | 25560 |
| | 2 months (1440 hours) | 5.4795 | 94.5205 | 24840 |
| | 3 months (2160 hours) | 8.2192 | 91.7808 | 24120 |
| Mexico | 0.5 weeks (84 hours) | 0.1765 | 99.8235 | 47508 |
| | 1 week (168 hours) | 0.3530 | 99.647 | 47424 |
| | 2 weeks (336 hours) | 0.7060 | 99.294 | 47256 |
| | 1 month/ 30 days (720 hours) | 1.5129 | 98.4871 | 46872 |
| | 2 months (1440 hours) | 3.0257 | 96.9743 | 46152 |
| | 3 months (2160 hours) | 4.5386 | 95.4614 | 45432 |
| Spain | 0.5 weeks (84 hours) | 0.1369 | 99.8631 | 61284 |
| | 1 week (168 hours) | 0.2738 | 99.7262 | 61200 |
| | 2 weeks (336 hours) | 0.5475 | 99.4525 | 61032 |
| | 1 month/ 30 days (720 hours) | 1.1733 | 98.8267 | 60648 |
| | 2 months (1440 hours) | 2.3465 | 97.6535 | 59928 |
| | 3 months (2160 hours) | 3.5197 | 96.4803 | 59208 |

# Chapter 4 Models and Methodology

Figure 2 shows the overall process, including data preparation, model training and testing, and finally the evaluation of the model's performance. The data pre-processing method was described in the previous section. The processed training datasets were used to train an LSTM (DL) model, with sequential training carried out via transfer learning across each dataset.  The testing datasets were used to test the fully trained model under various cold-start scenarios. Finally, performance metrics were used to analyze the performance of the fully trained model and compare its performance against a baseline LSTM model.

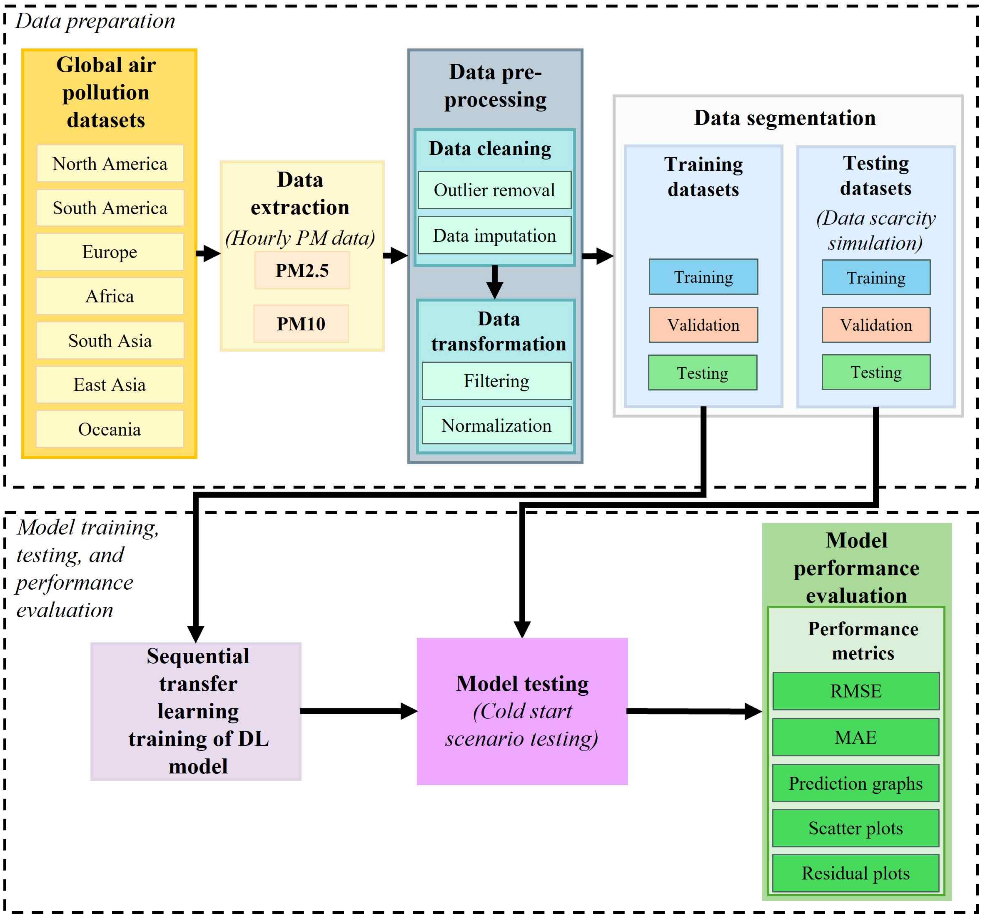


Figure 2. Overview of methodology.

## 4.1 Deep learning models

As discussed previously, the LSTM model is a particularly popular model in air pollution studies, which is why it was chosen as the main model for this study. Other popular models include the GRU and the

CNN which were used during the comparative analysis section to showcase the sequential transfer learning method's effectiveness.

### 4.1.1 LSTM model

The LSTM model, serving as the main model in this study, has undergone several transformations aimed at increasing its performance since its creation [54]. As a specialized form of RNN, this model is designed to effectively capture and learn sequential dependencies, particularly in time-series data. LSTM models are known for their ability to learn long-term temporal patterns, thereby making them well-suited for air pollution prediction. Figure 3 shows the architectural block diagram of an LSTM model, while equations (5)–(10) provide the mathematical formulation of the model.

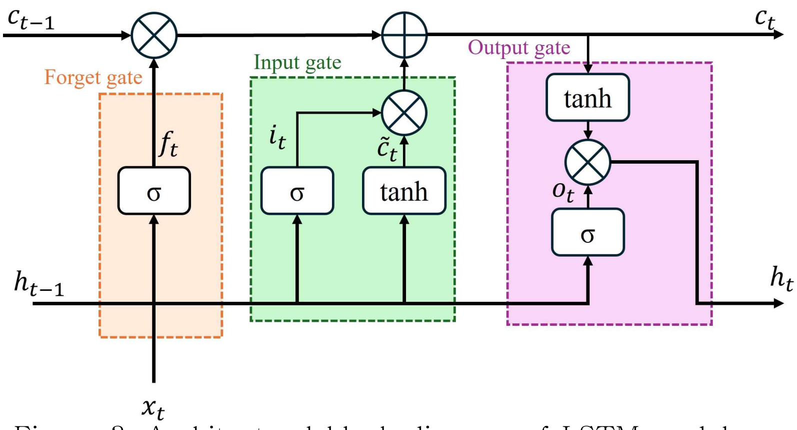


Figure 3. Architectural block diagram of LSTM model.

*Forget gate:*

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{5}$$

*Input gate:*

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{6}$$

*Candidate cell state:*

$$\tilde{c} = \tanh(W_c \bullet [h_{t-1}, x_t] + b_i) \tag{7}$$

*Cell state*

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{8}$$

*update:*

*Output gate:*

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

*Hidden state update:*

$$h_t = o_t * \tanh(c_t) \tag{10}$$

Where $f_t$ is the forget gate activation that determines which information is retained by the model, it is the input gate activation that sets the amount of added information, $\tilde{c}$ is the candidate cell state, and $c_t$ is the cell state used to determine the final output from the model. The variable $o_t$ represents the value of the output gate, $x_t$ is the input vector of the model and ht is the hidden state and output from the LSTM model. All of the $W$ variables with different subscripts represent the respective weight matrices for each gate, and the $b$ variables with different subscripts are the respective biases. These mechanisms enable the LSTM model to retain the essential temporal features of a time series while discarding irrelevant information.

Specifically, the forget gate removes the influence of past-learned PM information while making predictions. This helps improve prediction accuracy by reducing the influence of PM data no longer relevant to the current prediction. The input gate is responsible for determining the number of current PM values added to the cell state, which can help with filtering out noise such as short-term fluctuations, while retaining the values that significantly influence the current forecast. The output gate controls which information from the cell state is maintained as the hidden state. In the present study, this gate regulated how much of the discovered temporal patterns in the PM2.5

and PM10 data influence the model's predictions. These gates work together to enable the LSTM model to reliably forecast PM levels.

### 4.1.2. 1D-CNN model

One-dimensional CNN models have been applied in various ways in air pollution forecasting [23, 55]. While originally developed for image processing applications, 1D-CNNs can be effectively adapted for time-series prediction. Figure 4 showcases this model's architecture and equations (11) – (15) supply the mathematical formulation for the model.

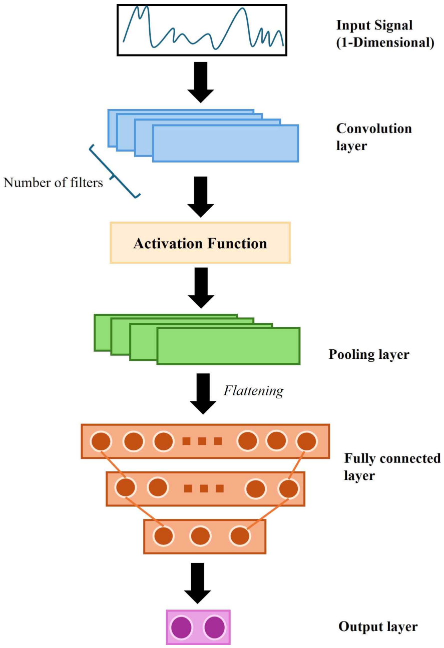


Figure 4. Architecture of 1D-CNN model.

*1D Convolution Layer:*

$$z_t^{(k)} = \sum_{i=0}^{K-1} \sum_{j=1}^{D} w_{i,j}^{(k)} \cdot x_{j,t+i} + b^{(k)} \tag{11}$$

*Activation Layer:*

$$a_t^k = ReLU\left(z_t^{(k)}\right) \tag{12}$$

*Pooling layer (Max pooling):*

$$p_j^k = \max\left\{a_{j\cdot S}^{(k)}, a_{j\cdot S+1}^{(k)}, \cdots, a_{j\cdot S+P-1}^{(k)}\right\} \tag{13}$$

*Flattening layer:* $$f = Flatten\left(\tilde{p}^{(1)}\ ,\tilde{p}^{(2)},\ \ldots\ ,\tilde{p}^{(F)}\right) \in R^{B\times(F\cdot L)} \tag{14}$$

*Fully connected layer:* $$\hat{y} = W\ \cdot f + b \tag{15}$$

In these equations, *k* denotes the index of the convolutional filter from a total of *F* filters, $z_t^{(k)}$ is the output of the convolution operation at time step t for the *k*-th filter. The indices *i* and *j* represent the filter index and feature/ channel index respectively, while *K* is the kernel size and *D* is the number of input features. $x_{j,t+i}$ represents the input signal at feature *j* at time index *t+i*, $w_{i,j}^{(k)}$ is the weight of the *i*-th position and *j*-th input feature for the *k*-th filter and $b^{(k)}$ is the bias term associated with the k-th filter. The output of the activation layer, $a_t^{(k)}$ is obtained by applying a ReLU function to $z_t^{(k)}$. In the pooling layer, max-pooling is applied such that $p_j^k$ denotes the pooled output at position *j* for filter *k*, P is the pooling window size and *S* is the stride. For the flattening operation, the pooled outputs are flattened into a feature vector *f* , where *F* is the number of filters, *L* is the length of each filter's pooled output, and *B* is the batch size. The final model output  is computed in the fully connected layer, where *W*  is the weight matrix, and b is the bias vector.

The 1D convolution layer learns the temporal patterns from the input PM data. Each filter in the layer slides over the input sequence created from the input data producing feature maps which capture relevant air pollution trends. The pooling layer then downsamples the resulting feature maps obtained, retaining the most important features pertinent to the discovered patterns. The activation layer applies the ReLU function, introducing non-linearity to the model, which enables it to learn complex patterns present in the PM data. The flattening layer converts the 3D output from the previous layers into a 2D format, making it suitable for input into the fully connected layer. Finally, the fully connected layer maps the extracted features to the output space through a linear transformation, generating the predicted PM values.

### 4.1.3. GRU model

The GRU unit has also been applied in various air pollution studies due to its efficiency and competitive predictive accuracy [55, 56]. GRUs are a simplified variant of LSTM networks that combine the forget and input gates into a single update gate, reducing computational complexity while maintaining effectiveness. Figure 5 illustrates the model architecture for the GRU model. Equations (16) – (19) provide the mathematical formulation of the model.

