Predicting PM2.5 Concentrations

Daniel Lovelock
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
lovelock@usc.edu

Ezra Magaram
Department of Computer Science
University of Southern California
Los Angeles, CA 90015
magaram@usc.edu

Oliver Swack
Department of Computer Science
University of Southern California
swack@usc.edu

Mitch Pi
Department of Computer Science
University of Southern California
mpi@usc.edu

Abstract

Air pollution causes more than 6.5 million deaths each year globally. Of these 6.5 million yearly deaths, 4.2 million are caused by outdoor air pollution. These deaths are the result of inhaling particulate matter (PM), with the most dangerous group being PM2.5 which refers to particulate matter with a diameter of 2.5 micrometers or less. At this diameter, the particulate matter can penetrate the lung barrier, and enter a person's blood system [OWID]. A lot of this pollution inhalation can be avoided by just remaining inside one's home. In this paper, we use machine learning techniques such as Convolutional LSTM, and Graph Nerual Networks to estimate the air quality index in the short term (5 day forecast). Producing this forecast would give people the opportunity to plan their near future around these estimates, and stay safe.

1 Introduction

1.1 Motivation

Air pollution causes more than 6.5 million deaths each year globally. Of these 6.5 million yearly deaths, 4.2 million are caused by outdoor air pollution. These deaths are the result of inhaling particulate matter (PM), with the most dangerous group being PM2.5 which refers to particulate matter with a diameter of 2.5 micrometers or less. At this diameter, the particulate matter can penetrate the lung barrier, and enter a person's blood system [OWID]. A lot of this pollution inhalation can be avoided by just remaining inside one's home. It's quite difficult to tell whether or not the air is safe to inhale just by looking outside.

1.2 Problem Statement

Given the substantial health risks associated with PM2.5 exposure and the limitations of current air quality monitoring systems, there is a pressing need for an advanced predictive model. This research aims to develop a robust, accurate forecasting model for aerosol quality, specifically focusing on PM2.5 concentrations, for a designated location. The objective is to provide a reliable forecast of air quality, empowering individuals, particularly those most vulnerable, with the foresight to plan their activities and minimize exposure during peak pollution hours. Such a predictive model would not only enhance public awareness of air quality risks but also serve as a valuable tool for health officials and policymakers in devising targeted strategies for reducing exposure to harmful air pollutants. The

anticipated outcome of this research is the creation of an actionable, user-friendly forecasting system that can significantly contribute to reducing the health burden of air pollution globally.

This research will address the following key questions:

- How can we accurately predict PM2.5 levels in a specific area?
- What are the most effective data sources and computational models for predicting PM2.5 concentrations?
- How does our model compare with existing solutions?

1.3 High-level Outline of Goals and Approaches

The biggest issue with current aerosol prediction models is their computational cost. They are extremely expensive and require hours to complete an accurate short term forecast. By implementing a graph neural network to produce short term, hour by hour forecasts of PM2.5 concentrations, we can promptly notify high risk individuals so that they may adjust their behavior in order to stay safe. Our goal is to build two models: a CNN-LSTM and a GNN, that can produce a 5 day hour by hour forecast after being fed the preceding days PM2.5 concentrations, and other factors such as wind speed and direction, temperature, and humidity.

2 Related Work

Short-Term Prediction of PM2.5 Using LSTM Deep Learning Methods by Kristiani, Lin, Lin et al. details multiple methods that the authors used to predict PM2.5 levels: CNN, RNN, LSTM, GRU, bi-LSTM, and bi-GRU models. Of these models, the LSTM performed the best, with an RMSE value of 1.9, while other models had higher values, such as the CNN, which had an RMSE of 3.5. The authors' motivation to try out these models was that their prior attempts on training a model to predict PM2.5 values with an RNN had suffered from gradient explosion and gradient disappearance during the training process, significantly hampering accuracy. The dataset used in this paper was a collection of data taken from observation stations throughout Taiwan, using the relevant variables of temperature, CO, NO, NO_2, NO_x, O_3, PM10, PM2.5, and SO_2 levels. This data was then normalized and fed into the model to predict PM2.5 levels 8 hours into the future.

