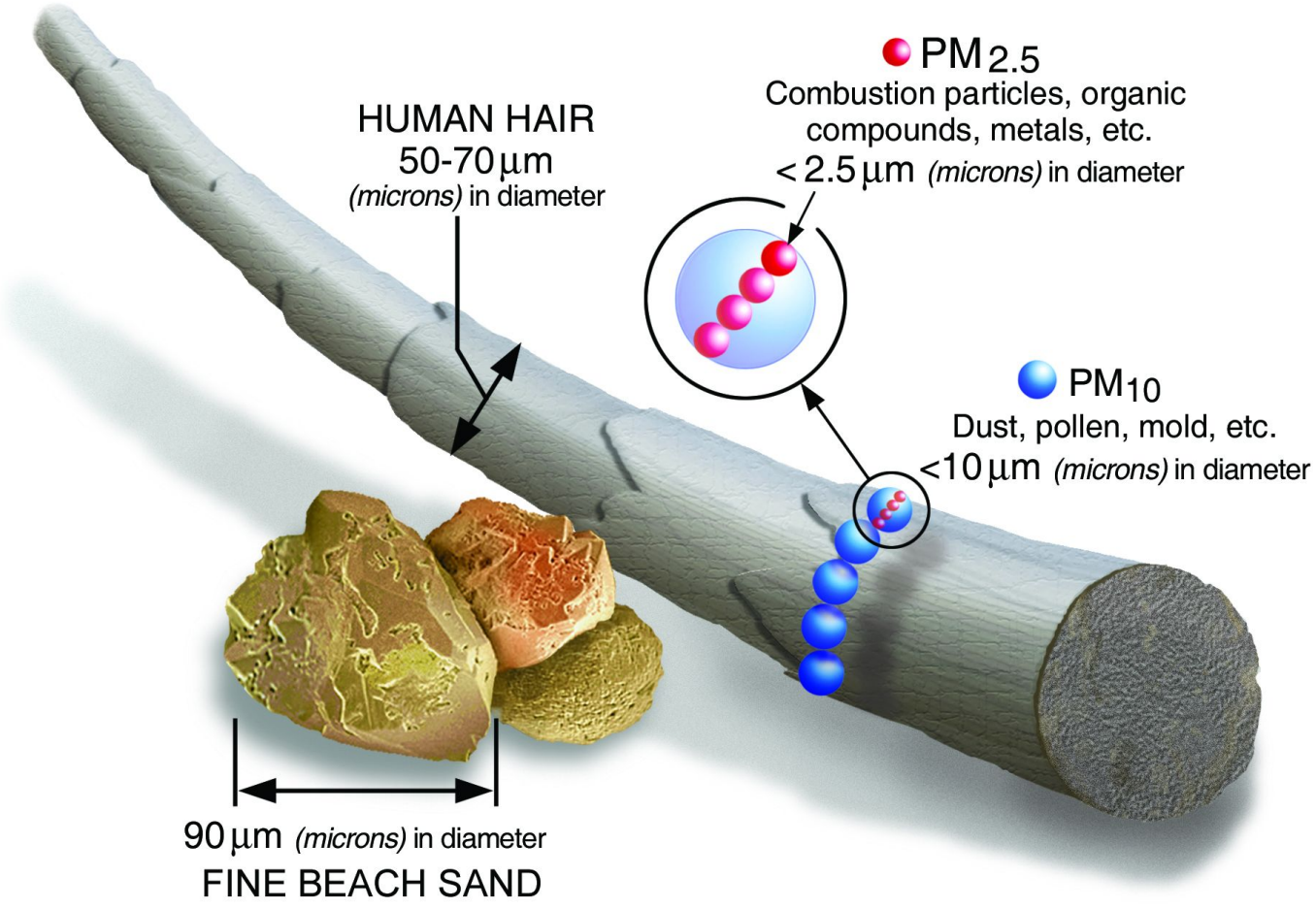
Data Augmentation Techniques for Improved $PM_{2.5}$ Forecasting Using Transformer Architectures

Phoebe Pan (Yilmaz 3)
05/21/2025

Problem

- Air pollution is the second leading global risk factor for death, accounting for 8.1 million deaths in 2021 (State of Global Air)
- Research has shown that breathing in particulate matter with diameters less than 2.5 micrometers (PM2.5) greatly increases risk for cardiovascular and respiratory diseases
- Ground monitoring sites are too sparsely located to capture the spatiotemporal variabilities of PM2.5
- Satellite data is often used to retrieve and predict PM2.5 concentrations as it provides higher spatial resolution