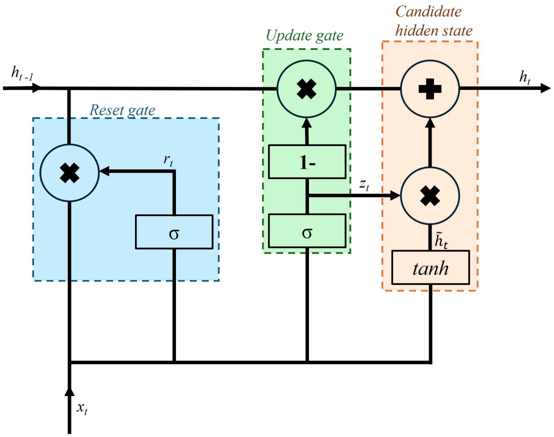


Figure 5. Architectural block diagram of GRU model.

*Update gate:* $$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \qquad (16)$$

*Reset gate:* $$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \qquad (17)$$

*Candidate hidden state:* $$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{18}$$

*Final hidden state:* $$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{19}$$

Where $x_t$ is the input vector at time step $t$, $h_t$ is the hidden state and $z_t$ is the update gate output. The variable $r_t$ represents the reset gate output, $\tilde{h}_t$ is the candidate hidden state. $W$ are the weight matrices for input, $U$ are the weight matrices for the previous hidden state, and $b$ are the bias vectors for each gate. The GRU model updates its hidden state based on these components, enabling the network to retain relevant temporal dependencies while discarding redundant or irrelevant past information.

Specifically, the reset gate controls how much of the previous hidden state is forgotten when calculating the candidate hidden state, enabling the model to discard irrelevant past information and concentrate on recent trends in the input PM data. The update gate determines how much of the previous hidden state is retained and how much of the new candidate hidden state is incorporated, enabling the model to preserve long-term dependencies present in the air pollution data when necessary. The candidate hidden state is responsible for producing the final hidden state which summarizes the captured temporal patterns learned from the input sequence of PM data. This final hidden state is then passed to a fully connected layer to generate the predicted PM values.

## 4.2. Sequential transfer learning method

Transfer learning enables models to take advantage of data from source datasets and apply it to target datasets containing less data [45]. This is accomplished by training a model on an initial data-rich dataset and subsequently utilizing the model along with the learned weights to generate predictions on a data-poor dataset. In this study, the TL process was repeated on each DL model for each dataset in sequence, thus the name sequential transfer learning. The model was trained on each training dataset while keeping the weights from each round of training, starting with the North America dataset, followed by the South American dataset, the European dataset, and so on, until the final dataset, Oceania. The fully trained TL model was formed once the DL model had been trained on all of the datasets. This model's weights now encompasses all of the learned information from regions throughout the world.

As the model is trained on subsequent datasets, TL enables the model to leverage the knowledge learned, thereby enhancing the final model's predictive performance. This approach was chosen as opposed to training the model on all of the datasets simultaneously to allow it to capture patterns that exist even in smaller datasets and prevent larger datasets from completely dominating the model weights, thereby causing overfitting. Progressively adding datasets from various regions reduced the model's dependence on any one specific dataset to predict air pollution trends. Instead, the model gradually learns generalizable global patterns. This structured approach helped to prevent the model from overfitting to any singular region while assisting the model to adapt to diverse PM characteristics worldwide. Figure 6 provides a visual overview of the sequential transfer learning process using the

LSTM model as an example. The datasets used in each stage are depicted on the left, while the central blocks represent successive iterations of the model, each incorporating the learned weights from the preceding dataset. Algorithm 1 explains how the entire process was implemented.

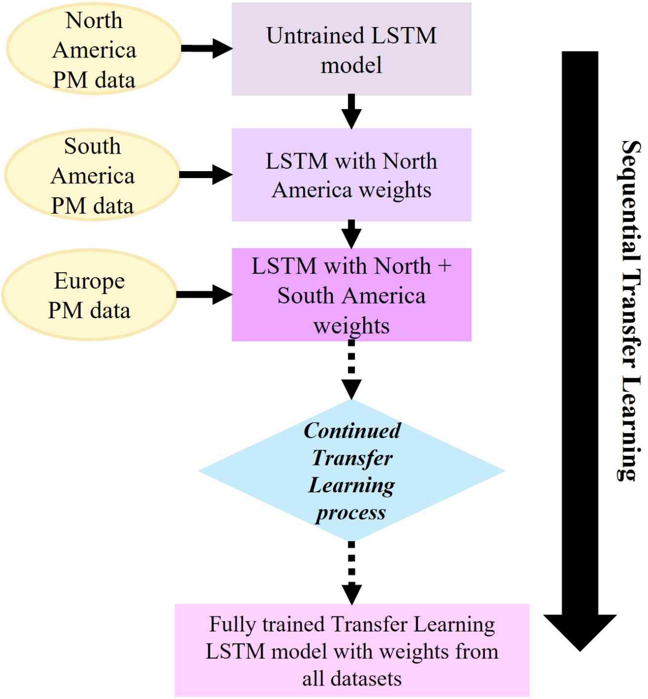

North America PM data
Untrained LSTM model
South America PM data
LSTM with North America weights
Europe PM data
LSTM with North + South America weights
Continued Transfer Learning process
Fully trained Transfer Learning LSTM model with weights from all datasets
Sequential Transfer Learning

Figure 6. Sequential transfer learning process.

Algorithm 1

```
Sequential transfer learning algorithm:
1. Let Ri represent the training datasets and Ti represent the testing datasets
2. Clean all datasets
   a. Apply STL to remove outliers from the dataset
   b. Apply KNN to impute missing data values
   c. Save all cleaned datasets
3. Define the DL model
   a. Initialize the model with specified weights and parameters
4. Load training dataset Ri
5. Apply data transformation and segmentation to dataset Ri
   a. Apply the filter to dataset Ri
   b. Normalize dataset Ri
   c. Split the dataset into training, validation, and testing subsets using a 6:2:2 ratio
6. Train the model
   a. Load the pre-trained weights from the previous training session (if available)
   b. Train the model using the training subset from dataset Ri
   c. Validate the model using the validation subset from dataset Ri
   d. Test the model using the testing subset from dataset Ri to check for erroneous performance
   e. Save the updated weights
7. Repeat steps 4-6 until the model has been trained on all Ri datasets
8. Evaluate model performance using testing datasets Ti
```

The North American dataset was selected as the starting point for the sequential transfer learning process since it is the largest dataset that spans the longest period of time. This allowed the model to learn foundational patterns from a large dataset before adapting to regional variations. The following regions were added based on their proximity, moving roughly from west to east across the globe. This was intended to help the model capture any similar patterns that may exist between nearby regions, thereby enabling incremental discovery of regional trends while still maintaining global patterns. Table 5 reports the order in which each dataset was added to the model.

TABLE 5. Order of addition of datasets during sequential training

| Dataset | Order of addition |
|---|---|
| North America | 1st |
| South America | 2nd |
| Europe | 3rd |
| Africa | 4th |
| South Asia | 5th |
| East Asia | 6th |
| Oceania | 7th |

This method of training was selected to encourage model generalizability, thus enabling the model to be used globally. Prior studies have shown that regional transfer learning increases prediction accuracy [3, 46]. In addition, training the model on such large and diverse datasets potentially enhances its applicability to regions with limited PM data by leveraging the global patterns discovered.

## 4.3. Model parameters

In this study, Pytorch 2.3.0 was used to implement each DL model via its dedicated library. Both PM2.5 and PM10 were used as input features, so the input and output dimensions were set to 2. These parameter values were selected since predicting PM levels is the focus of this study. The batch size was set to 32, and the sequence size was selected as 24 to enable the model to pick up patterns from the last 24 hours of data. These parameter values produced sufficient model performance hence their selection.

Random search was implemented to help identify the optimal hyperparameters for each model. For testing, fifty trials were conducted to explore various combinations of hyperparameters randomly selected from a set range. A random seed was set to ensure reproducibility across the trials. The same seed was subsequently used during model training and testing to maintain consistency across experiments.

The validation loss was evaluated using the mean square error and an Adam optimizer was utilized for fine-tuning the model parameters. During each trial, the validation was measured after five epochs to quickly evaluate each model and determine which hyperparameters would potentially yield the best results. Multiple parameter ranges were tested and it was found during testing that less complex models generally yielded better results. Test “a” evaluated more complex models that were slightly larger and utilized weight decay and dropout. Test “b” was assessed less complex models that were generally smaller and did not use weight decay or dropout. Tables 6 - 8 report the parameter ranges and optimal parameters obtained from each test.

TABLE 6. Random search test ranges and optimal hyperparameters for LSTM model

| Test | Hyperparameter | Hyperparameter range | Optimal value | Best validation loss | Mean validation loss |
|---|---|---|---|---|---|
| a | Hidden size | 50 - 200 | 119 | 0.00038032 | 0.002709 |
| | Number of layers | 2 - 4 | 2 | | |
| | Weight decay | 0.0001 - 0.01 | 0.00003 | | |
| | Dropout | 0.1 - 0.5 | 0.19 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00797 | | |
| b | Hidden size | 50 - 120 | 94 | 6.80547e-06 | 0.000023 |
| | Number of layers | 1 - 2 | 1 | | |
| | Weight decay | 0 | 0 | | |
| | Dropout | 0 | 0 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00161 | | |

TABLE 7. Random search test ranges and optimal hyperparameters for 1D-CNN model

| Test | Hyperparameter | Hyperparameter range | Optimal value | Best validation loss | Mean validation loss |
|---|---|---|---|---|---|
| a | Kernel size | 3 - 6 | 4 | 0.00038056 | 0.001089 |
| | Number of filters | 32 - 128 | 68 | | |
| | Weight decay | 0.0001 - 0.01 | 0.00001 | | |
| | Dropout | 0.0 - 0.5 | 0.35 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00262 | | |
| b | Kernel size | 1 - 3 | 1 | 6.86935e-05 | 6.86935e-05 |
| | Number of filters | 8 - 32 | 22 | | |
| | Weight decay | 0 | 0 | | |
| | Dropout | 0 | 0 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00765 | | |

TABLE 8. Random search test ranges and optimal hyperparameters for GRU model

| Test | Hyperparameter | Hyperparameter range | Optimal value | Best validation loss | Mean validation loss |
|---|---|---|---|---|---|
| a | Hidden size | 50 - 200 | 125 | 2.06077e-05 | 0.001386 |
| | Number of layers | 2 - 4 | 2 | | |
| | Weight decay | 0.0001 - 0.01 | 0.00065 | | |
| | Dropout | 0.1 - 0.5 | 0.18 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00036 | | |
| b | Hidden size | 50 - 120 | 85 | 9.79405e-06 | 0.000187 |
| | Number of layers | 1 - 2 | 1 | | |
| | Weight decay | 0 | 0 | | |
| | Dropout | 0 | 0 | | |
| | Learning rate | 0.00001 - 0.001 | 0.00187 | | |

Comparing test “a” to test “b”, the results show that across all models the simpler models, on average, produced better results than the more complex models, hence the optimal hyperparameters from test “b” were chosen as the model parameters. The simpler models also allowed for reduced training time compared to the more complex models, which would help with real-world applications. Analysis of validation loss during training and testing indicated that it generally stabilized after approximately 50 epochs. Therefore, this epoch count was selected as it allowed the models to achieve acceptable prediction accuracy while mitigating the risk of overfitting.

## 4.4. Performance evaluation metrics

Evaluation metrics are needed to validate a model’s performance and demonstrate that it is capable of generating accurate predictions. In this study, the chosen metrics were root mean square error (RMSE) and mean average error (MAE), which have been used in the majority of air pollution studies to showcase a model's performance [13]. Prediction graphs were selected for this study to clearly show the performance capability of the created model.

RMSE is commonly utilized as an evaluation metric in prediction studies. This metric calculates the average overall error of the model by taking the square root of the squared differences between the actual value and the prediction. This method of evaluation punishes larger errors, thereby making it particularly robust in detecting outliers. The lower the RMSE, the better the model's predictive ability. Equation (20) shows how RMSE is calculated.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n} \left(y_i - \hat{y_i}\right)^2} \tag{20}$$

Where $n$ represents the total number of values in the dataset, $y_i$ denotes the actual value of the data, and $\hat{y}$ is the model's predicted value.