In this 2022 paper, *A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5)*, a hybrid CNN-LSTM model is devised to marry the strengths of convolutional neural networks (CNN) and long short-term memory (LSTM) neural networks to predict PM2.5 concentration in Beijing over a 24-hour time period. The CNN aptly captures air quality-related features, while the LSTM excels at capturing long-term historical patterns in time series data. The study's model incorporates two one-dimensional convolutional layers and a MaxPooling layer for feature extraction. To prepare the data for LSTM processing, the authors flatten the CNN outputs. To address the risk of overfitting, the authors utilize dropout, which they state to be a simple and efficient technique for the sake of their study. To address missing values in the dataset, the authors decided to fill the gaps with zeroes. For training the model, they did a multitude of tests and discovered that they received the best results (after testing 1-14 day-long sequences) that training with 7-day sequences followed by making the 24-hour prediction yielded the best results. The paper used RMSE and MAE for their loss functions and recorded an average of 17.93 and 13.97 respectively.

In a June 2023 paper titled, *Regional Aerosol Forecasts Based on Deep Learning and Numerical Weather Prediction*, the researchers built a spatial-temporal deep learning framework called PPN, or Pollution-Prediction Net for PM2.5. The goal of this model was to produce a 3-day PM2.5 forecast over the Beijing-Tianjin-Hebei region in China on a three-hour by three-hour basis. The researchers built their model by injecting the feature variables in different convolutional layers in order of their impacts on PM2.5, to imitate the behavior of CTMs. Also, the researchers used the PM2.5 values from multiple preceding timesteps to provide an accurate initial field. By implementing all of these novel approaches to aerosol forecasting, the researchers were able to achieve decent accuracy with an R^2 value of 0.7 and an RMSE of 17.7 micrograms per cubic meter.

3 Data

Description
In this project, we use climate data from 2018-2023. The features that we decided to include are air temperature, air humidity, wind speed and direction, and PM2.5 concentrations. The data was stored in nc4 files with a resolution of 0.25×0.25 degrees. The files spanned 360 degrees in width and 180 degrees in height. Lastly, the files were created at 30-minute intervals.

Data sources
The data for this project was provided by NASA's global modeling and assimilation office. All of the files can be downloaded at the following link: *https://fluid.nccs.nasa.gov/cf/*. We specifically used the Replay data.

Data statistics on the data source size
The initial dataset consisted of 5 years of climate data taken at 30-minute increments 24/7. Each 30-minute increment time shot was an nc4 file size 15.6-15.7 megabytes. However, each of these files contained unnecessary variables such as other aerosol concentrations that we are not trying to predict such as CO2 and SO2.

Extended Dataset for GNN
For the GNN model, we have developed a dataset that combines the ERA5 reanalysis data for weather parameters with PM2.5 concentration data obtained from NASA's GEOS CF model, specifically tailored for the US. This dataset covers the year 2018 and includes information for the top 100 most populated cities in the US, such as New York and Los Angeles. The data is recorded in 3-hour intervals (2920 3-hour timesteps) for each city. The weather data, extracted from ERA5, encompasses atmospheric pressure levels of 925, 950, and 975 hectopascals (hPa), including `'relative_humidity'`, `'specific_humidity'`, `'temperature'`, `'u_component_of_wind'`, `'v_component_of_wind'`, `'vertical_velocity'`, and `'vorticity'`. Additionally, the dataset includes crucial atmospheric variables not specific to pressure levels, such as `'100m_u_component_of_wind'`, `'100m_v_component_of_wind'`, `'2m_dewpoint_temperature'`, `'2m_temperature'`, `'boundary_layer_height'`, `'k_index'`, `'surface_pressure'`, and `'total_precipitation'`. The PM2.5 values are derived from NASA's GEOS CF model, ensuring accurate and high-resolution pollution data. The methodology for compiling and analyzing this dataset draws parallels to the approach described in the PM2.5GNN paper, with adaptations made to incorporate the combined datasets effectively.

In addition, to detect barriers between cities spanning 1200 m, we acquired elevation data for the entire United States. This data was obtained at a resolution of $0.05°$ longitude by $0.05°$ latitude. For this purpose, we utilized the Open Elevation API, which provided comprehensive and detailed topographical information necessary for our analysis.

Preprocessing
To convert our data which was in the form of .nc4 files into something usable for our model, we had to create a custom dataset designed to hold sequences in the form of numpy arrays of 23 hours worth of data (consisting of 20x20 geographic coordinates, each with 5 variables: pm25 levels, u and v as wind vectors, humidity, and temperature), as well as the associated label for that sequence, which would be the last hour of the day (so each day is divided into 23 hours of a sequences and the last hour being the label). The label for each sequence was the pm25 levels for each coordinate for the 24th hour. For our baseline approach, we did a proof-of-concept on a small set of data (ten days) on a model with CNN and LSTM layers.