HUMAN HAIR
50-70 µm
(microns) in diameter
PM2.5
Combustion particles, organic
compounds, metals, etc.
< 2.5 µm (microns) in diameter
PM10
Dust, pollen, mold, etc.
<10 µm (microns) in diameter
90 µm (microns) in diameter
FINE BEACH SAND

Other Solutions

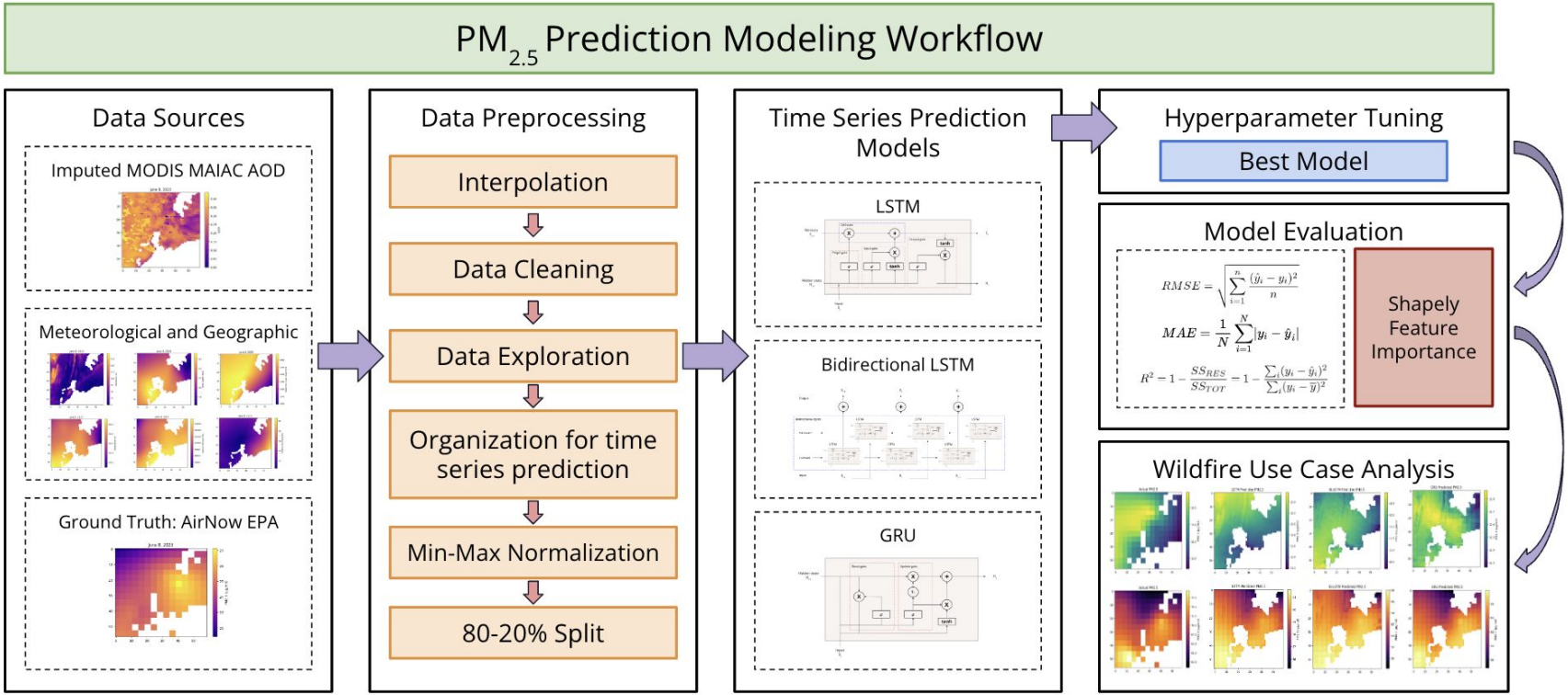
- At first, statistical models such as linear regression and ARIMA (autoregressive integrated moving average) have been used to predict PM2.5 (Graupe et al. 1975; Cekim 2020; Jian et al. 2012; Abedi et al. 2020)
- Simple machine learning methods such as SVR, Random Forest, and XGBoost have also been tested (Chu et al. 2021; Yang et al. 2018; Agarwal et al. 2020)
- Later on, deep learning techniques such as CNN, RNN, and LSTM offered more advanced learning capabilities (Hinton & Salakhutdinov 2006; LeCun et al. 2015)