MAE, as implied by its name, measures the average of the absolute differences between the predicted and actual values. Since this metric does not square errors, it is not sensitive to outliers, but provides a balanced view of how the model performs. Therefore, this metric can be used to complement the RMSE metric to provide another perspective on the model's predictive accuracy. The formula for MAE is illustrated in equation (21).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y_i} \right| \tag{21}$$

Prediction graphs, utilized in this study to provide a visual representation of the comparison of the models, are plots of the predicted values against the actual values of PM data over time. This was carried out to move away from just a numerical showcase of the model's performance and highlight the performance differences between the transfer learning model and the baseline model. The prediction graphs clearly showcase the discrepancies between the base model and the transfer learning model, offering a clear visual assessment of the model's ability to address the cold-start problem.

In addition to the prediction graphs, scatter plots were also utilized to visually compare model performance. This graph plots the actual values (ideal predictions) on the x-axis against predicted values on the y-axis, alongside an ideal prediction reference line (y =x). The closer the model's predicted values are to the actual values, the closer the plotted points will be to the ideal line. However, scatter plots alone inadequately depict temporal variations in a model's performance. To address this limitation, residual plots were employed. These plots illustrate temporal deviations between actual and model predicted values (residuals). This visualization enables the identification of systematic biases and performance patterns over time facilitating the comparative evaluation of the seasonal performance of the models.

# Chapter 5 Results and Discussion

After fully training the LSTM model, tests were conducted to evaluate its ability to generate predictions accurately, particularly in scenarios with minimal training data, determining its ability to address the cold-start problem. This was accomplished using data-scarcity simulations on the testing datasets. The results using the proposed model were assessed against those of a baseline LSTM model. A comparative analysis of the sequential transfer learning method was then conducted to evaluate its effectiveness across different models.

## 5.1 Fully trained sequential transfer learning model results

The performance of the fully trained LSTM model was evaluated against the base model using RMSE and MAE. Figure 7 displays the prediction graphs for each dataset under the most severe data scarcity conditions (0.5 weeks), illustrating the performance gap between the base and TL models. These results demonstrate the extent to which the sequential transfer learning process enhances the ability of the LSTM to track air pollution trends with minimal training data.

Figure 8 presents prediction scatter plots of each dataset under the least data-scarce conditions tested. This figure further highlights the performance differences between the TL and base models, providing a comprehensive assessment across both extremes of the data scarcity spectrum. Figure 9 shows the residual plots under the same minimal scarcity conditions. These plots are all centered around zero, indicating no apparent systematic bias in the predictions of either model. Detailed analysis of Figures 7 – 9 is provided in the subsequent sections.

The remaining results are divided based on the three testing datasets (Korea, Mexico and Spain). For each section, a table is provided to show a side-by-side comparison of the RMSE and MAE for the fully trained model against the base LSTM model. Through analysis of these results, the performance of the fully trained model was assessed, and its viability for handling cold-start scenarios was determined.

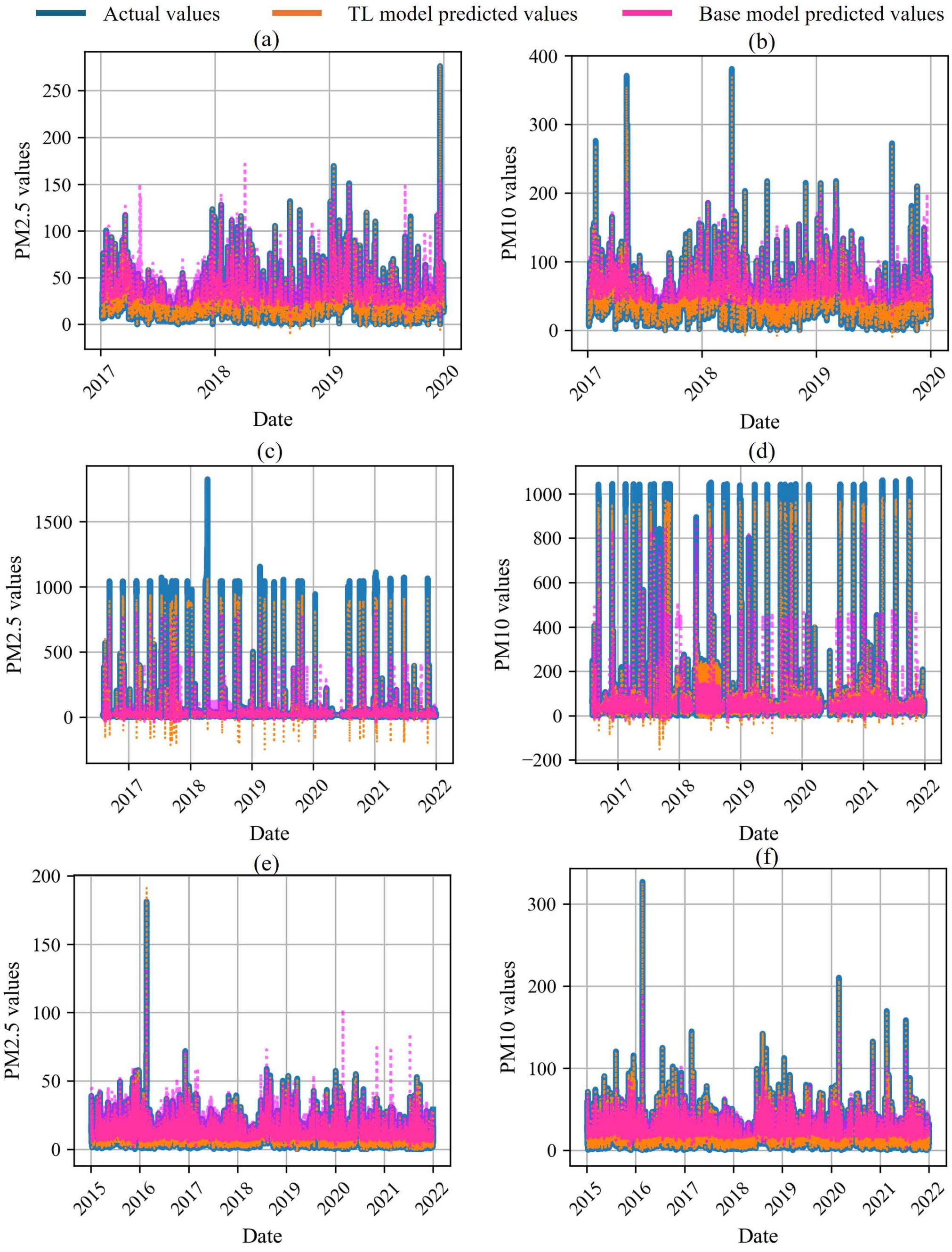


Figure 7. Prediction graphs under 0.5 weeks data scarcity level: (a) Korean dataset PM2.5 predictions, (b) Korean dataset PM10 predictions, (c) Mexican dataset PM2.5 predictions, (d) Mexican dataset PM10 predictions, (e) Spanish dataset PM2.5 predictions, and (f) Spanish dataset PM10 predictions

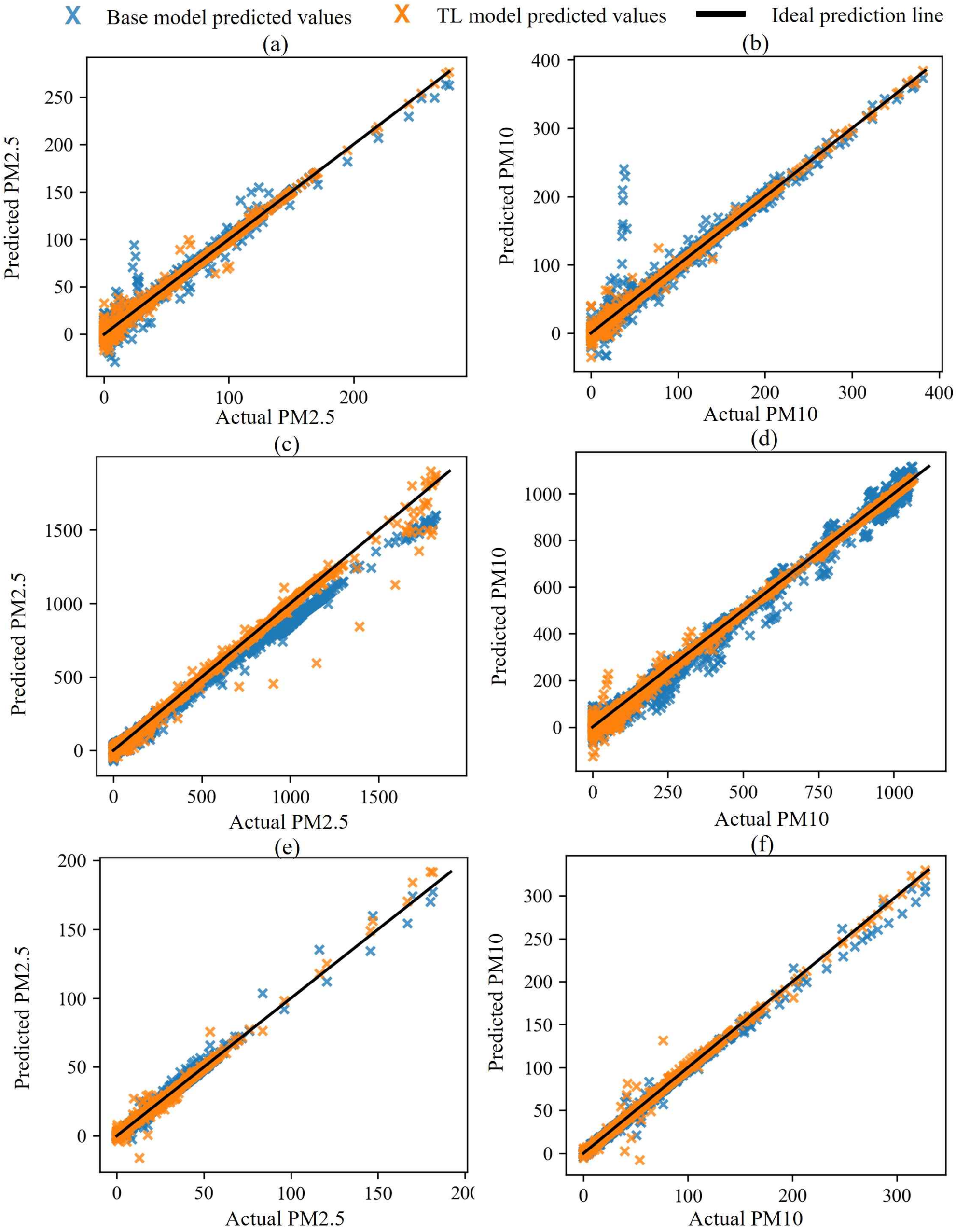


Figure 8. Comparison prediction scatter plots of TL model vs base model at 3 months data scarcity level: (a) Korean dataset PM2.5 predictions, (b) Korean dataset PM10 predictions, (c) Mexican dataset PM2.5 predictions, (d) Mexican dataset PM10 predictions, (e) Spanish dataset PM2.5 predictions and (f) Spanish dataset PM10 predictions

