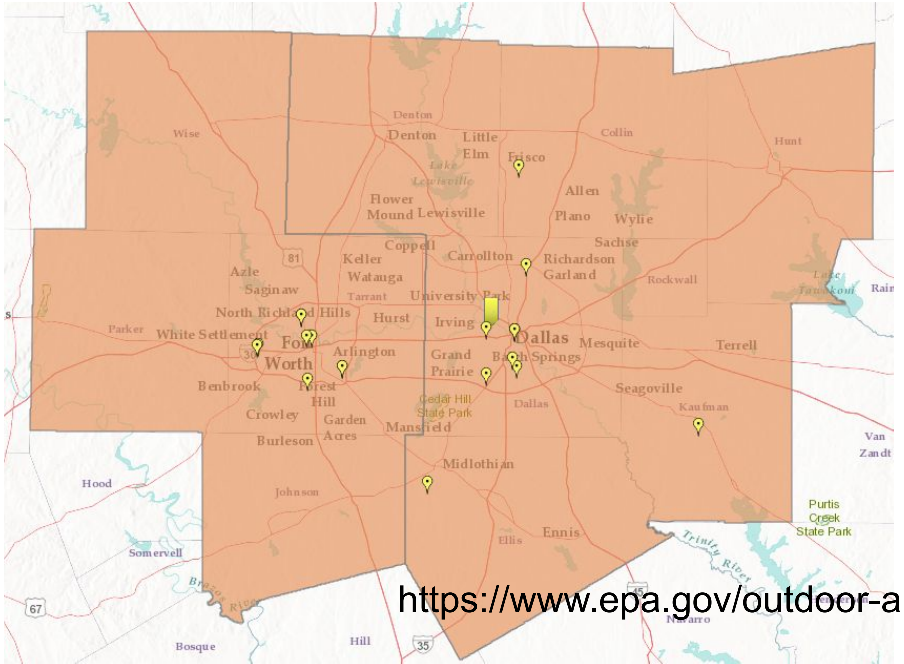
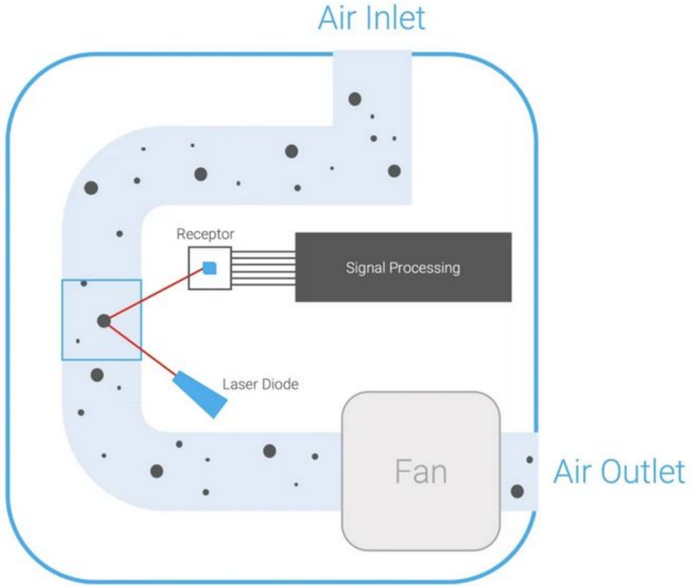
Air-Quality (PM2.5) prediction with Machine-learning models

Prahlad Siwakoti
NSS Data Science Cohort 7

Overview

- Local air-quality variations are difficult to monitor.
- Inexpensive optical-scattering sensors offer a solution.
  - Granular insights from a denser array of sensors





Laser sensors schematic

https://www.epa.gov/outdoor-air-quality-data/interactive-map-air-quality-monitors