Originally, our data was of the dimension 721*1440*5 for each file, with 721 latitude, 1440 longitude, and 5 variables for each point. We quickly found that the data size was going to be way too much for our computational resources, and we decided to take a smaller set of coordinates. We decided to focus on a region partly bound by the location of the USC Campus, and we did this by choosing the coordinate closest to USC campus. As the data resolution is 0.25 for both longitude and latitude, this coordinate was (34,-118). Then we iterated in 19 steps north and 19 steps east (these directions were chosen because our 20x20 grid would not be over the ocean). We then were able to iterate through our data to produce numpy arrays that would be used for our dataset.

Our custom dataset would hold both of these in a list, and the indices for the sequence and labels would correspond with each other. So essentially they were being held in a dictionary. The *__getitem__* function of our dataset would convert these sequences and labels into tensors.

Data statistics of processed dataset
After processing, the size of the nc4 file went from 15 megabytes to 7 megabytes. Our total dataset for the year was around 122 GB.

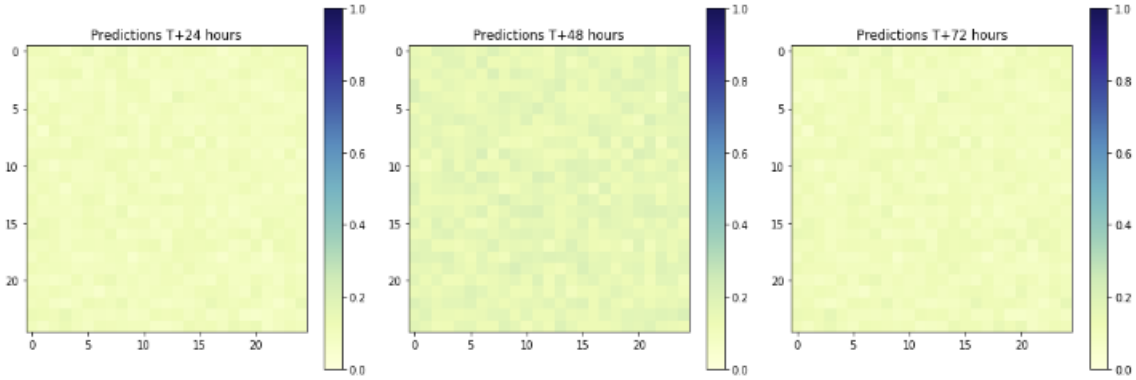
4 Task + Approach + Results

4.1 CNN Task

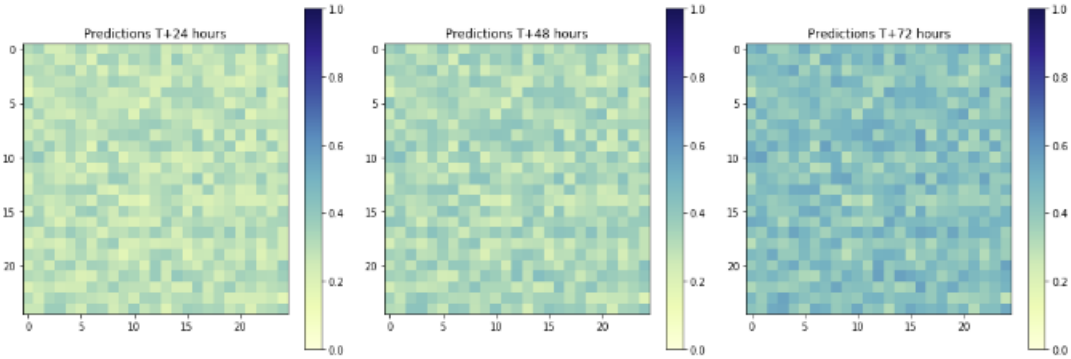
The objective is to create a way to predict short-term PM2.5 concentration data in the Los Angeles region. We did this by constructing two distinct models for predicting PM2.5 concentrations. The first model employs a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. This model takes as input a sequence of PM2.5 concentration data spanning 3 days, recorded at 30-minute intervals, and generates an output forecasting PM2.5 concentrations for the subsequent 3 days, maintaining the 30-minute granularity. The space that we are predicting for is a scaled-down version of the original dataset, with a size of 25×25 grid where each cell represents $1/4$ longitude by $1/4$ latitude totaling an area of 6.25 longitude by 6.25 latitude, or about 90,000 square miles with the bottom left of the grid in Los Angeles. The reason we added a CNN layer is to allow the influence of the neighboring regions to affect a given cell's PM2.5 concentration. Given the PM2.5 Concentrations for each point in the 25×25 grids in the previous 3-day sequence, we want to predict future values for each time step in the grid for the next 3 days. For our loss function, we used Mean Squared Error (MSE), as this is a regression task. MSE was chosen over other loss functions such as MAE because we wanted to weigh the outliers more.