Previous Work

Objectives

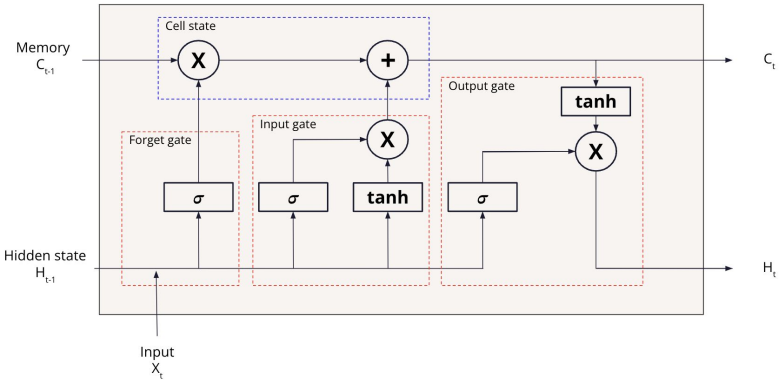
1. PM2.5 Prediction: Leverage deep learning models to capture the temporal and spatial dependencies of PM2.5 and model its intricate relationships with meteorological and geographical variables
2. Time Series Deep Learning Models: Compare prediction performances of LSTM, Bi-LSTM and GRU using data from New York City
3. 2023 Canadian Wildfires Use Case: Predict PM2.5 levels in the New York City during the wildfires to facilitate public health policies and measures against air pollution

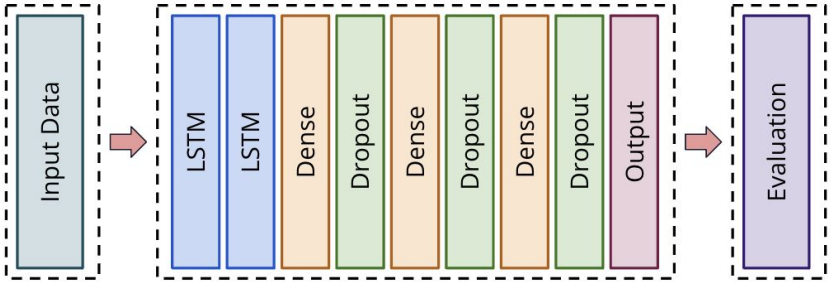
Methodology



Long Short-Term Memory (LSTM)

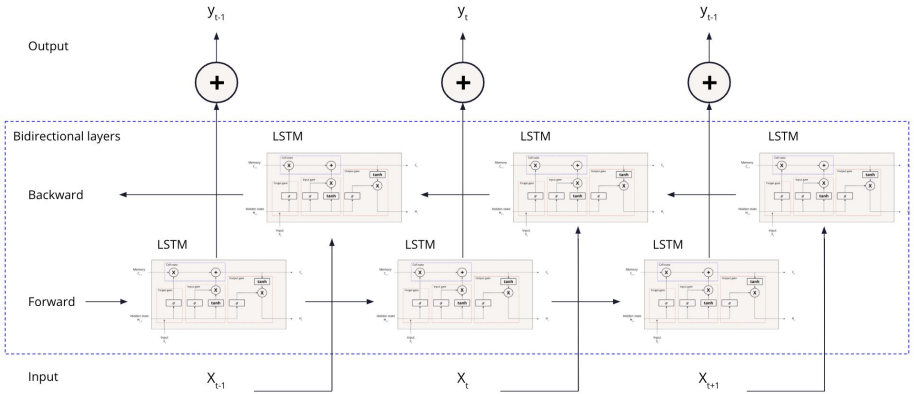
- Designed to handle long-range dependencies and mitigate the vanishing gradient problem in traditional RNNs (Hochreiter, 1997)
- Consists of three gates: forget, input, and output
- Shown by previous studies to be effective in air pollution prediction while using continuous historical data (Kim et al., 2022)