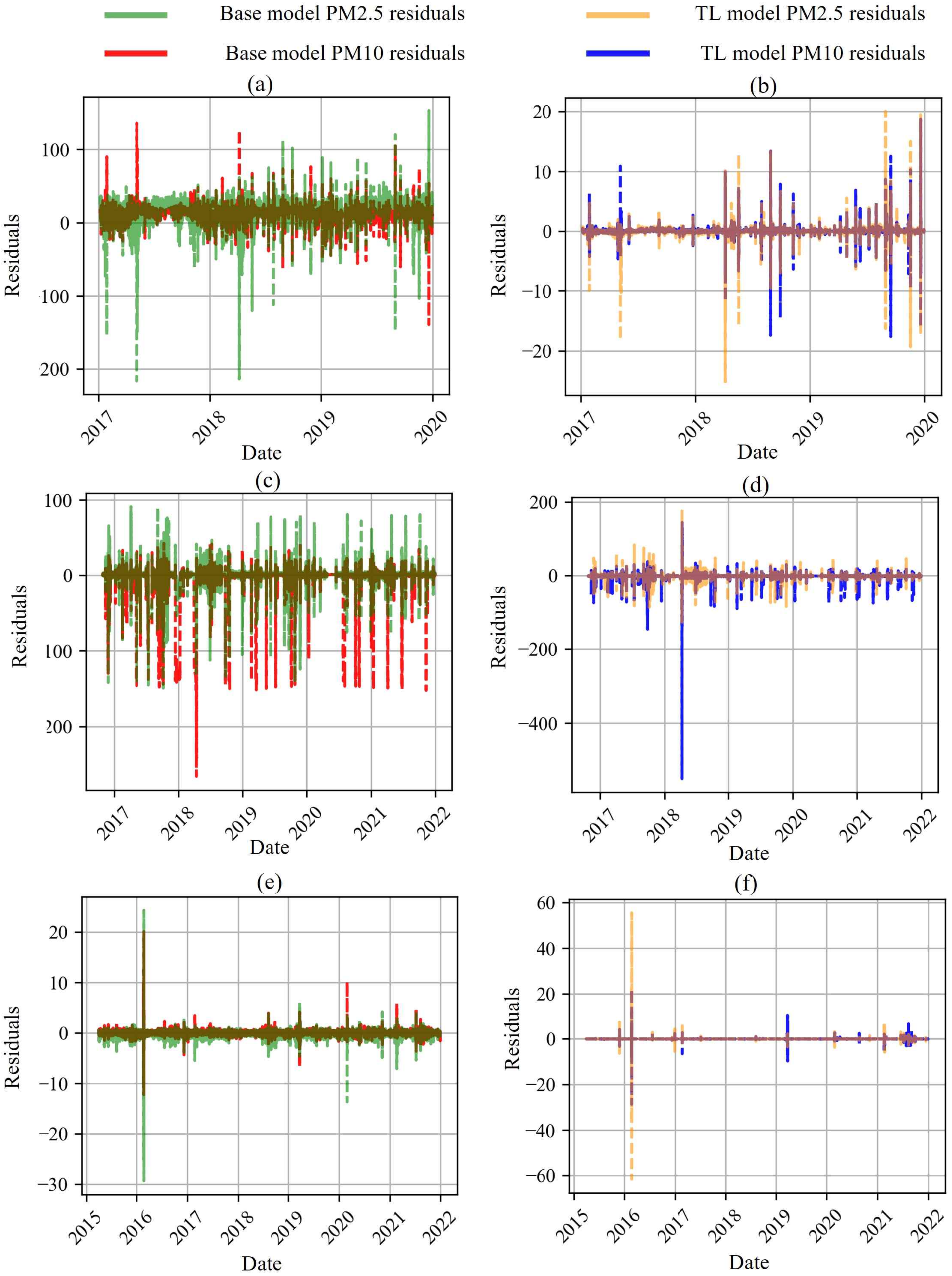


Figure 9. Residual plots of the TL and base model at 3 months data scarcity level: (a) Korean dataset base model residual plot, (b) Korean dataset TL model residual plot, (c) Mexican dataset base model residual plot, (d) Mexican dataset TL model residual plot, (e) Spanish dataset base model residual plot and (f) Spanish dataset TL model residual plot

### 5.1.1 LSTM model performance on the South Korean dataset

Table 9 provides a comparison of the fully trained TL model and the base LSTM model under simulated data-scarce conditions using the South Korean dataset. The results show that the proposed model consistently outperformed the base model across all data-scarcity levels. Notably, in the most extreme scarcity condition (0.5 weeks), the TL model achieved an RMSE of 0.00572 and an MAE of 0.00167, even surpassing the base model's best performance under the least severe scarcity condition (3 months). These results underscore the robustness of the proposed approach in cold-start conditions, as evidenced by the consistently low RMSE and MAE values (below 0.002) across all simulations.

TABLE 9. Summary of the South Korean dataset results

| Data scarcity time period (Training-Validation data split percentage) | Base Model RMSE | Transfer Learning Model RMSE | Base Model Mean Average Error (MAE) | Transfer Learning Model MAE |
|---|---|---|---|---|
| 0.5 weeks (0.3196%) | 0.090999715 | 0.005721834 | 0.06420611 | 0.001674956 |
| 1 week (0.6393%) | 0.06848902 | 0.005151839 | 0.040081218 | 0.000907381 |
| 2 weeks (1.2785%) | 0.056696735 | 0.005481194 | 0.030633263 | 0.000924808 |
| 1 month (2.7397%) | 0.014864958 | 0.004400508 | 0.010172323 | 0.000538203 |
| 2 months (5.4795%) | 0.01104961 | 0.005463058 | 0.005055063 | 0.000592693 |
| 3 months (8.2192%) | 0.012316924 | 0.005676969 | 0.003386124 | 0.00070767 |

Figure 7 (a) and (b) visually confirm these findings under the most extreme data scarcity conditions. These prediction graphs show that the proposed model closely tracks PM values, but the base model fails to completely capture key data trends. These results demonstrate the benefit provided by the sequential transfer learning process in improving early-stage predictions. Figures 8 and 9 further support these findings by illustrating the TL model's performance on the most lenient data scarcity conditions. Figure 8 (a) and (b) show that even though both models generally follow the ideal prediction line, the proposed model exhibits tighter alignment with fewer outliers compared to the base model. Likewise, residual plots in Figure 9 (a) and (b) reinforce this result as the range of residuals for the TL model is lower than that of the base model.

### 5.1.2 LSTM model performance on the Mexican dataset

The proposed model can only be considered globally applicable if it demonstrates consistent performance across datasets from diverse geographical regions.  Table 10 presents a comparison of the performance of the fully trained model with that of the base LSTM model in terms of RMSE and MAE on the Mexican dataset. Both models performed less effectively on this dataset compared to the Korean and Spanish datasets, as shown by the generally higher RMSE and MAE values. This is likely attributable to both domain dissimilarity and greater irregularities in the pollution data. Transfer learning typically relies on trends captured in the source domain in order to make accurate predictions in target domains. However, this benefit may be reduced if the target domain contains substantial deviations from the established trends [57]. Comparing the appearance of the prediction graphs (Figure 7) shows that the Mexican dataset (Figure 7 (c) and (d)) exhibits a notably different temporal pattern from that of the Korean (Figure 7 (a) and (b)) and Spanish (Figure 7 (e) and (f)) datasets suggesting the presence of domain dissimilarity. Despite the overall higher values, the TL model continued to consistently outperform the base model across all data scarcity levels.

TABLE 10. Summary of the Mexican dataset results

| Data scarcity time period (Training-Validation data split percentage) | Base Model RMSE | Transfer Learning Model RMSE | Base Model Mean Average Error (MAE) | Transfer Learning Model MAE |
|---|---|---|---|---|
| 0.5 weeks (0.1765%) | 0.66157526 | 0.18666399 | 0.2531395 | 0.06367647 |
| 1 week (0.353%) | 0.40647915 | 0.11711181 | 0.15683462 | 0.03653864 |
| 2 weeks (0.706%) | 0.23002201 | 0.048291344 | 0.072740585 | 0.01648878 |
| 1 month (1.5129%) | 0.10263728 | 0.021225423 | 0.035882555 | 0.007105837 |
| 2 months (3.0257%) | 0.034600586 | 0.010601467 | 0.01007723 | 0.002508329 |
| 3 months (4.5386%) | 0.022593964 | 0.005965399 | 0.007580475 | 0.002044728 |

Figure 6 (c) and (d) illustrate the performance gap between the base and proposed model for this dataset under stringent data constraints. Both models struggle to fully capture the pollution trends accurately in this region, particularly during sharp spikes. However, the TL model demonstrates comparatively greater accuracy as it is able to better predict those spikes, although it underestimates their magnitude. Additionally, both models also exhibit some negative pollution predictions, highlighting a limitation of the LSTM model in this study.

The advantages of the proposed approach are also evident on this dataset under less stringent data scarcity conditions. Prediction scatter plots (Figure 8 (c) and (d)) demonstrate that the TL model outperforms the base model as its predictions closely align with the ideal prediction line, with a few outliers. Conversely, the base model continues to struggle making accurate predictions, with more of its prediction points falling further away from the ideal. The residual plot for this region (Figure 9 (b) and (c)) reinforce these findings, as while the TL model has a larger spike residual value on average it has lower residual values evidenced by the lower MAE of 0.002 compared to the base model's MAE of 0.0075 for the 3 month data split (Table 10). These results affirm that the sequential TL framework enhances the prediction accuracy of the base model in regions outside of the training datasets.

### 5.1.3 LSTM Model performance on the Spanish dataset

The third testing dataset (Spanish dataset) was used to further evaluate the predictive capabilities of the sequential transfer learning model. As with the previous cases, the proposed model demonstrates a consistent performance advantage over the base model, evidenced by the consistently lower TL RMSE and MAE values shown in Table 11. These results further support the TL model's effectiveness, particularly in handling the cold-start problem. The TL model also maintained strong performance on this dataset with RMSE values below 0.005 and MAE values below 0.00085 across all conditions. Notably, on this dataset, the TL model under the most severe data scarcity conditions also outperforms the best-performing base model under the most lenient data-limited conditions. These results not only highlight the model's excellent prediction accuracy but also underscore the model's ability to maintain its performance across geographically diverse regions. This suggests that the model captures transferable pollution trends that are applicable across multiple regions.

TABLE 11. Summary of Spain dataset results

| Data scarcity time period (Training-Validation data split percentage) | Base Model RMSE | Transfer Learning Model RMSE | Base Model Mean Average Error (MAE) | Transfer Learning Model MAE |
|---|---|---|---|---|
| 0.5 weeks (0.1369%) | 0.10298807 | 0.003475171 | 0.08416645 | 0.000847659 |
| 1 week (0.2738%) | 0.097825974 | 0.003223221 | 0.08108435 | 0.000735591 |
| 2 weeks (0.5475%) | 0.04250151 | 0.003208223 | 0.030142585 | 0.000391079 |
| 1 month (1.1733%) | 0.025380466 | 0.003694691 | 0.016366677 | 0.00034414 |
| 2 months (2.3465%) | 0.008298404 | 0.004493136 | 0.00386322 | 0.000399762 |
| 3 months (3.5197%) | 0.006602294 | 0.004921514 | 0.003538401 | 0.000355875 |

The prediction graphs for this dataset under the most limited data conditions (Figure 7 (e) and (f)) again illustrate the extent to which the proposed model outperforms the base model. While the base model again fails to capture PM variation effectively, the TL model is able to better capture those trends. Figures 8 - 9 (e) and (f) continue to show the distinction between the base model and the created model. The prediction scatter plot (Figure 8 (e) and (f)) confirms the superior prediction accuracy of the TL model, as the proposed model has points that are generally closer to the ideal line compared to the base model. The residual plots (Figure 9 (e) and (f)) again highlight the accuracy of both models, as they both have low residual values. Although the TL model exhibits a slightly higher peak residual, its residuals are more tightly clustered around zero, while the base model shows greater dispersion and more frequent small deviations. These findings further demonstrate the effectiveness of the proposed model in addressing the cold-start problem, as it consistently maintained strong predictive performance across all tested conditions.

## 5.2. Comparative analysis

This section compares the performance of the transfer learning method across each DL model (LSTM, 1D-CNN and GRU). Through this comparison, the effectiveness of the proposed sequential transfer learning method is assessed.