Then for our error analysis, we chose to use RMSE over MSE as it would contextualize our results better than the MSE, essentially giving us the advantages of both the Mean Absolute Error (MAE) and Mean Squared Error (MSE).

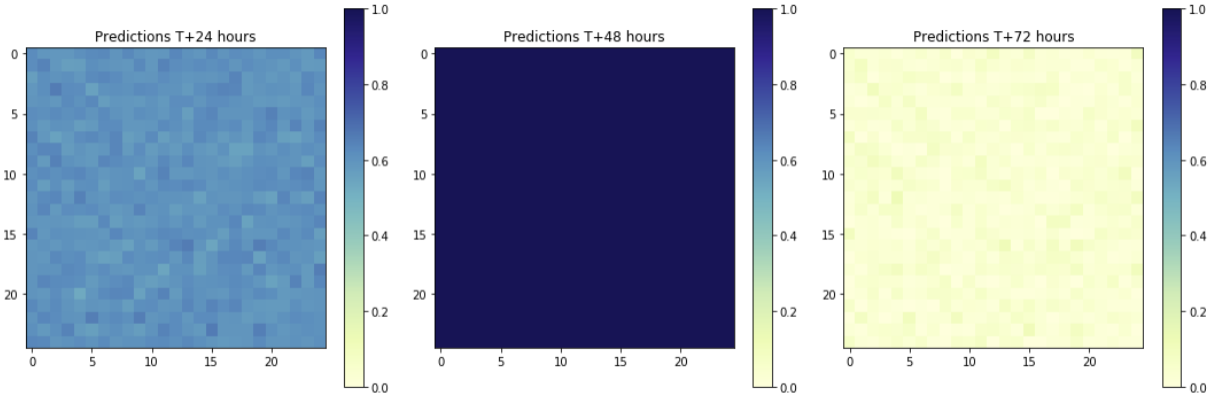
4.2 CNN-LSTM Results



The first trend in the results reveals a remarkable accuracy in predictions closely aligning with the actual true values. This success can be attributed to the absence of significant events influencing PM2.5 concentrations during the forecast period. In the absence of external disruptions, the model excels in capturing the regular patterns and trends in air quality, leading to highly accurate predictions.This is the most common scenario, reflected by MSE value of .0276 and a RMSE of 0.1661. To put these error metrics in context, the mean PM2.5 levels for our data were 6.21 μg/m3, the max value was 50.17 μg/m3, min was 1.36 μg/m3, and the standard deviation was 4.77 μg/m3.



The second trend indicates a notable decline in accuracy as the forecasting horizon extends. Following an initial day of relatively accurate predictions, there is a gradual decrease in precision on day two, further exacerbated on day three. This decline is attributed to the model's reliance on previously predicted values for subsequent forecasts. Any inaccuracies in the initial predictions are compounded over time, resulting in a cascading effect that leads to less reliable forecasts as the prediction horizon progresses.



The third trend underscores a substantial disparity between predicted and real values, primarily triggered by external events such as a forest fire or other significant pollution-contributing incidents. In these cases, the model faces challenges in adapting to sudden and substantial changes in the environment, leading to a divergence between predicted and actual PM2.5 concentrations. The model's performance is particularly impacted when confronted with unforeseen events that deviate significantly from the training data, emphasizing the importance of accounting for such disruptive events in enhancing model robustness.

4.2.1 GNN Task

Following the architecture of Wang et. al (2020), we created a directed graph with cities as nodes, using PM2.5 data from the NASA GEOS-CF Dataset. The PM2.5-GNN model combines a knowledge-enhanced GNN for horizontal pollutant transport and a spatio-temporal GRU for vertical pollutant dynamics under weather conditions. This model predicts PM2.5 levels up to 72 hours ahead, outperforming previous methods limited to small-scale datasets.