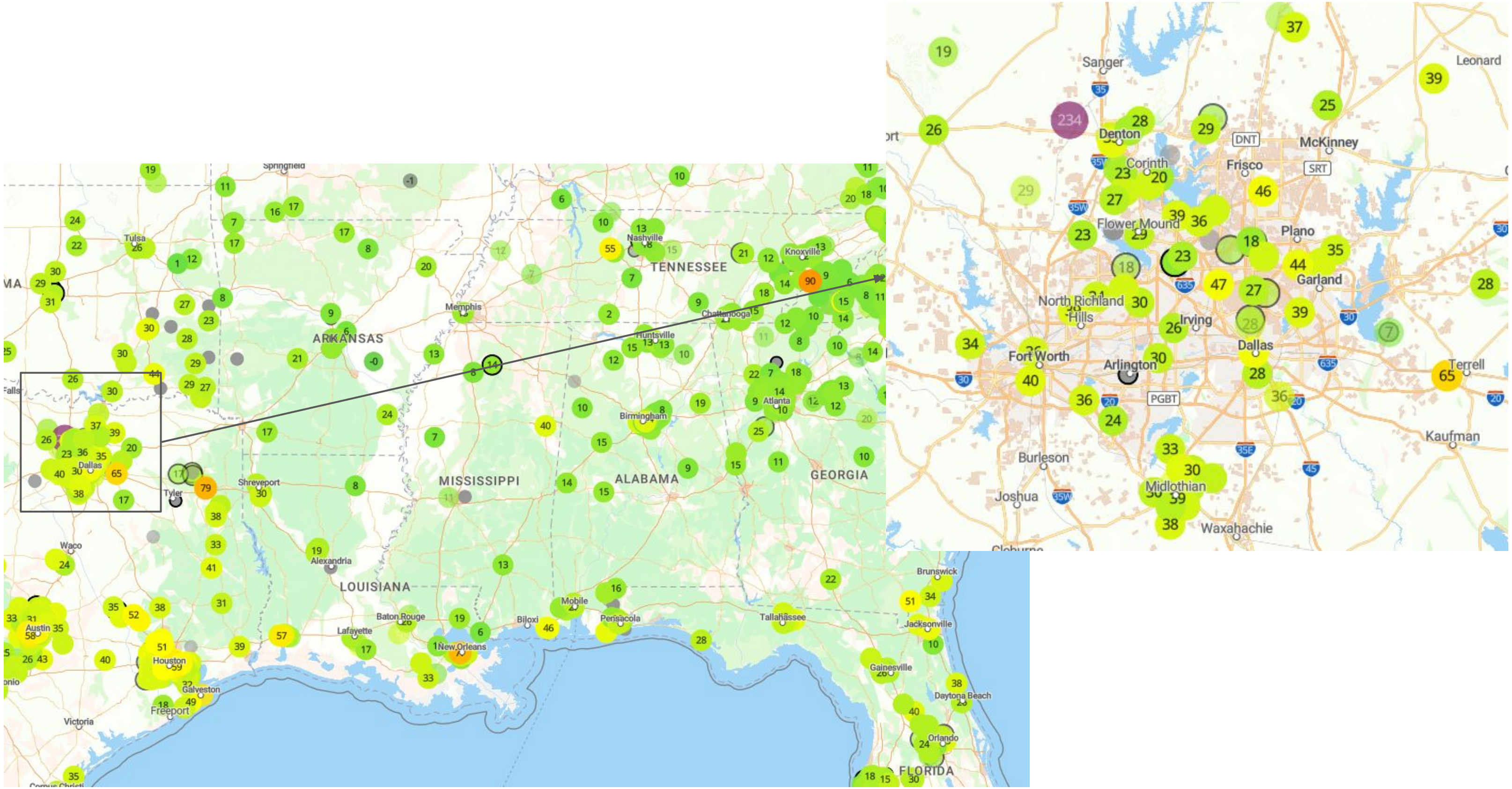
Data Importing and Cleaning

Used PurpleAir api[1] to collect historical data for Dallas Metropolitan region for the date range : 2022/04/01 to 2024/04/01

- ❖ First a list of sensors in the area defined by a geographical bounding box were identified.
- ❖ Historical data for a selected date range requested
- ❖ Filter, clean and load into a sqlite database
- ❖ Weather data for the date range is gathered from NOAA[2]

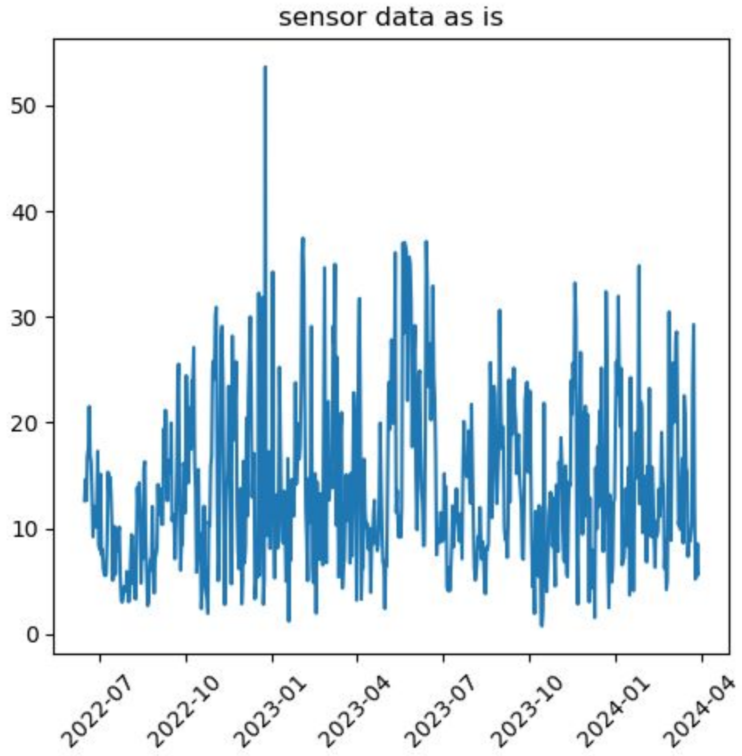
- ◆ 1. https://api.purpleair.com
- ◆ *2. https://www.ncei.noaa.gov/cdo-web/api/v2/{endpoint}*

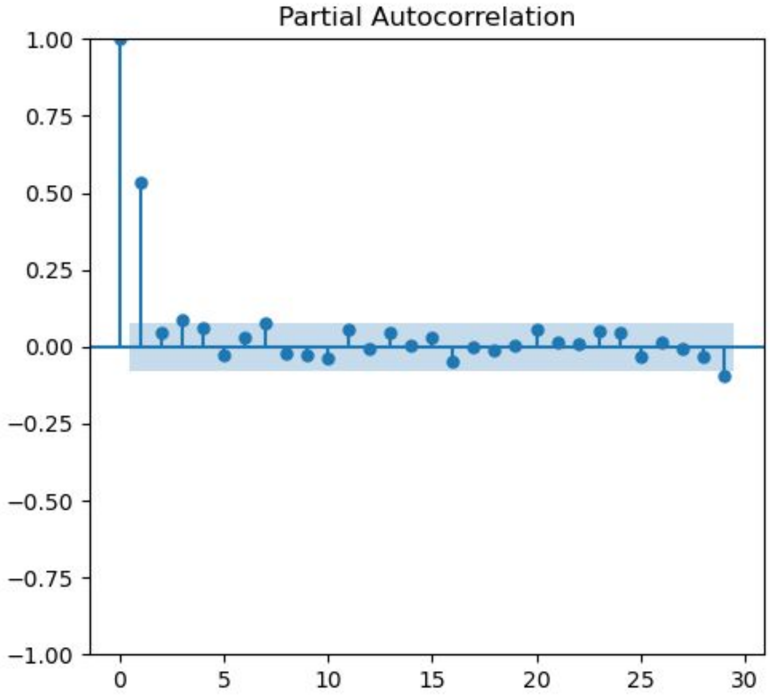
Real-time map with PM2.5 values



Daily PM2.5 data for a single sensor

- ❏ stationary
- ❏ dependent





Approach

Modelling Approaches

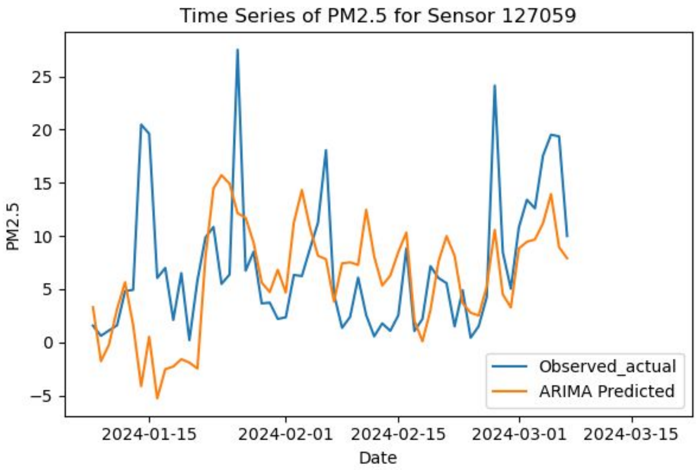
Temporal model: Time series forecasting

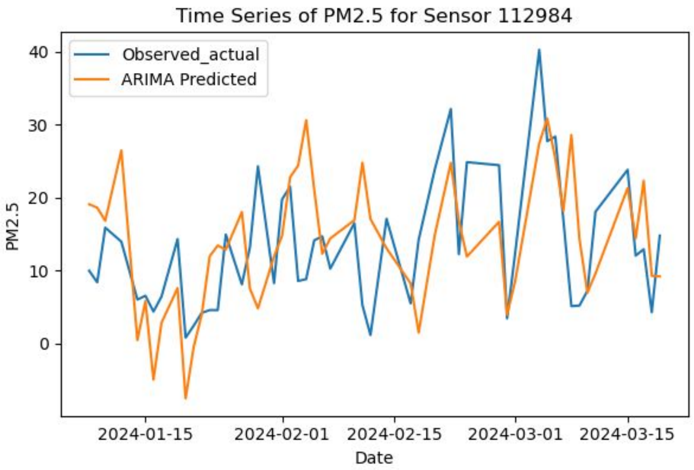
- Weather data included : Avg temp and Weather type (e.g. Fog, Rain etc)
- Multivariate ARIMA (1 0 1)
- XGBoost (with Temporal lag feature)
- No access to the spatial dimension