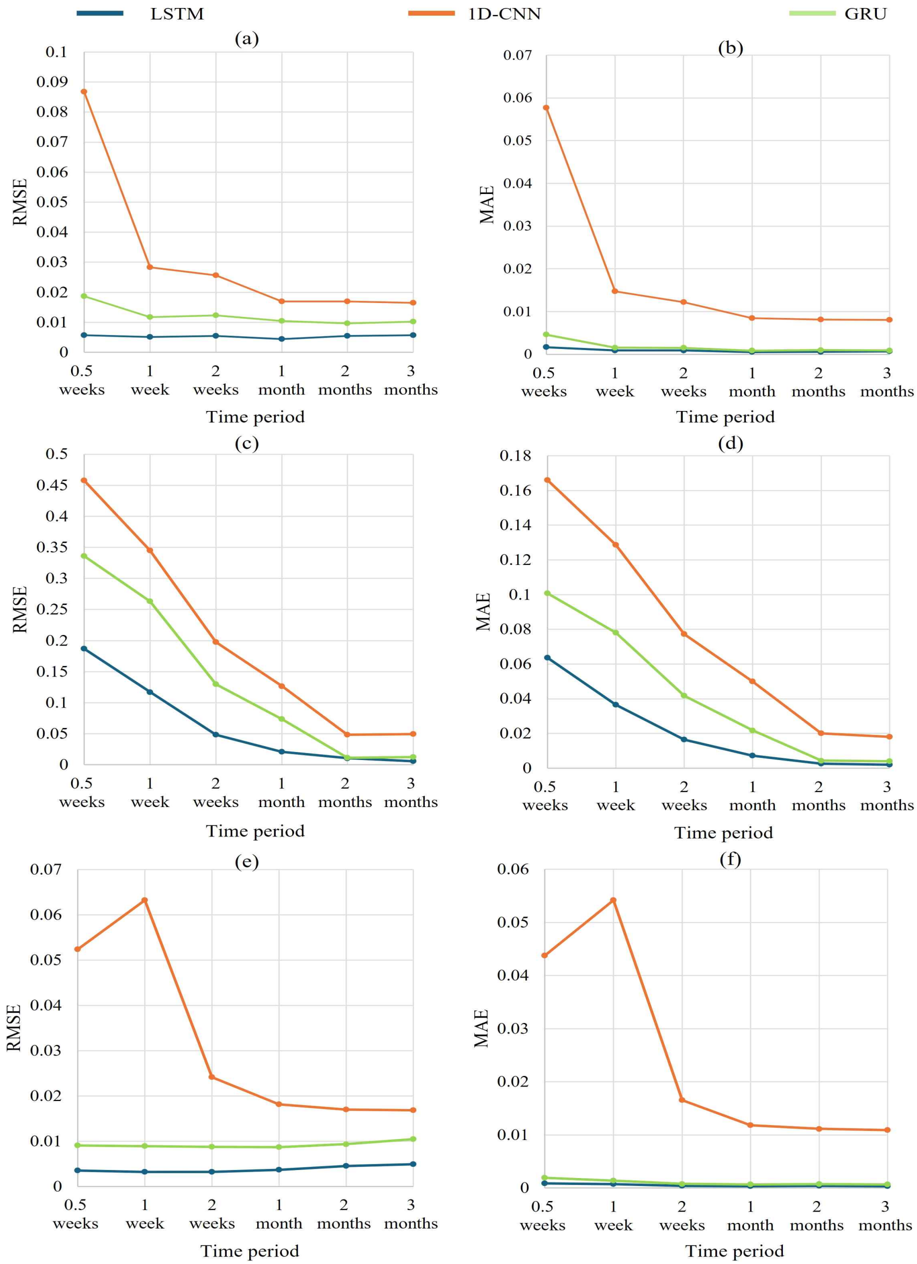


Figure 10. DL models' error versus data scarcity time period graphs: (a) DL models' RMSE performance on the Korean dataset, (b) DL models' MAE performance on the Korean dataset, (c)DL models' RMSE performance on the Mexican dataset, (d) DL models' MAE performance on the Mexican dataset, (e)DL models' RMSE performance on the Spanish dataset and (f) DL models' MAE performance on the Spanish dataset

Figure 10 presents the RMSE and MAE of all tested models across each data scarcity time period. These results demonstrate that the LSTM model consistently outperformed the other two models in all regions across all data scarcities. The 1D-CNN model exhibited the weakest overall performance across all regions, consistent with findings from prior studies [55], suggesting that it is less suited to capturing the complex temporal patterns in air pollution data.

Figure 11 visualizes the percentage improvement in RMSE and MAE for each model and dataset under cold start conditions. Each point in the figure corresponds to the percentage improvement for a specific data scarcity level, as outlined in Table 4. This visual representation illustrates the effectiveness of the sequential transfer learning approach under all defined data scarcities.

For the LSTM model on the Korean dataset, there is a general decrease in performance improvement as the amount of data available for training increases. The decrease is more pronounced in the RMSE, which started around 93% and dropped to 53%. In MAE, the improvement percentage decreased from 97% to 79%. Similar trends were also observed on the Spanish dataset; however, there was a much sharper decrease in RMSE performance improvement from 96% down to 25%, while the MAE improvement remained higher, ranging from 99% to just below 90%. This general decrease in performance improvement suggests that as the base model gains more data, the relative benefit of the TL model diminishes, since the base model becomes increasingly capable of learning the target domain independently. Also, the sharper decrease in RMSE over MAE may suggest that while the TL process enhances general predictive accuracy, it is less effective in handling extreme pollution spikes or outliers, which disproportionately affect RMSE.

In contrast, for the Mexican dataset, the percentage improvement remains around 75% across all metrics with slight fluctuation. It was not as affected by data scarcity conditions compared to the other two regions. This may stem from the dataset exhibiting a higher frequency of extreme pollution events (as seen in the prediction graphs, Figure 6), which can make forecasting more challenging. Furthermore, as previously discussed, the domain dissimilarity may limit the model's ability to fully leverage TL, resulting in more stable but modest performance gains.

The 1D-CNN model also benefited from the proposed method, with improvements ranging from 21% up to 68% in RMSE and 20% to 75% in MAE across all datasets. On this model, the error performance trend closely followed that of the LSTM, though with lower overall performance. The performance improvement again roughly decreased as data scarcity increased, with minor deviations observed in the Spanish dataset.

The GRU model benefited the least from the full sequential model training. In several cases, the base model marginally outperformed the TL model, indicative of negative transfer [58], where the addition of new source domain data disrupts the learned representations, leading to increased error. This phenomenon may be due to the GRU's simpler architecture, which may lack the capacity to retain complex multi-domain temporal features introduced during sequential training. Further exploration of this issue is provided in the following section on partial model evaluation.

Finally, Figure 12 illustrates the average percentage improvement for each model across all datasets. This bar graph highlights the overall effectiveness of the sequential transfer learning approach across all models. This summary again highlights that, on average, the LSTM benefited the most from the proposed method, with an average improvement ranging from 73% to above 95% across both performance metrics. The 1D-CNN model had moderate gains of 35% to 55%, and the GRU benefited the least with an average increase of 0.4% to 16% in error improvement.

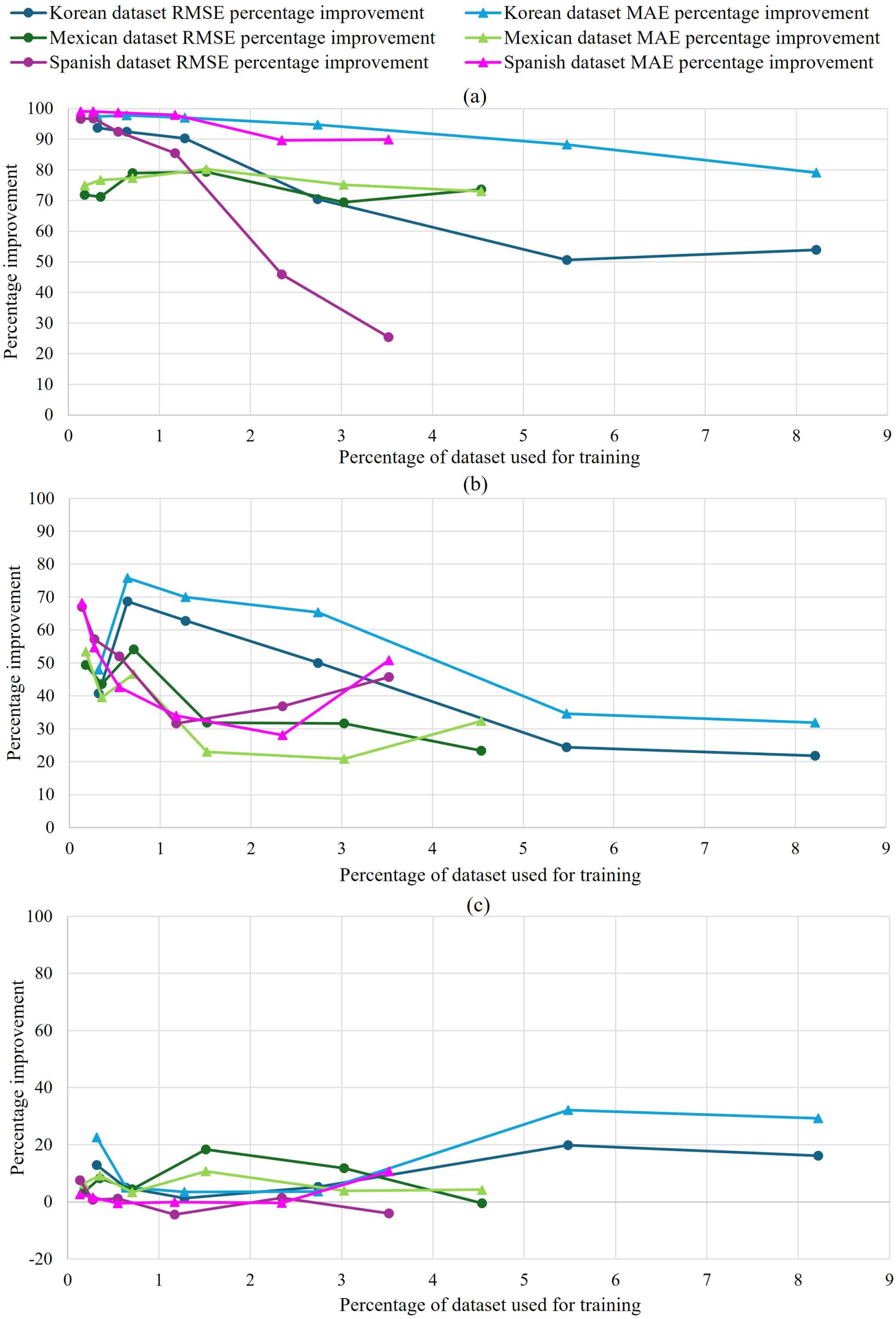


Figure 11. Percentage improvement in performance metrics (RMSE) and (MAE) across all data scarcity levels: (a) LSTM model, (b) 1D-CNN model, and (c) GRU model.