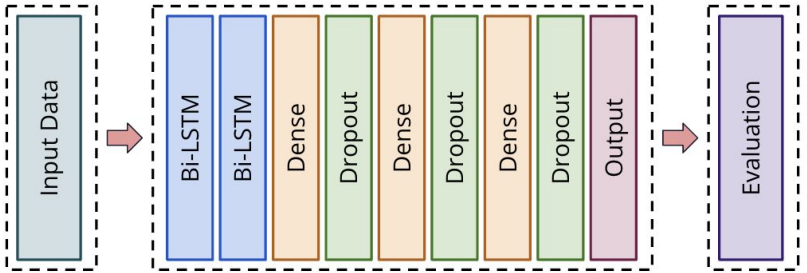




Bidirectional LSTM

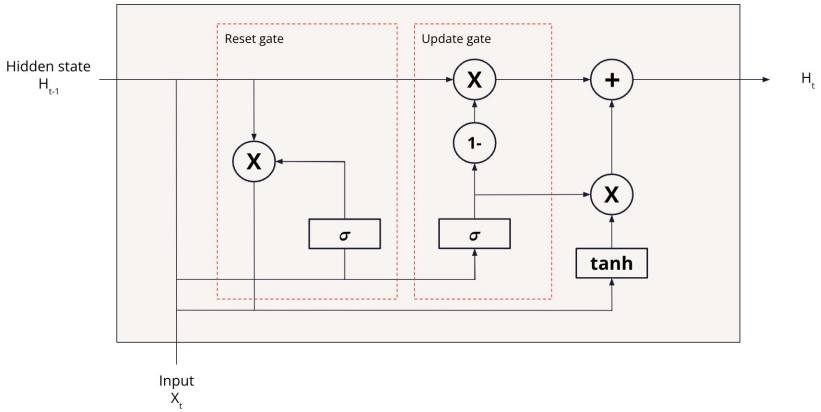
- Input sequence is processed in two directions: from the beginning to the end and from the end to the beginning (Graves, 2005)
- Consists of a forward layer and a backward layer which are concatenated together in the end
- Performance does not deteriorate for lengthy datasets (Kim et al., 2022)

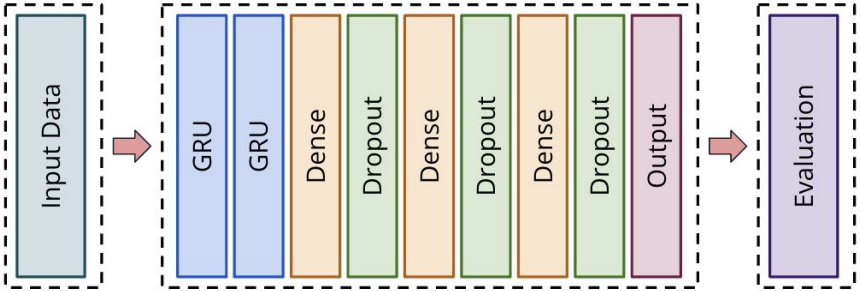




Gated Recurrent Unit (GRU)

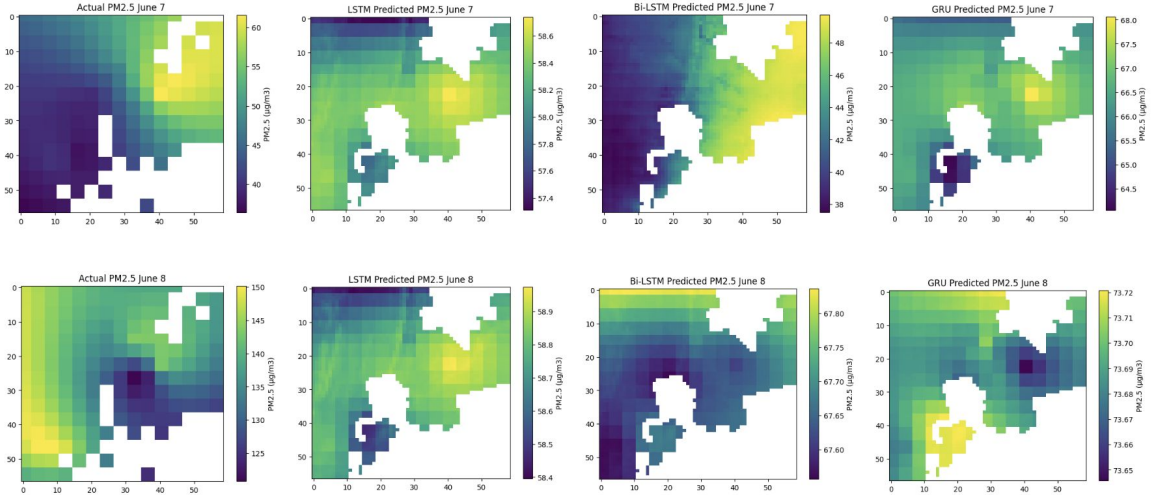
- Simplifies the LSTM architecture while maintaining performance (Cho, 2014)
- Greater operating speed makes GRUs faster and easier to train -> more optimal in PM2.5 prediction scenarios (Kim et al., 2022)
- Two gates: reset gate and update gate