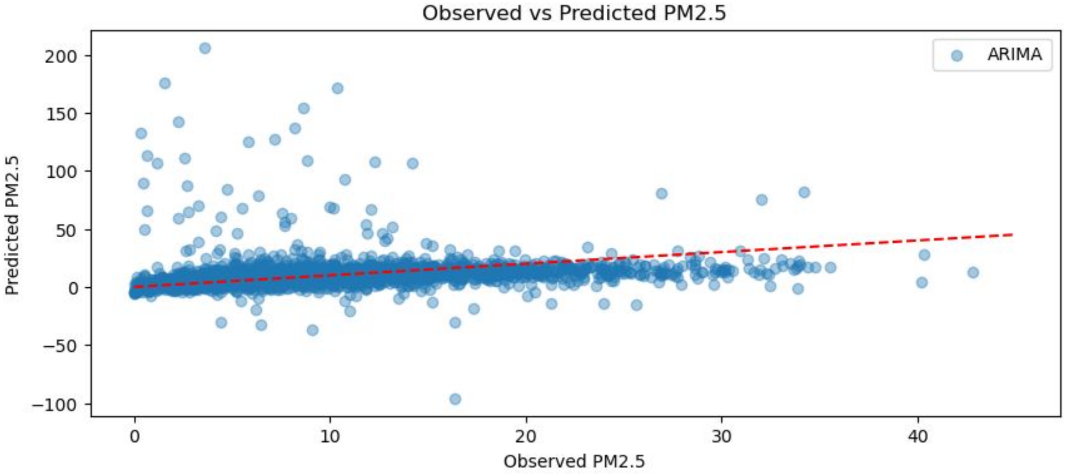
Spatial model:

- Feature engineering to create spatial_feature to fit a regression model:
    - Use spatial_weights to create spatial_features corresponding to the PM2.5 values.
    - Train by leaving a sensor out
- Does not have access to the weather data

Time Series Forecasting with ARIMA

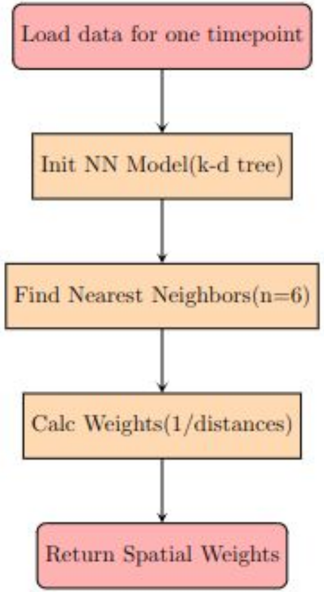






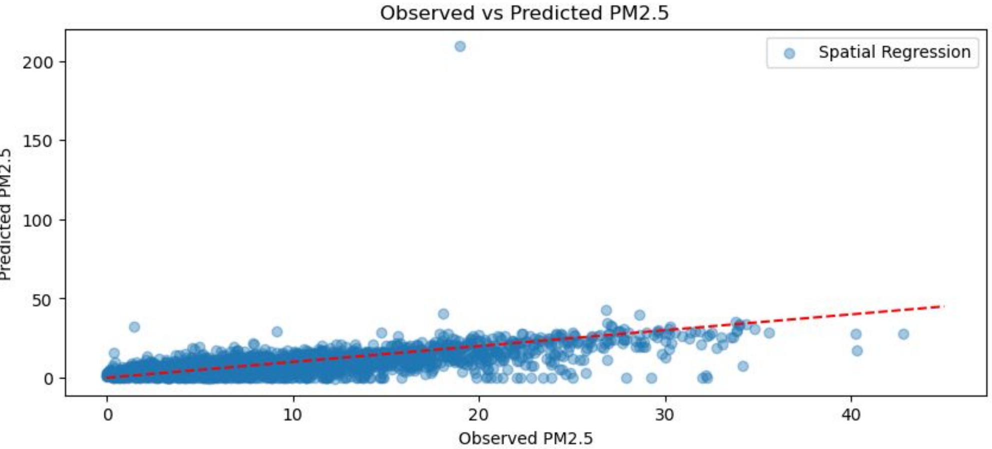
Humidity is the only variable that is significant at the 5% level

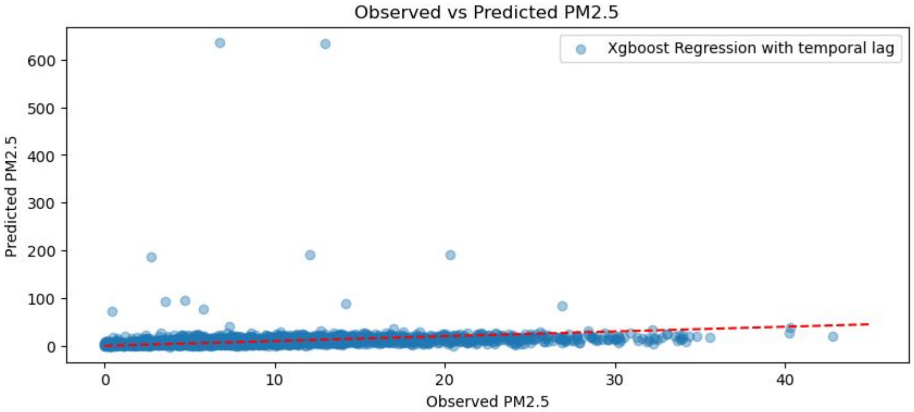
Spatio-Temporal prediction with XGBoost Regression