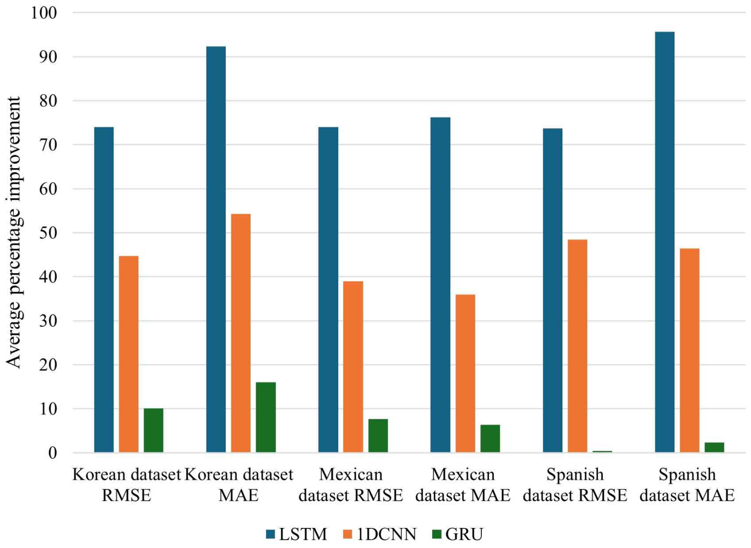
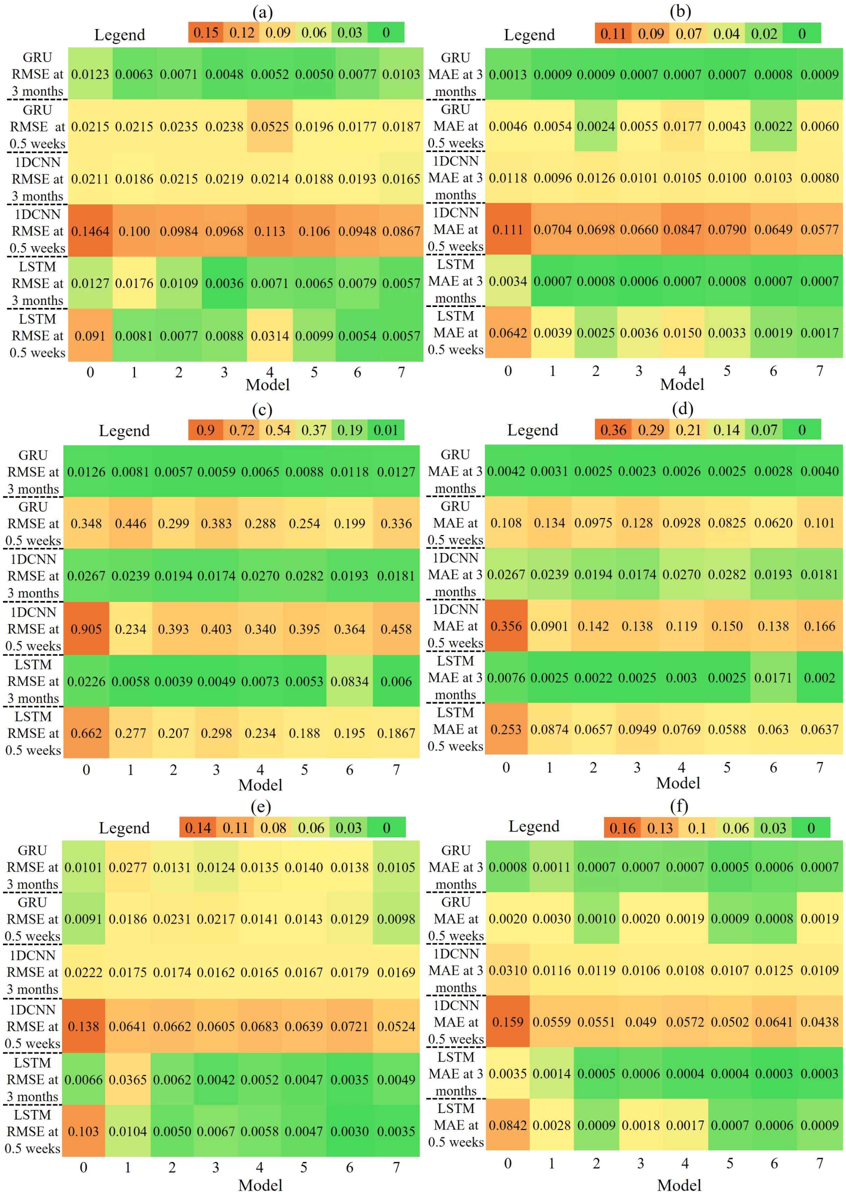


Figure 12. DL models' average percentage improvement in performance metrics

### 5.2.1 Partial transfer learning models' performance

Several partial TL models were developed during the process of creating the full sequential transfer learning model using the seven datasets (Figure 6). To evaluate the effect that the sequential transfer learning process had on the model's prediction performance, each partial version of each model (LSTM, 1D-CNN and GRU) was tested on the testing datasets (South Korea, Mexico and Spain), under both high and low data-scarcity conditions (0.5 weeks and 3 months, respectively). This analysis provides insight into how the incremental addition of regional knowledge affects predictive accuracy across different DL model architectures.



(a)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU RMSE at 3 months | 0.0123 | 0.0063 | 0.0071 | 0.0048 | 0.0052 | 0.0050 | 0.0077 | 0.0103 |
| GRU RMSE at 0.5 weeks | 0.0215 | 0.0215 | 0.0235 | 0.0238 | 0.0525 | 0.0196 | 0.0177 | 0.0187 |
| 1DCNN RMSE at 3 months | 0.0211 | 0.0186 | 0.0215 | 0.0219 | 0.0214 | 0.0188 | 0.0193 | 0.0165 |
| 1DCNN RMSE at 0.5 weeks | 0.1464 | 0.100 | 0.0984 | 0.0968 | 0.113 | 0.106 | 0.0948 | 0.0867 |
| LSTM RMSE at 3 months | 0.0127 | 0.0176 | 0.0109 | 0.0036 | 0.0071 | 0.0065 | 0.0079 | 0.0057 |
| LSTM RMSE at 0.5 weeks | 0.091 | 0.0081 | 0.0077 | 0.0088 | 0.0314 | 0.0099 | 0.0054 | 0.0057 |

(b)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU MAE at 3 months | 0.0013 | 0.0009 | 0.0009 | 0.0007 | 0.0007 | 0.0007 | 0.0008 | 0.0009 |
| GRU MAE at 0.5 weeks | 0.0046 | 0.0054 | 0.0024 | 0.0055 | 0.0177 | 0.0043 | 0.0022 | 0.0060 |
| 1DCNN MAE at 3 months | 0.0118 | 0.0096 | 0.0126 | 0.0101 | 0.0105 | 0.0100 | 0.0103 | 0.0080 |
| 1DCNN MAE at 0.5 weeks | 0.111 | 0.0704 | 0.0698 | 0.0660 | 0.0847 | 0.0790 | 0.0649 | 0.0577 |
| LSTM MAE at 3 months | 0.0034 | 0.0007 | 0.0008 | 0.0006 | 0.0007 | 0.0008 | 0.0007 | 0.0007 |
| LSTM MAE at 0.5 weeks | 0.0642 | 0.0039 | 0.0025 | 0.0036 | 0.0150 | 0.0033 | 0.0019 | 0.0017 |

(c)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU RMSE at 3 months | 0.0126 | 0.0081 | 0.0057 | 0.0059 | 0.0065 | 0.0088 | 0.0118 | 0.0127 |
| GRU RMSE at 0.5 weeks | 0.348 | 0.446 | 0.299 | 0.383 | 0.288 | 0.254 | 0.199 | 0.336 |
| 1DCNN RMSE at 3 months | 0.0267 | 0.0239 | 0.0194 | 0.0174 | 0.0270 | 0.0282 | 0.0193 | 0.0181 |
| 1DCNN RMSE at 0.5 weeks | 0.905 | 0.234 | 0.393 | 0.403 | 0.340 | 0.395 | 0.364 | 0.458 |
| LSTM RMSE at 3 months | 0.0226 | 0.0058 | 0.0039 | 0.0049 | 0.0073 | 0.0053 | 0.0834 | 0.006 |
| LSTM RMSE at 0.5 weeks | 0.662 | 0.277 | 0.207 | 0.298 | 0.234 | 0.188 | 0.195 | 0.1867 |

(d)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU MAE at 3 months | 0.0042 | 0.0031 | 0.0025 | 0.0023 | 0.0026 | 0.0025 | 0.0028 | 0.0040 |
| GRU MAE at 0.5 weeks | 0.108 | 0.134 | 0.0975 | 0.128 | 0.0928 | 0.0825 | 0.0620 | 0.101 |
| 1DCNN MAE at 3 months | 0.0267 | 0.0239 | 0.0194 | 0.0174 | 0.0270 | 0.0282 | 0.0193 | 0.0181 |
| 1DCNN MAE at 0.5 weeks | 0.356 | 0.0901 | 0.142 | 0.138 | 0.119 | 0.150 | 0.138 | 0.166 |
| LSTM MAE at 3 months | 0.0076 | 0.0025 | 0.0022 | 0.0025 | 0.003 | 0.0025 | 0.0171 | 0.002 |
| LSTM MAE at 0.5 weeks | 0.253 | 0.0874 | 0.0657 | 0.0949 | 0.0769 | 0.0588 | 0.063 | 0.0637 |

(e)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU RMSE at 3 months | 0.0101 | 0.0277 | 0.0131 | 0.0124 | 0.0135 | 0.0140 | 0.0138 | 0.0105 |
| GRU RMSE at 0.5 weeks | 0.0091 | 0.0186 | 0.0231 | 0.0217 | 0.0141 | 0.0143 | 0.0129 | 0.0098 |
| 1DCNN RMSE at 3 months | 0.0222 | 0.0175 | 0.0174 | 0.0162 | 0.0165 | 0.0167 | 0.0179 | 0.0169 |
| 1DCNN RMSE at 0.5 weeks | 0.138 | 0.0641 | 0.0662 | 0.0605 | 0.0683 | 0.0639 | 0.0721 | 0.0524 |
| LSTM RMSE at 3 months | 0.0066 | 0.0365 | 0.0062 | 0.0042 | 0.0052 | 0.0047 | 0.0035 | 0.0049 |
| LSTM RMSE at 0.5 weeks | 0.103 | 0.0104 | 0.0050 | 0.0067 | 0.0058 | 0.0047 | 0.0030 | 0.0035 |

(f)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| GRU MAE at 3 months | 0.0008 | 0.0011 | 0.0007 | 0.0007 | 0.0007 | 0.0005 | 0.0006 | 0.0007 |
| GRU MAE at 0.5 weeks | 0.0020 | 0.0030 | 0.0010 | 0.0020 | 0.0019 | 0.0009 | 0.0008 | 0.0019 |
| 1DCNN MAE at 3 months | 0.0310 | 0.0116 | 0.0119 | 0.0106 | 0.0108 | 0.0107 | 0.0125 | 0.0109 |
| 1DCNN MAE at 0.5 weeks | 0.159 | 0.0559 | 0.0551 | 0.049 | 0.0572 | 0.0502 | 0.0641 | 0.0438 |
| LSTM MAE at 3 months | 0.0035 | 0.0014 | 0.0005 | 0.0006 | 0.0004 | 0.0004 | 0.0003 | 0.0003 |
| LSTM MAE at 0.5 weeks | 0.0842 | 0.0028 | 0.0009 | 0.0018 | 0.0017 | 0.0007 | 0.0006 | 0.0009 |

**0** – Base Model| **1** – Model with North America Weights| **2** – North + South America Weights| **3** – North and South America + Europe Weights| **4** – North and South America, Europe + Africa Weights| **5** - North and South America, Europe, Africa + South Asia Weights| **6** - North and South America, Europe, Africa, South + East Asia Weights | **7** – Fully Trained Transfer Learning Model/ All Weights

Figure 13. Partial transfer learning models' performance heatmaps: (a) RMSE of each partial model on Korean dataset, (b) MAE of each partial model on Korean dataset, (c) RMSE of each partial model on Mexican dataset, (d) MAE of each partial model on Mexican dataset, (e) RMSE of each partial model on Spanish dataset, and (f) MAE of each partial model on Spanish dataset.

Figure 13 presents the performance of the partial models in terms of RMSE and MAE through heatmaps. This format facilitates direct comparison of model performance across domains, data scarcity levels, and architectures. Across all models and all regions, the addition of new regional weights occasionally resulted in increased error, indicating the presence of negative transfer.

For the LSTM model, across all datasets, the fully trained TL model ranked within the top three performing models, typically yielding the lowest or second-lowest RMSE and MAE values. This indicates that the LSTM model is better able to capture and take advantage of the patterns from all seven regions and confirms the overall effectiveness of the proposed TL strategy. While isolated instances of negative transfer occurred, they were outweighed by the broader gains provided by the full TL process.

A similar trend was observed with the 1D-CNN model. The fully trained TL model mostly had the lowest or second-lowest error values, with the exception of the Mexico dataset at 0.5 weeks, where the model with North American weights performed the best. Despite this outlier, the results support the conclusion that the sequential transfer approach is also well-suited to the 1D-CNN model.

In contrast, the GRU model deviated from the patterns observed in the other two models. As previously mentioned, the full sequential model did not provide as much performance improvements to predictions as in the case of the other two models. Instead, in this model, the partial TL models consistently outperformed the full model.

This suggests that GRU's simpler memory architecture may be more susceptible to disruption when exposed to heterogeneous domain characteristics, limiting its capacity to benefit from sequential transfer. Despite this, the fully trained TL LSTM models consistently outperformed the best-performing GRU partial models under equivalent conditions. This result further solidifies the conclusion that the LSTM model is most compatible with the proposed sequential transfer learning method.