Results

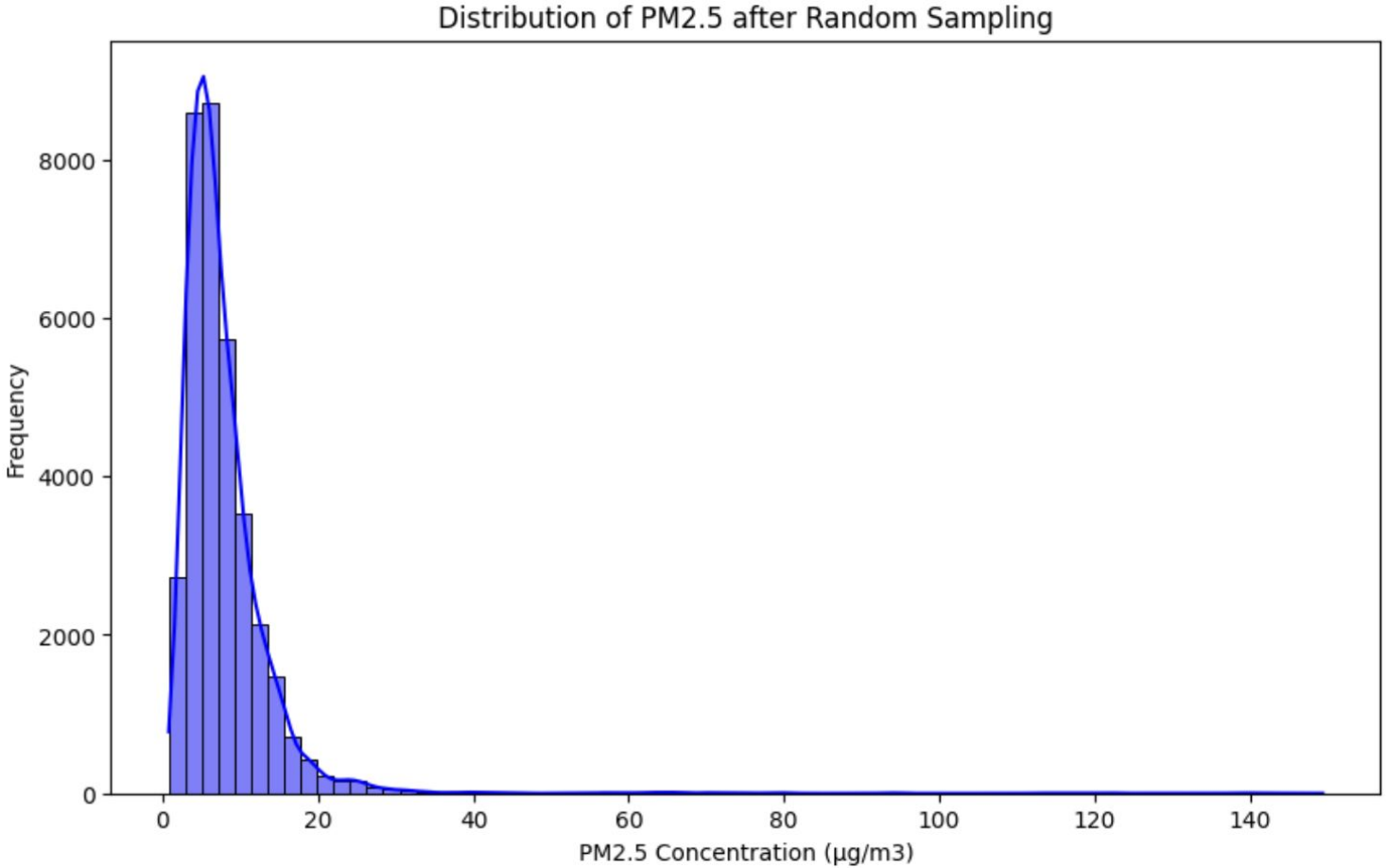
- Compared to the LSTM and Bi-LSTM, the GRU performed marginally better
- The three models were able to capture the pattern of PM2.5 pretty well
- However, they struggled with really high observations during extreme air pollution scenarios