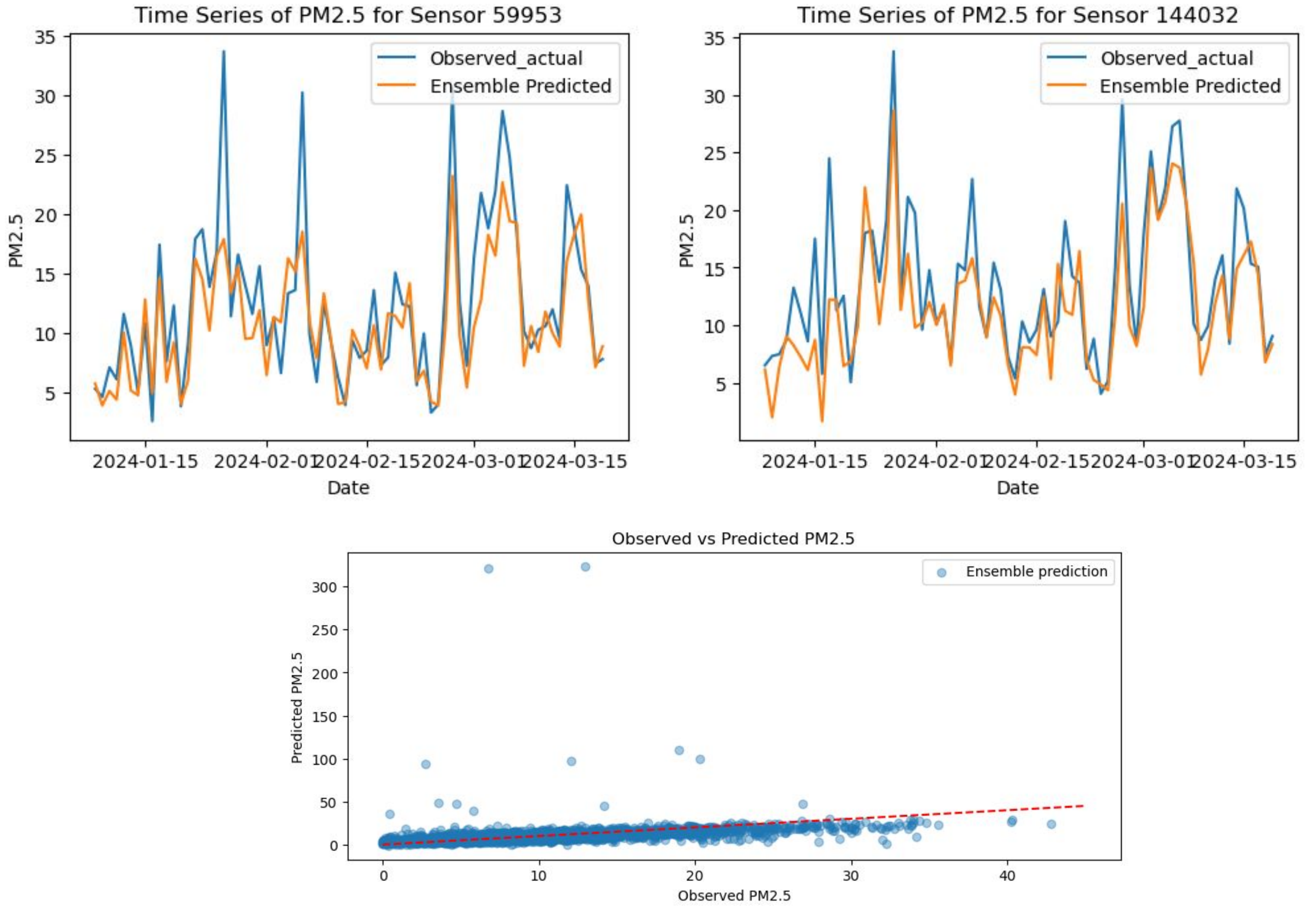


Feature Engineering :

1) Spatial lag feature based on proximity to sensors in the geospatial dimension:
    a) Does not have weather data
2) Temporal lag with a shift of 1 :
    a) Has weather data
3) Spatio-Temporal prediction : Average of the two

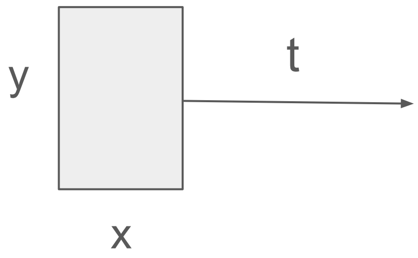


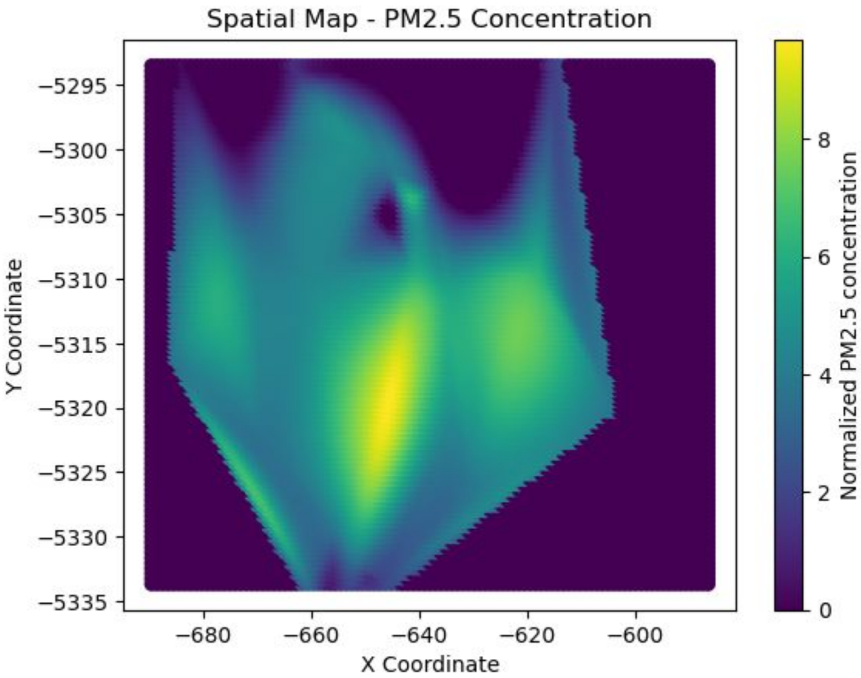



Time Series of PM2.5 for Sensor 59953
Observed_actual
Ensemble Predicted
PM2.5
Date
2024-01-15
2024-02-01
2024-02-15
2024-03-01
2024-03-15
Time Series of PM2.5 for Sensor 144032
Observed vs Predicted PM2.5
Ensemble prediction
Predicted PM2.5
Observed PM2.5

3D CNN for spatio-temporal models

- Temporal sequence of two-dimensional snapshots of values in geographical space





Single day map

- Stack 2d images over time
- Capture both spatial and tempora tendencies

Design and Workflow

Spatial Map Generation:

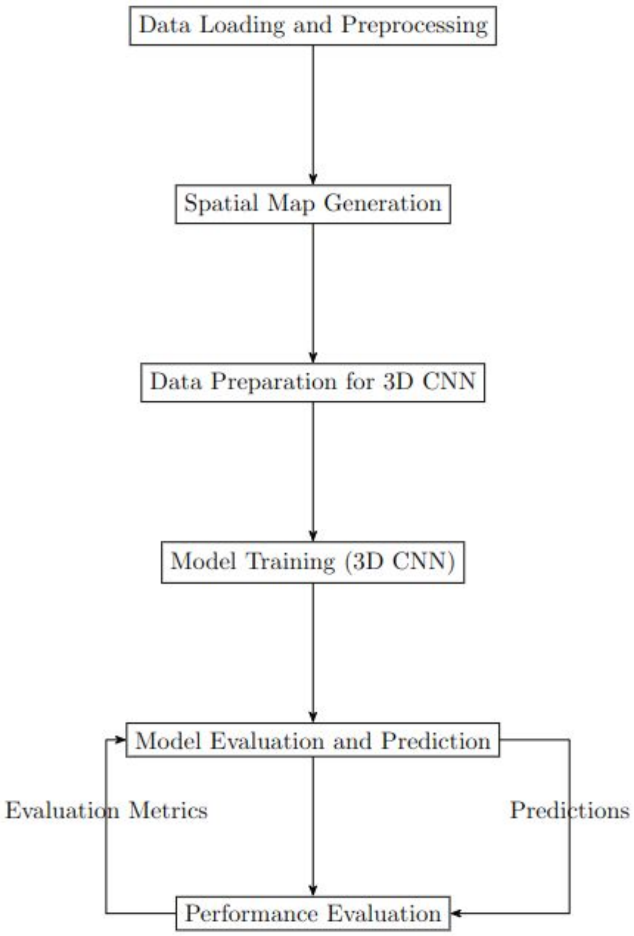
- Generate a series of spatial maps for each timestamp, capturing the spatial distribution of PM2.5 levels.

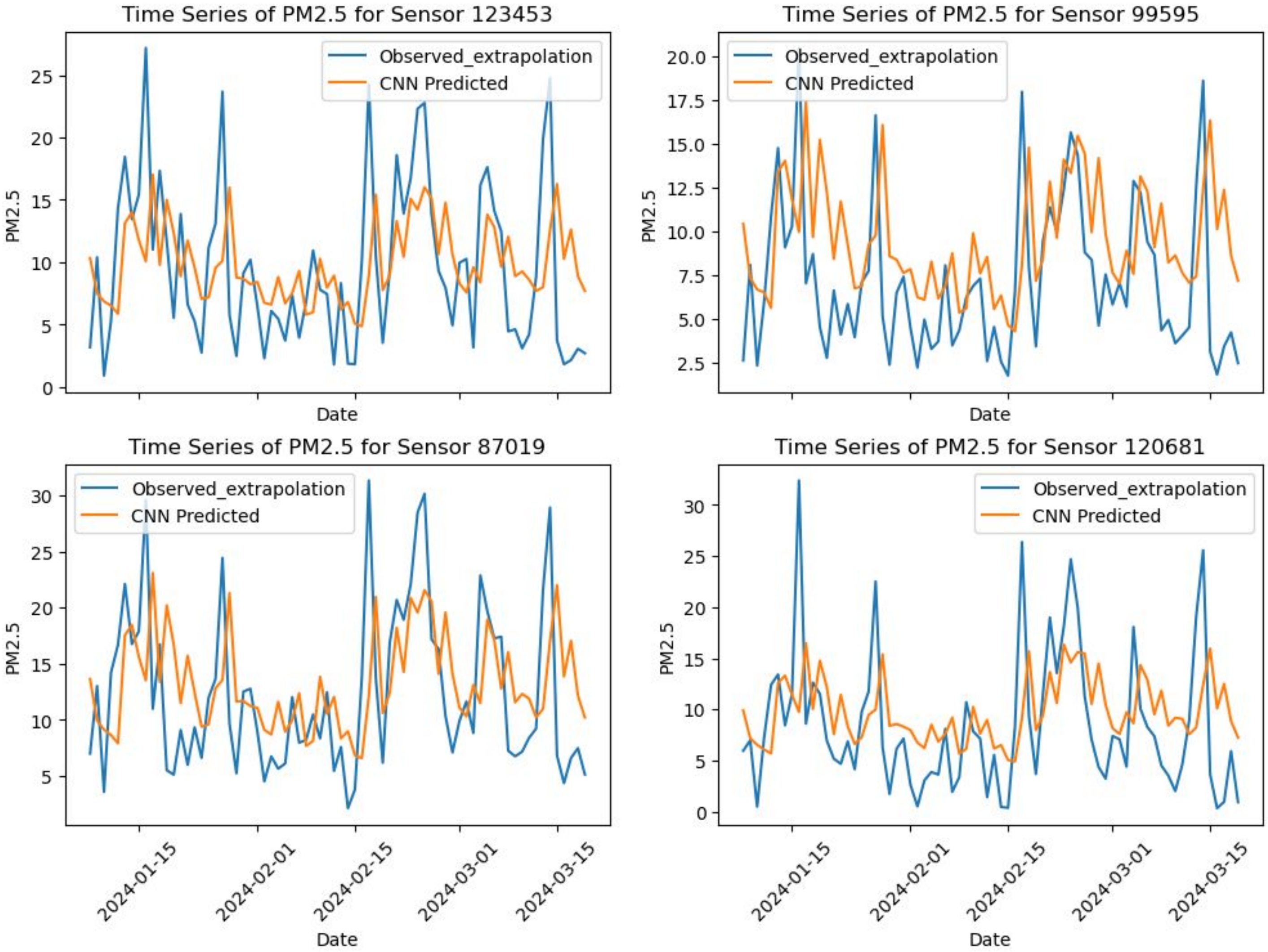
Data Preparation for 3D CNN:

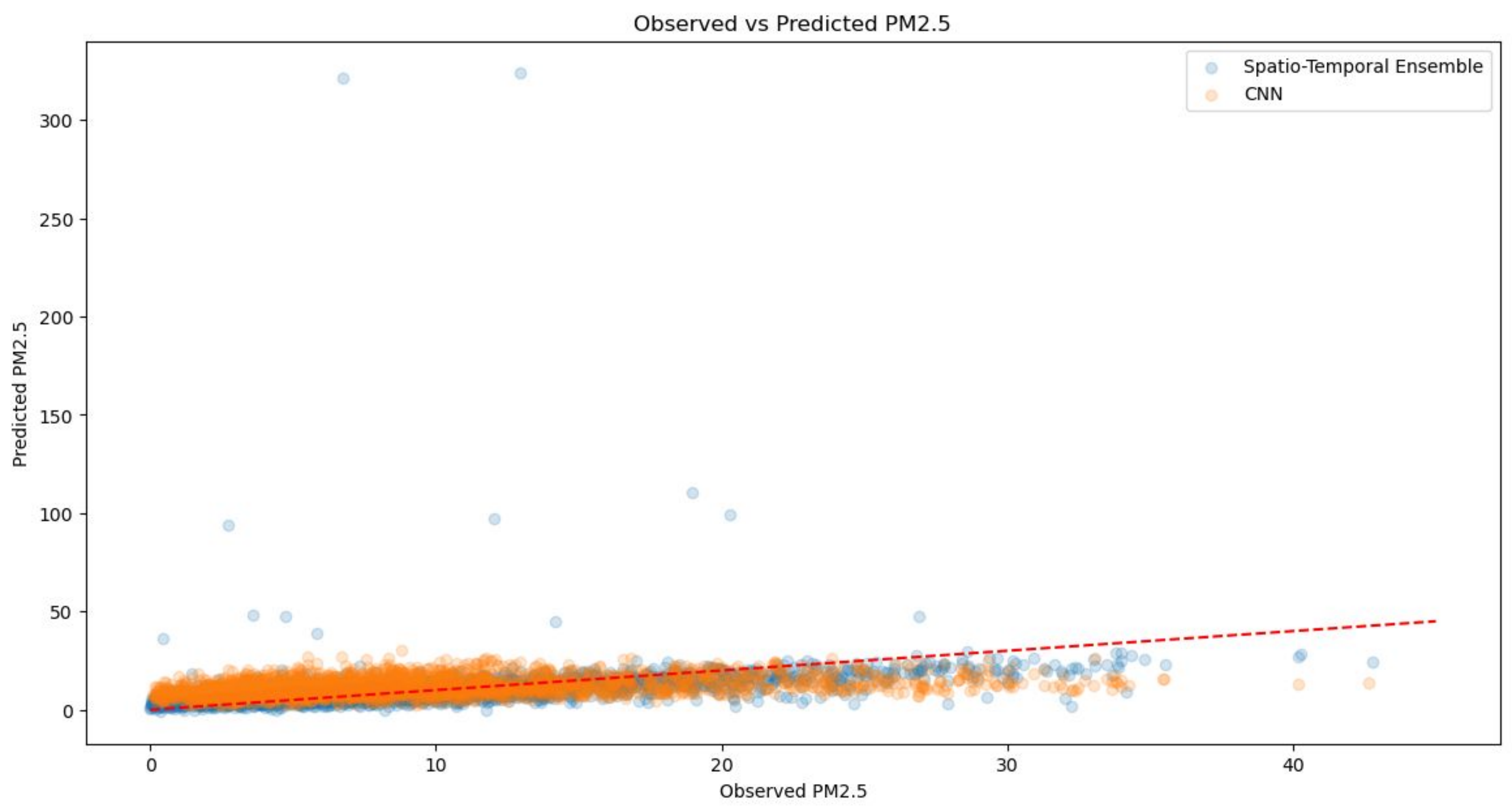
- Create sequences of spatial maps for input into the 3D CNN.
- Split the data into training and testing sets.

Model Training:

- Define and compile the 3D CNN architecture.
- Train the model on the training set.