### 5.2.2 Statistical reliability analysis

To assess the statistical reliability of the reported results, each model was re-evaluated using four additional random seeds, for a total of five independent runs. These experiments were conducted under the most data-scarce and data-rich conditions respectively to evaluate performance consistency across extremes. The mean and standard deviation of RMSE and MAE across these runs are presented in the error bar plots in Figure 14. Across all datasets and both data scarcity conditions (Fig. 14(a) and 14(b)), the TL models consistently outperformed their base counterparts, reinforcing the effectiveness of the proposed sequential transfer learning method. The TL LSTM model again demonstrated superior performance by yielding the lowest mean errors and the smallest standard deviations across nearly all instances, affirming its accuracy and robustness. Although the TL GRU model exhibited competitive mean performance in certain instances, it showed greater performance variability across runs, particularly under low data scarcity on the Korean dataset (Figure 14(a)), where it recorded the highest errors among the three models. The TL 1D-CNN model yielded moderately stable results but generally produced higher RMSE and MAE values compared to both models. These results further validate the LSTM model as the most reliable and robust architecture for implementing the proposed sequential transfer learning approach.

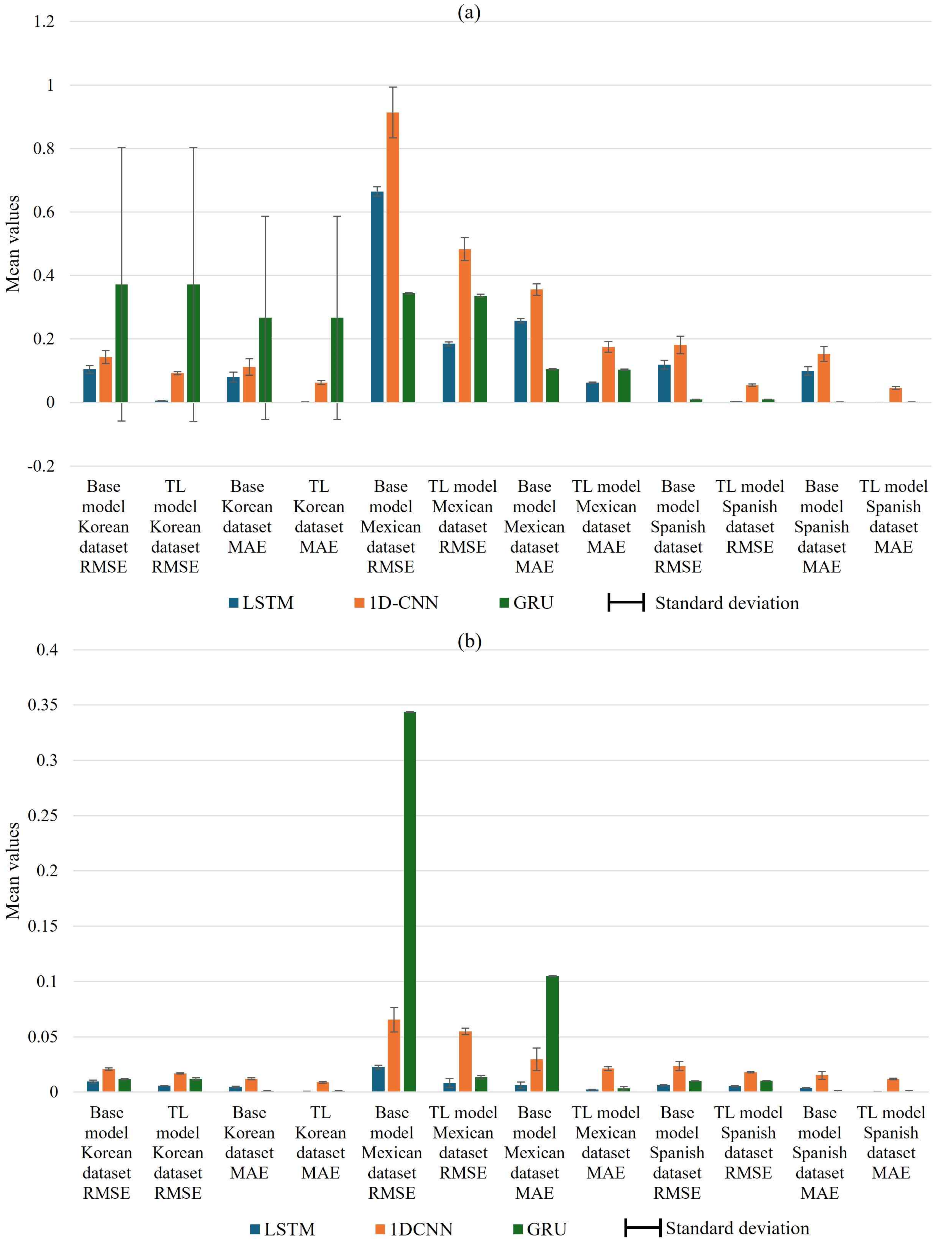


Figure 14. Comparative error bar graphs: (a) 0.5 weeks data scarcity level and (b) 3 months data scarcity level

# Chapter 6 Conclusion

This study investigated a globally generalizable approach for tackling the cold-start problem in air pollution prediction. An LSTM model was trained via sequential transfer learning on PM datasets from seven major regions around the world. These datasets were pre-processed while maintaining their base characteristics to ensure effective model training. This method enabled the model to capture any global patterns existing within the PM data. The model was subsequently tested on the three independent datasets taken from countries excluded from the training phase. These datasets were segmented to simulate several degrees of data scarcity to assess the model's ability to address the cold-start problem and its applicability to predict PM levels in new geographic regions.

The results demonstrate that the fully trained TL model consistently outperformed the base LSTM model across all datasets, with a performance improvement ranging from 25% to over 99%, depending on the dataset and scarcity condition. The proposed model consistently achieved lower RMSE and MAE values than the base model across all datasets and data splits, with an RMSE range from 0.187 down to 0.0032 and an MAE range from 0.0637 down to 0.00034. These findings validate the model's capacity to predict PM values accurately even under extreme data-scarcity conditions. Prediction graphs, scatter plots, and residual plots further support these results, showing that the developed model was capable of accurately predicting PM values with limited data, while the base LSTM model struggled under such conditions. These results demonstrate that the proposed model can effectively handle the cold-start problem while being globally generalizable, addressing this gap in air pollution studies.

The effectiveness of the sequential transfer learning method was also evaluated on a 1D-CNN and GRU models. It was found that the LSTM model outperformed these models across all data scarcity levels. Also, the partial sequential models were tested, and for the LSTM, despite the instances of negative transfer, the full TL model mostly outperformed the partial models, indicating the effectiveness of the proposed method.

While the proposed model demonstrated strong predictive capabilities, several limitations present opportunities for further research. Although random search was used for hyperparameter tuning, alternative optimization strategies such as Bayesian optimization or newer frameworks like Optuna could yield improved performance (Du et al., 2022; Hanifi et al., 2024), including optimizing the number of training epochs. The fixed training duration of 50 epochs, though based on convergence observations, could be replaced with adaptive strategies like early stopping to mitigate overfitting and enhance efficiency. Despite the LSTM model's established effectiveness in air pollution studies, recent findings suggest that hybrid and ensemble models may outperform standalone LSTMs (Sakar et al., 2022; Yang et al., 2024); thus, the current results should be considered a performance baseline for sequential transfer learning, with future work encouraged to extend the method to more advanced architectures. Moreover, while smaller models performed well, their real-time deployment capabilities were not assessed, warranting future investigation into adaptations for faster inference. Negative transfer was also observed, particularly in the GRU model, highlighting the importance of domain compatibility; future studies could address this by incorporating domain selection methods used in time series forecasting (Ye and Dai, 2021; 2022). Lastly, the exclusion of local environmental factors (e.g., wind speed, temperature, traffic) limited the model's fine-tuning potential; future work could integrate such variables through ensemble methods or use the proposed model as a generalizable base for region-specific fine-tuning.

Despite the limitations faced, the sequential transfer learning model shows potential for global applicability while effectively addressing the cold-start problem. The results suggest the existence of global trends in air pollution that enabled the model to generalize across diverse regions. This research presents a viable model for air quality forecasting in regions with limited air quality monitoring infrastructure, thereby contributing to better air pollution management and public health protection.

# References

1. Dominski, Fábio Hech, Joaquim Henrique Lorenzetti Branco, Giorgio Buonanno, Luca Stabile, Manuel Gameiro da Silva, and Alexandro Andrade. "Effects of air pollution on health: A mapping review of systematic reviews and meta-analyses." Environmental research 201 (2021): 111487.
2. Mannucci, Pier Mannuccio, Sergio Harari, Ida Martinelli, and Massimo Franchini. "Effects on health of air pollution: a narrative review." Internal and emergency medicine 10 (2015): 657-662.
3. Fong, Iat Hang, Tengyue Li, Simon Fong, Raymond K. Wong, and Antonio J. Tallon-Ballesteros. "Predicting concentration levels of air pollutants by transfer learning and recurrent neural network." Knowledge-Based Systems 192 (2020): 105622.
4. Pope 3rd, C. A., David V. Bates, and Mark E. Raizenne. "Health effects of particulate air pollution: time for reassessment?." Environmental health perspectives 103, no. 5 (1995): 472-480.
5. Zaini, Nur'atiah, Lee Woen Ean, Ali Najah Ahmed, and Marlinda Abdul Malek. "A systematic literature review of deep learning neural network for time series air quality forecasting." Environmental Science and Pollution Research (2022): 1-33.
6. Li, Ranran, Yuqi Dong, Zhijie Zhu, Chen Li, and Hufang Yang. "A dynamic evaluation framework for ambient air pollution monitoring." Applied Mathematical Modelling 65 (2019): 52-71.
7. Bai, Lu, Jianzhou Wang, Xuejiao Ma, and Haiyan Lu. "Air pollution forecasts: An overview." International journal of environmental research

and public health 15, no. 4 (2018): 780.

8. Kim, Bubryur, K. R. Sri Preethaa, Sujeen Song, R. R. Lukacs, Jinwoo An, Zengshun Chen, Euijung An, and Sungho Kim. "Internet of things and ensemble learning-based mental and physical fatigue monitoring for smart construction sites." Journal of Big Data 11, no. 1 (2024): 115.
9. Chen, Zengshun, Likai Zhang, Ke Li, Xuanyi Xue, Xuelin Zhang, Bubryur Kim, and Cruz Y. Li. "Machine-learning prediction of aerodynamic damping for buildings and structures undergoing flow-induced vibrations." Journal of Building Engineering 63 (2023): 105374.
10. Kim, Bubryur, Dong-Eun Lee, Gang Hu, Yuvaraj Natarajan, Sri Preethaa, and Arun Pandian Rathinakumar. "Ensemble machine learning-based approach for predicting of FRP–concrete interfacial bonding." Mathematics 10, no. 2 (2022): 231.
11. Masood, Adil, and Kafeel Ahmad. "A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance." Journal of Cleaner Production 322 (2021): 129072.
12. Li, Yanzhao, Ju-E. Guo, Shaolong Sun, Jianing Li, Shouyang Wang, and Chengyuan Zhang. "Air quality forecasting with artificial intelligence techniques: A scientometric and content analysis." Environmental Modelling & Software 149 (2022): 105329.
13. Kaur, Manjit, Dilbag Singh, Mohamed Yaseen Jabarulla, Vijay Kumar, Jusung Kang, and Heung-No Lee. "Computational deep air quality prediction techniques: a systematic review." Artificial Intelligence Review 56, no. Suppl 2 (2023): 2053-2098.
14. Yan, Rui, Jiaqiang Liao, Jie Yang, Wei Sun, Mingyue Nong, and Feipeng

Li. "Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering." Expert Systems with Applications 169 (2021): 114513.

15. Yoo, Jaeseong, and Jihoon Moon. "Bayesian Model Selection for Addressing Cold-Start Problems in Partitioned Time Series Prediction." Mathematics (2227-7390) 12, no. 17 (2024).
16. Sokhi, Ranjeet S., Nicolas Moussiopoulos, Alexander Baklanov, John Bartzis, Isabelle Coll, Sandro Finardi, Rainer Friedrich et al. "Advances in air quality research–current and emerging challenges." Atmospheric Chemistry and Physics Discussions 2021 (2021): 1-133.
17. Li, Xiang, Ling Peng, Xiaojing Yao, Shaolong Cui, Yuan Hu, Chengzeng You, and Tianhe Chi. "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation." Environmental pollution 231 (2017): 997-1004.
18. Bai, Yun, Bo Zeng, Chuan Li, and Jin Zhang. "An ensemble long short-term memory neural network for hourly PM2. 5 concentration forecasting." Chemosphere 222 (2019): 286-294.
19. Liu, Hui, and Rui Yang. "A spatial multi-resolution multi-objective data-driven ensemble model for multi-step air quality index forecasting based on real-time decomposition." Computers in Industry 125 (2021): 103387.
20. Zhou, Yanlai, Fi-John Chang, Li-Chiu Chang, I-Feng Kao, and Yi-Shin Wang. "Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts." Journal of cleaner production 209 (2019): 134-145.
21. Zhang, Luo, Peng Liu, Lei Zhao, Guizhou Wang, Wangfeng Zhang, and

Jianbo Liu. "Air quality predictions with a semi-supervised bidirectional LSTM neural network." Atmospheric Pollution Research 12, no. 1 (2021): 328-339.