To optimize the performance of our PM2.5-GNN model, we employed the following training parameters: a batch size of 32, over a total of 50 epochs. The model was subjected to 10 experimental repetitions to ensure robustness and reliability of the results. We set the historical length to 1, allowing the model to use immediate past data for predictions, and the prediction length was set to 24, enabling the model to forecast PM2.5 levels up to 24 hours ahead. To prevent overfitting, we applied a weight decay of 0.0005 and implemented an early stopping mechanism triggered after 10 epochs

without improvement in validation loss. The learning rate was set to a modest 0.0005, balancing the speed of convergence and the risk of overshooting minimal loss values.

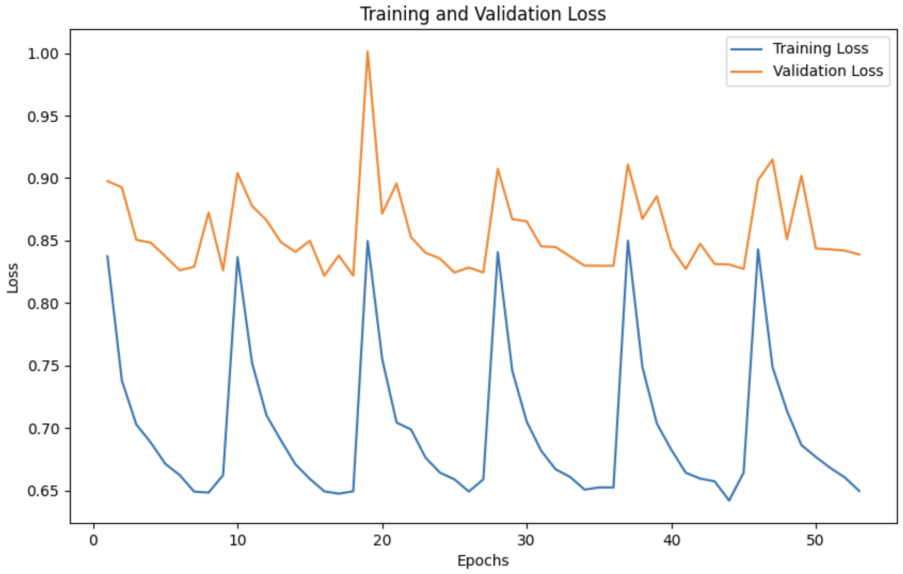
Given our limited dataset of just 2018 (we couldn't pull more data from ERA5 in time), we made our train/validation/test split as follows:

- **Train:** January 1st to August 31st
- **Validation:** September 1st to October 31st
- **Test:** November 1st to December 31st

4.3 GNN Results

We evaluated the performance of our proposed model across multiple metrics. The model exhibited a mean training loss of 0.6076 with a standard deviation of 0.0117, indicating a consistent learning pattern across different training iterations. The validation and test losses were higher, with means of 0.8046 and 0.8752, respectively, and relatively low standard deviations (0.0048 for validation and 0.0128 for test), suggesting a stable performance in unseen data scenarios. The model achieved a mean Root Mean Square Error (RMSE) of 19.9010 and a mean Mean Absolute Error (MAE) of 15.4997, with standard deviations of 0.3796 and 0.3277, respectively. These values reflect the model's accuracy in predicting PM2.5 concentrations. In terms of categorical metrics, the Critical Success Index (CSI) had a mean of 0.2430 with a standard deviation of 0.0059, while the Probability of Detection (POD) and False Alarm Ratio (FAR) recorded means of 0.4746 and 0.6656, with standard deviations of 0.0279 and 0.0211, respectively. These results indicate a moderate level of effectiveness in the model's ability to correctly identify hazardous PM2.5 levels, albeit with a relatively high rate of false alarms.

In the process of training our GNN, we noted a pattern of volatility in the training loss, characterized by sharp declines followed by abrupt increases. This could be indicative of several potential issues, such as overfitting, an aggressive learning rate, or insufficiently representative data. To mitigate these spikes in loss, we propose several strategies beyond merely increasing the quantity of training and validation data. Firstly, a more diverse dataset spanning a full year could provide a broader range of examples to enhance the model's generalization capabilities. Secondly, we suggest implementing regularization techniques and possibly adjusting the learning rate to achieve a more stable descent in loss. Thirdly, we would consider employing a validation strategy that includes cross-validation to ensure the robustness and reliability of the model's performance. By combining these approaches, we aim to create a more robust GNN that is capable of effectively learning from its training environment and generalizing well to unseen data.