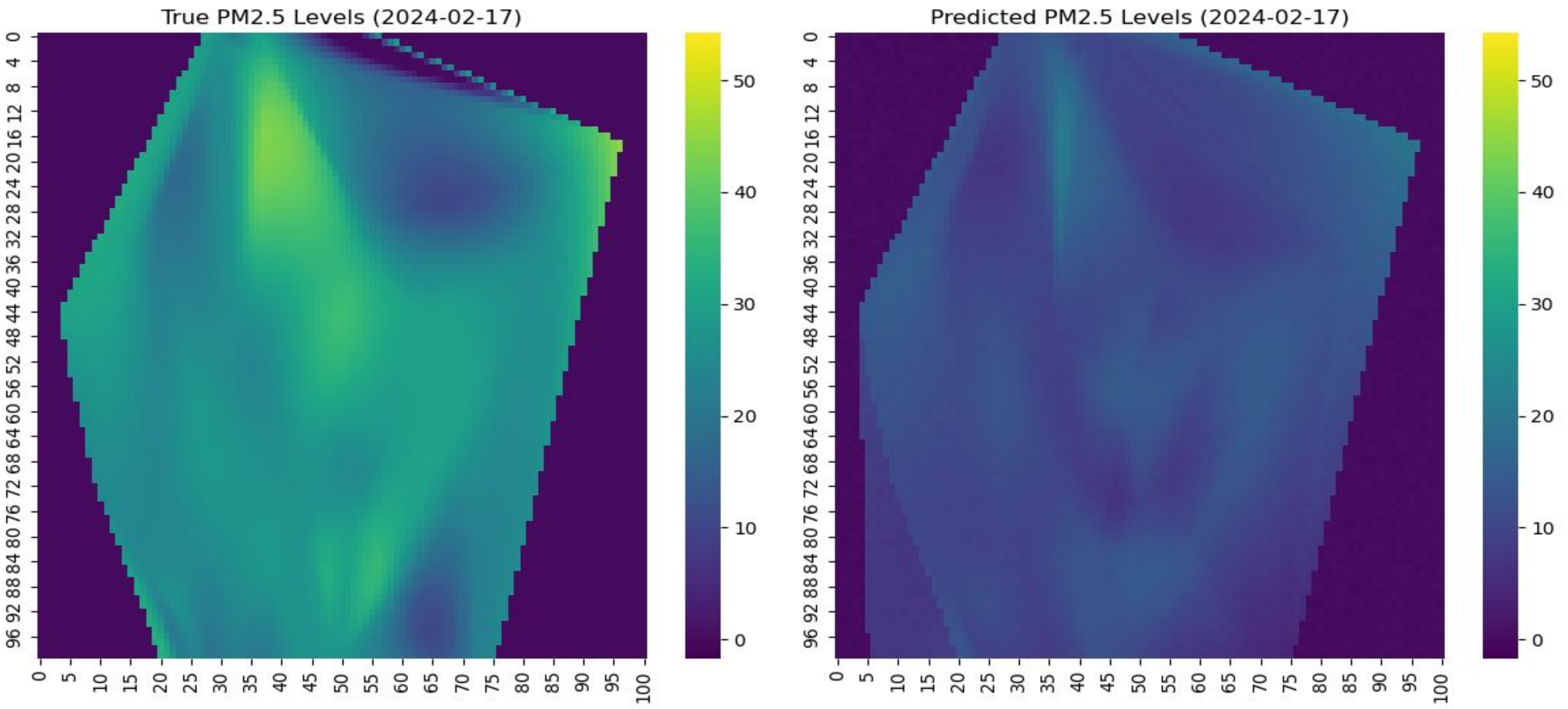


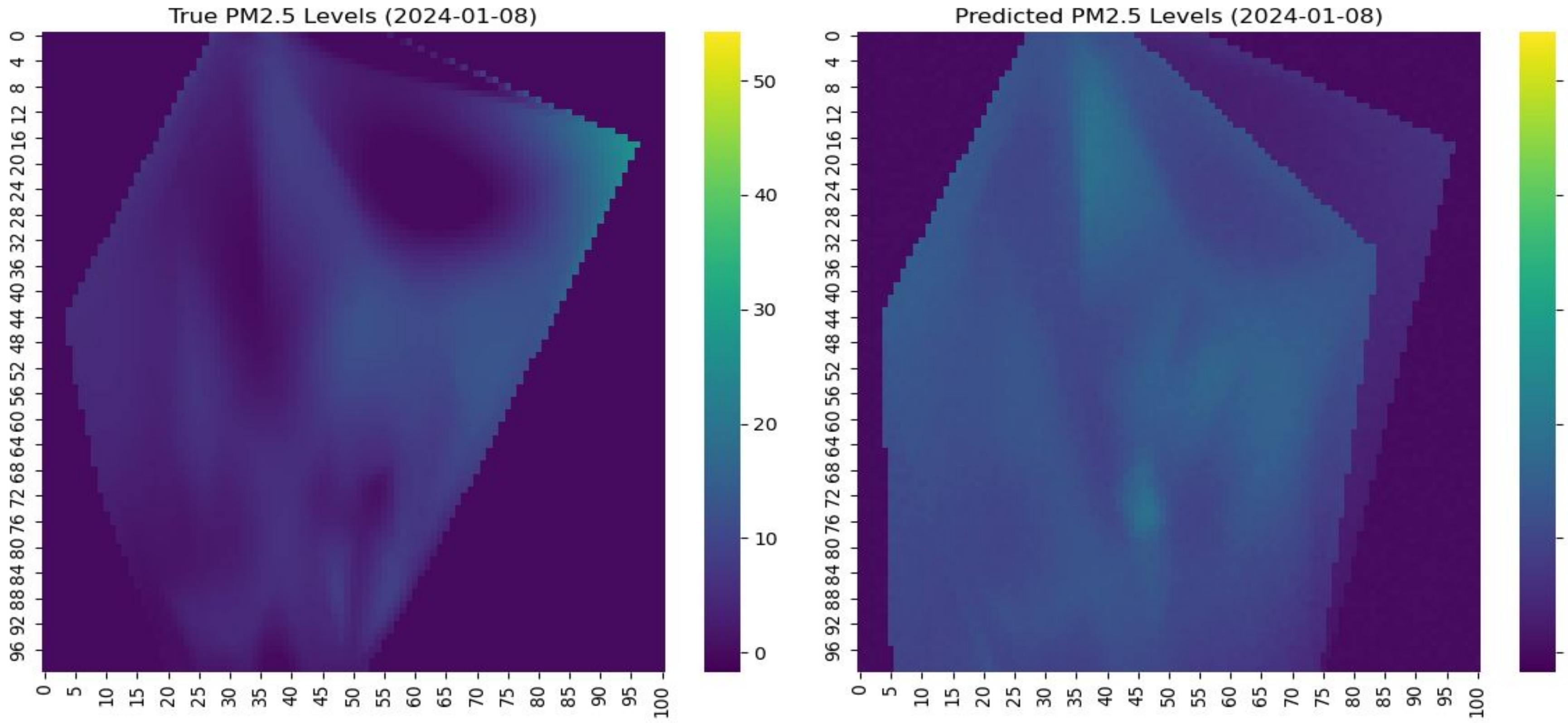

Time Series of PM2.5 for Sensor 123453
Observed_extrapolation
CNN Predicted
PM2.5
Date
Time Series of PM2.5 for Sensor 99595
Observed_extrapolation
CNN Predicted
PM2.5
Date
Time Series of PM2.5 for Sensor 87019
Observed_extrapolation
CNN Predicted
PM2.5
2024-01-15
2024-02-01
2024-02-15
2024-03-01
2024-03-15
Date
Time Series of PM2.5 for Sensor 120681
Observed_extrapolation
CNN Predicted
PM2.5
2024-01-15
2024-02-01
2024-02-15
2024-03-01
2024-03-15
Date