22. Pak, Unjin, Jun Ma, Unsok Ryu, Kwangchol Ryom, U. Juhyok, Kyongsok Pak, and Chanil Pak. "Deep learning-based PM2. 5 prediction considering the spatiotemporal correlations: A case study of Beijing, China." Science of the Total Environment 699 (2020): 133561.
23. Du, Shengdong, Tianrui Li, Yan Yang, and Shi-Jinn Horng. "Deep air quality forecasting using hybrid deep learning framework." IEEE Transactions on Knowledge and Data Engineering 33, no. 6 (2019): 2412-2424.
24. Wu, Zhiyuan, Ning Liu, Guodong Li, Xinyu Liu, Yue Wang, and Lin Zhang. "Meta-Learning-Based Spatial-Temporal Adaption for Coldstart Air Pollution Prediction." International Journal of Intelligent Systems 2023, no. 1 (2023): 3734557.
25. Ma, Jun, Zheng Li, Jack CP Cheng, Yuexiong Ding, Changqing Lin, and Zherui Xu. "Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network." Science of The Total Environment 705 (2020): 135771.
26. Gugnani, Veena, and Rajeev Kumar Singh. "Analysis of deep learning approaches for air pollution prediction." Multimedia Tools and Applications 81, no. 4 (2022): 6031-6049.
27. Jairi, Idriss, Sarah Ben-Othman, Ludivine Canivet, and Hayfa Zgaya-Biau. "Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants." Environmental Technology & Innovation 36 (2024): 103793.

28. Guo, Zhuoyue, Canyun Yang, Dongsheng Wang, and Hongbin Liu. "A novel deep learning model integrating CNN and GRU to predict particulate matter concentrations." Process Safety and Environmental Protection 173 (2023): 604-613.
29. Huang, Guoyan, Xinyi Li, Bing Zhang, and Jiadong Ren. "PM2. 5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition." Science of the Total Environment 768 (2021): 144516.
30. Jin, Xue-Bo, Nian-Xiang Yang, Xiao-Yi Wang, Yu-Ting Bai, Ting-Li Su, and Jian-Lei Kong. "Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction." Mathematics 8, no. 2 (2020): 214.
31. Yan, Rui, Jiaqiang Liao, Jie Yang, Wei Sun, Mingyue Nong, and Feipeng Li. "Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering." Expert Systems with Applications 169 (2021): 114513.
32. Wen, Congcong, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu, and Tianhe Chi. "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction." Science of the total environment 654 (2019): 1091-1099.
33. Gilik, Aysenur, Arif Selcuk Ogrenci, and Atilla Ozmen. "Air quality prediction using CNN+ LSTM-based hybrid deep learning architecture." Environmental science and pollution research (2022): 1-19.
34. Sun, Qiang, Yanmin Zhu, Xiaomin Chen, Ailan Xu, and Xiaoyan Peng. "A hybrid deep learning model with multi-source data for PM 2.5 concentration forecast." Air Quality, Atmosphere & Health 14 (2021):

503-513.

35. Li, Taoying, Miao Hua, and X. U. Wu. "A hybrid CNN-LSTM model for forecasting particulate matter (PM2. 5)." Ieee Access 8 (2020): 26933-26940.
36. Huang, Chiou-Jye, and Ping-Huan Kuo. "A deep CNN-LSTM model for particulate matter (PM2. 5) forecasting in smart cities." Sensors 18, no. 7 (2018): 2220.
37. Seng, Dewen, Qiyan Zhang, Xuefeng Zhang, Guangsen Chen, and Xiyuan Chen. "Spatiotemporal prediction of air quality based on LSTM neural network." Alexandria Engineering Journal 60, no. 2 (2021).
38. Qi, Yanlin, Qi Li, Hamed Karimian, and Di Liu. "A hybrid model for spatiotemporal forecasting of PM2. 5 based on graph convolutional neural network and long short-term memory." Science of the Total Environment 664 (2019): 1-10.
39. Mao, Wenjing, Limin Jiao, Weilin Wang, Jianlong Wang, Xueli Tong, and Suli Zhao. "A hybrid integrated deep learning model for predicting various air pollutants." GIScience & Remote Sensing 58, no. 8 (2021): 1395-1412.
40. Srivastava, Harshit, and Santos Kumar Das. "Air pollution prediction system using XRSTH-LSTM algorithm." Environmental Science and Pollution Research 30, no. 60 (2023): 125313-125327.
41. Aggarwal, Apeksha, and Durga Toshniwal. "A hybrid deep learning framework for urban air quality forecasting." Journal of Cleaner Production 329 (2021): 129660.
42. Ahani, Ida Kalate, Majid Salari, and Alireza Shadman. "An ensemble multi-step-ahead forecasting system for fine particulate matter in urban

areas." Journal of cleaner production 263 (2020): 120983.

43. Xu, Xinghan, and Minoru Yoneda. "Multitask air-quality prediction based on LSTM-autoencoder model." IEEE transactions on cybernetics 51, no. 5 (2019): 2577-2586.
44. Liu, Duen-Ren, Shin-Jye Lee, Yang Huang, and Chien-Ju Chiu. "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning." Expert Systems 37, no. 3 (2020): e12511.
45. Lu, Jie, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. "Transfer learning using computational intelligence: A survey." Knowledge-Based Systems 80 (2015): 14-23.
46. Ma, Jun, Jack CP Cheng, Changqing Lin, Yi Tan, and Jingcheng Zhang. "Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques." Atmospheric Environment 214 (2019): 116885.
47. Yang, Junzi, Ajune Wanis Ismail, Yingying Li, Limin Zhang, and Fazliaty Edora Fadzli. "Transfer Learning-Driven Hourly PM2. 5 Prediction Based on a Modified Hybrid Deep Learning." IEEE Access (2023).
48. Yadav, Nishant, Meytar Sorek-Hamer, Michael Von Pohle, Ata Akbari Asanjan, Adwait Sahasrabhojanee, Esra Suel, Raphael E. Arku et al. "Using deep transfer learning and satellite imagery to estimate urban air quality in data-poor regions." Environmental Pollution 342 (2024): 122914.
49. Malhotra, Meenakshi, Savita Walia, Chia-Chen Lin, Inderdeep Kaur Aulakh, and Saurabh Agarwal. "A systematic scrutiny of artificial intelligence-based air pollution prediction techniques, challenges, and viable solutions." Journal of Big Data 11, no. 1 (2024): 142.

50. Zhang, Bo, Yi Rong, Ruihan Yong, Dongming Qin, Maozhen Li, Guojian Zou, and Jianguo Pan. "Deep learning for air pollutant concentration prediction: A review." Atmospheric Environment 290 (2022): 119347.

51. Jana, Swadesh, Asif Iqbal Middya, and Sarbani Roy. "Short-term air pollution prediction using graph convolutional neural networks." Technological Forecasting and Social Change 208 (2024): 123684.

52. Cleveland, Robert B., William S. Cleveland, Jean E. McRae, and Irma Terpenning. "STL: A seasonal-trend decomposition." J. off. Stat 6, no. 1 (1990): 3-73.

53. Ahn, Hyun, Kyunghee Sun, and K. Pio Kim. "Comparison of missing data imputation methods in time series forecasting." Computers, Materials & Continua 70, no. 1 (2022): 767-779.

54. Yu, Yong, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. "A review of recurrent neural networks: LSTM cells and network architectures." Neural computation 31, no. 7 (2019): 1235-1270.

55. Espinosa, Raquel, José Palma, Fernando Jiménez, Joanna Kamińska, Guido Sciavicco, and Estrella Lucena-Sánchez. "A time series forecasting based multi-criteria methodology for air quality prediction." Applied Soft Computing 113 (2021): 107850.

56. Qing, L. PM2. 5 Concentration Prediction Using GRA-GRU Network in Air Monitoring, Sustainability 15 (2023) 1973.

57. Ye, Rui, and Qun Dai. "Implementing transfer learning across different datasets for time series forecasting." Pattern Recognition 109 (2021): 107617.

58. Zhang, Wen, Lingfei Deng, Lei Zhang, and Dongrui Wu. "A survey on negative transfer." IEEE/CAA Journal of Automatica Sinica 10, no. 2

(2022): 305-329.

59. Du, Liang, Ruobin Gao, Ponnuthurai Nagaratnam Suganthan, and David ZW Wang. "Bayesian optimization based dynamic ensemble for time series forecasting." Information Sciences 591 (2022): 155-175.
60. Hanifi, Shahram, Andrea Cammarono, and Hossein Zare-Behtash. "Advanced hyperparameter optimization of deep learning models for wind power prediction." Renewable Energy 221 (2024): 119700.
61. Sarkar, Nairita, Rajan Gupta, Pankaj Kumar Keserwani, and Mahesh Chandra Govil. "Air Quality Index prediction using an effective hybrid deep learning model." Environmental Pollution 315 (2022): 120404.
62. Yang, Hong, Wenqian Wang, and Guohui Li. "Multi-factor PM2. 5 concentration optimization prediction model based on decomposition and integration." Urban Climate 55 (2024): 101916.
63. Ye, Rui, and Qun Dai. "A relationship-aligned transfer learning algorithm for time series forecasting." Information Sciences 593 (2022): 17-34.

# 순차적 전이 학습을 활용한 전 지구 대기오염 예측: 콜드 스타트 문제 해결을 중심으로

**애쉬 크리스토퍼 다니엘**

경북대학교 대학원 로봇및스마트시스템공학과
(지도교수 김법렬)

(초 록)

대기오염이 인체 건강에 부정적인 영향을 미친다는 것은 잘 알려져 있으며, 신뢰할 수 있는 대기오염 예측 모델은 이러한 영향을 완화하는 데 도움을 줄 수 있다. 기존의 대기오염 관련 연구들은 대부분 특정 지역이나 국가를 대상으로 딥러닝(Deep Learning, DL) 기법을 활용하여 대기오염을 예측해왔다. 이러한 접근 방식은 지역별 기후, 교통량 등의 요인이 대기오염에 영향을 미치기 때문에, 일반화에 한계가 있으며 대규모의 지역 특화 데이터셋을 요구한다. 본 연구는 이러한 한계를 극복하기 위해 보편적으로 적용 가능한 DL 모델을 개발하였으며, 동시에 콜드스타트(cold-start) 문제에 대한 잠재적 해결책을 제시한다. 본 연구에서는 장단기 기억(Long Short-Term Memory, LSTM) 기반의 DL 모델을 전 세계 7개 주요 지역의 대기오염 데이터셋에 대해 순차 전이 학습(sequential transfer learning) 방식으로 학습시켰다. 이와 같은 학습 방식은 다양한 지리적 지역의 패턴을 모델에 반영할 수 있게 하여, 지역 간 적용 가능성을 향상시켰다. 세 개의 훈련 데이터셋 외부 지역을 대상으로 한 데이터 부족 상황을 모의한 실험에서, 제안된 모델은 강력한 예측 성능을 보였으며 지역별 전이 가능성 차이를 확인할 수 있었다. 루트 평균 제곱 오차(RMSE) 및 평균 절대 오차(MAE) 지표를 통해 모델 성능을 평가한 결과, 제안된 전이 학습 기반 모델은 지역에 따라 기존 LSTM 모델 대비 약 73%에서 95%까지 향상된 성능을 보였다. 또한, 제안된 접근 방식은 (Gated Reccurent Unit, GRU) 및 (One-dimensional Convolutional Neural Network, 1D-CNN)을 포함한 다른 DL 구조에도 적용되어 그 유효성이 추가로 입증되었다. 이러한 결과는 본 모델이 콜드스타트 문제를 효과적으로 처리할 수 있음을 보여주며, 대기오염에 있어 글로벌한 패턴이 존재할 가능성을 시사한다.