5 Conclusion

5.1 Contributions

The biggest contribution that our code provides is speed. Currently, the only aerosol prediction model that is accurate is NASA-GEOS's weather prediction systems. However, running this model could take days. Our CNN-LSTM provides a great improvement to that time, in that the time to make a prediction in around 30 seconds. With this improvement, we can update those who are in affected areas much sooner than we would be able to if we were just using the GEOS model. With these shorter update times, people would have more time to adjust their schedules, inhaling less pollutants, resulting in a healthier population.

5.2 Limitations CNN

The CNN-LSTM does not take any features into account other than past PM2.5 values. Since we rana time series, the model might be able to implicitly take into account other features like wind speed and direction. This is because wind typically follows certain patterns, so by modeling PM2.5 with a time series, the model might be taking those patterns into account. However, a model that actually took in these values as input, especially if it also had access to accurate forecasted wind data would be a lot more successful at finding the relationships between the weather features and PM2.5 concentrations. Another huge drawback is that the model can't take into account any big events like forest fires or house fires.

5.3 Summary

In this research paper, we have embarked on a significant journey to develop a predictive model for PM2.5 concentrations, a crucial factor in air pollution and public health. While our model is still in the developmental stage, it is essential to acknowledge the limitations and assumptions that currently shape our project. One of the primary limitations is the availability and granularity of data. Our model's effectiveness hinges on the accuracy and comprehensiveness of the input data, yet we have encountered challenges in obtaining real-time, high-resolution data that can significantly impact the model's precision. Additionally, our understanding of the complex interactions between various atmospheric factors and PM2.5 levels is still evolving, which may affect the model's predictive accuracy.

Looking forward, we envision several promising extensions to our work. Once the model is operational, a critical next step will be to refine it through iterative testing and validation against real-world scenarios. This process will not only improve the model's accuracy but also help identify new data sources and variables that could enhance its predictive capabilities. Furthermore, we plan to explore the integration of our model into public health advisories and planning tools, making it a practical asset for communities and policymakers. By continually evolving our model to adapt to new scientific findings and technological advancements, we aim to make a meaningful contribution to managing the health impacts of air pollution.

5.4 Possible Extensions

We propose several enhancements to our current methodology to deepen the study's insights and broaden its applicability. One significant modification involves shifting the focus from the most popular cities to the most polluted ones. This change, coupled with an increase in the number of nodes, would provide a more comprehensive understanding of environmental challenges in areas most affected by pollution. By expanding our scope to include a larger and more diverse set of locations, we can gain a more nuanced understanding of the global environmental situation.

Additionally, extending the temporal range of our study beyond just the year 2018 would allow for a more thorough analysis of trends and patterns over time. This longitudinal approach could uncover more detailed insights into the evolution of environmental conditions and the effectiveness of various interventions. However, it's important to note that our current study was constrained by the time-intensive process of downloading data from the ERA5 database. This limitation highlights the need for more efficient data retrieval methods in future research.

Another promising extension involves incorporating a traffic congestion dataset into our analysis. By adding traffic-related features for each city, we can explore the relationship between traffic patterns and environmental conditions. This integration could reveal significant correlations and provide a more holistic view of the factors contributing to urban pollution. The inclusion of traffic data would enrich our understanding of the complex interplay between human activities and environmental health, offering valuable insights for policymakers and urban planners. Overall, these proposed extensions aim to enhance the depth and breadth of our study, making it a more robust and comprehensive resource for understanding and addressing environmental challenges.

References

[1] Kristiani, E., Lin, H., Lin, J., Chuang, Y.-H., Huang, C.-Y., & Yang, C.-T. (2022) *Short-Term Prediction of PM2.5 Using LSTM Deep Learning Methods.* Sustainability, 14(2068). doi:10.3390/su14042068.

[2] Li, T., Hua, M., & Wu, X. (2020) *A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5).* IEEE Access, 8, 26933-26940. doi:10.1109/ACCESS.2020.2971348.

[3] Qiu, Y., Feng, J., Zhang, Z., et al. (2023) *Regional aerosol forecasts based on deep learning and numerical weather prediction.* npj Climate and Atmospheric Science, 6(71). doi:10.1038/s41612-023-00397-0.

[4] Wang et al. (2020). *PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2.5 Forecasting*. https://doi.org/10.1145/3397536.3422208