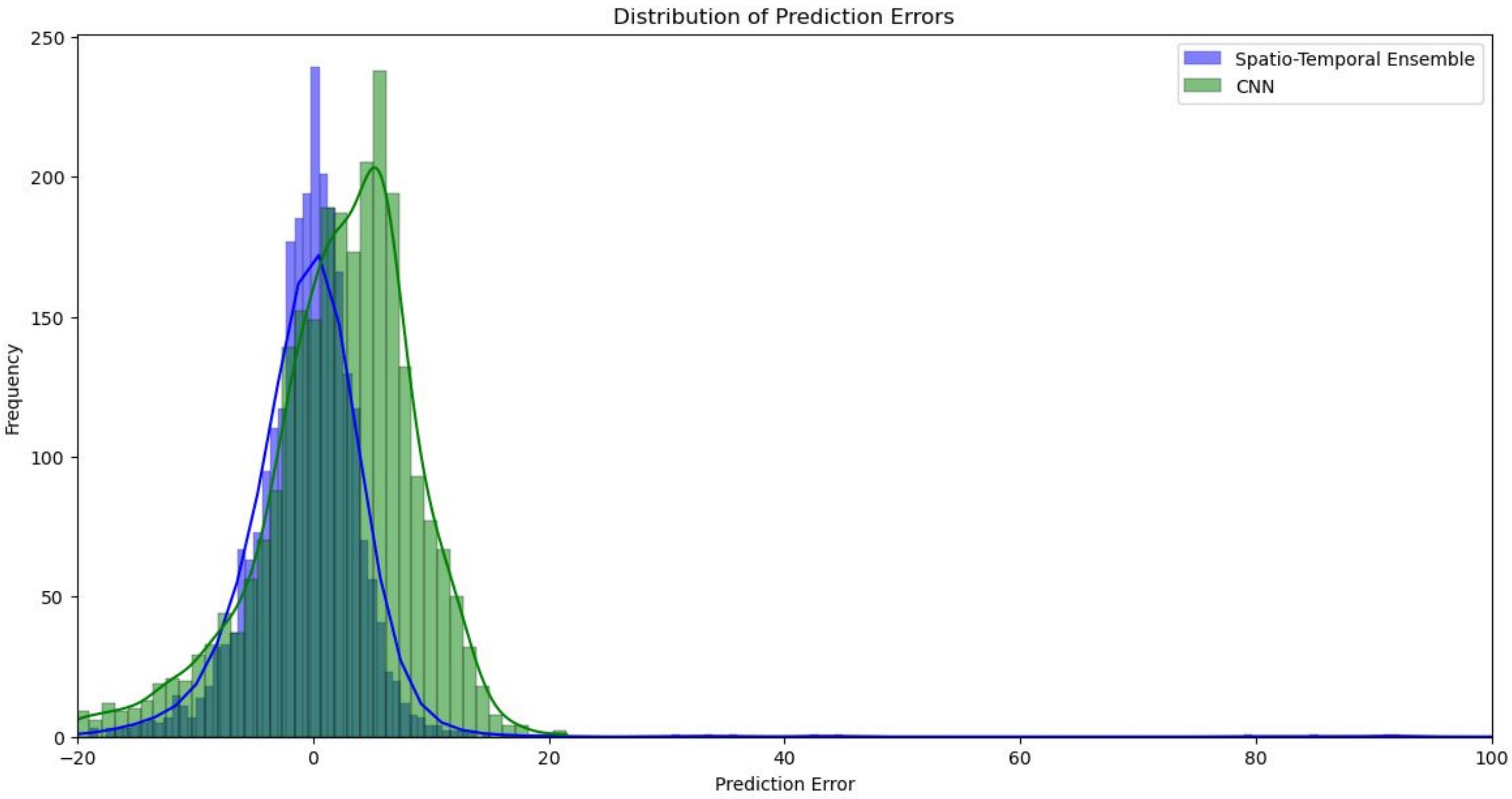


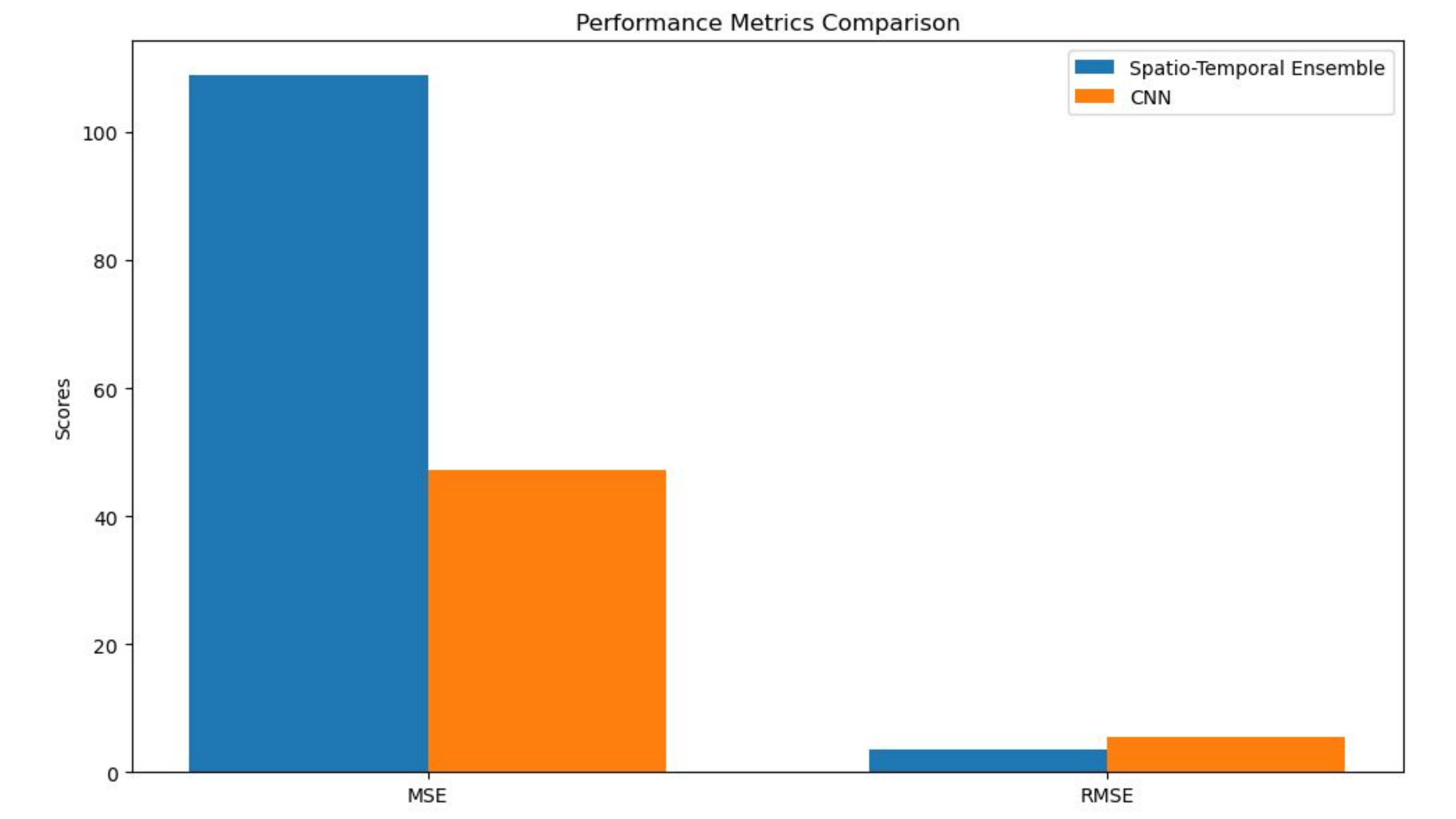
Observed vs Predicted PM2.5
Spatio-Temporal Ensemble
CNN
Predicted PM2.5
Observed PM2.5
0
50
100
150
200
250
300
0
10
20
30
40

3D CNN prediction and observed (interpolated) PM2.values




True PM2.5 Levels (2024-01-08)
Predicted PM2.5 Levels (2024-01-08)
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
0 10 20 30 40 50


Distribution of Prediction Errors
Spatio-Temporal Ensemble
CNN
Frequency
Prediction Error
250
200
150
100
50
0
−20
0
20
40
60
80
100



Performance Metrics Comparison
Spatio-Temporal Ensemble
CNN
Scores
0
20
40
60
80
100
MSE
RMSE

Conclusions:

- XGBoost regression provided the closest prediction to test data
- 3D CNN underperformed compared to spatio-temporal regression
- Average wind speed, the generated lag features (both spatial and time), max and min temperatures, and precipitation most important features in regression
- For ARIMA model, only humidity was found to be significant