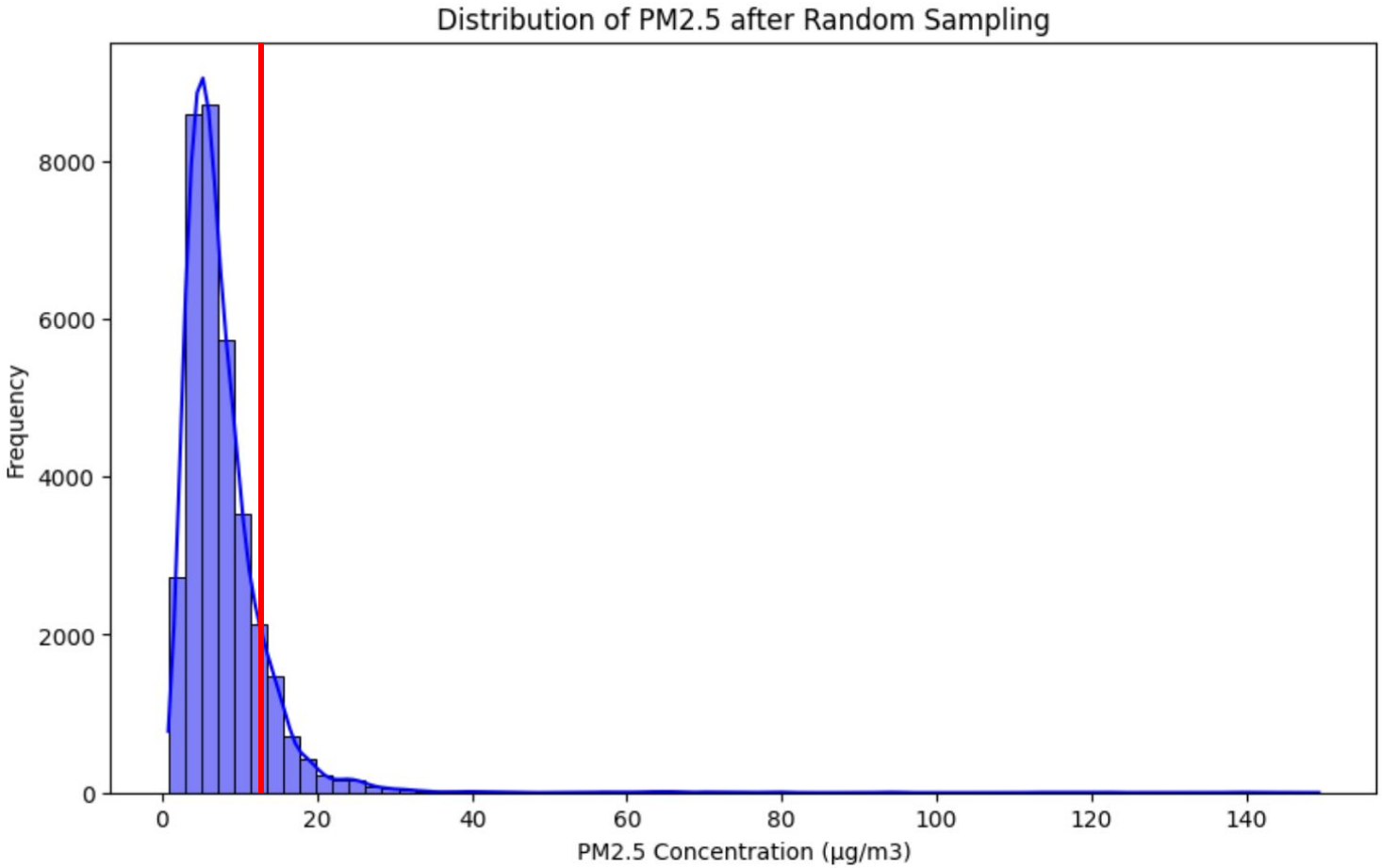


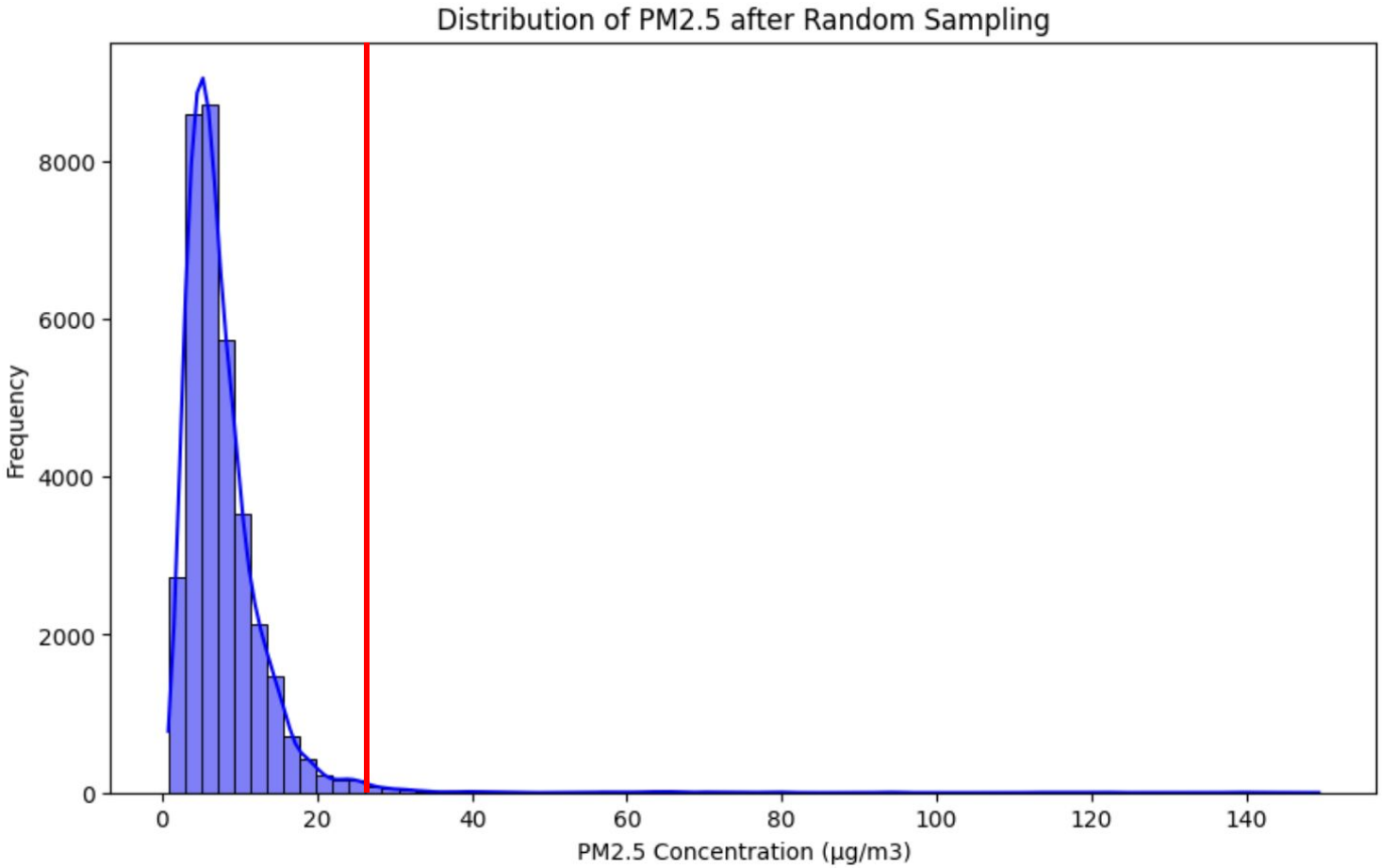
This Project

Objectives

1. Investigate the impact of cutoff thresholds based on limits set by the Environmental Protection Agency (EPA) on model performance
2. Augment imbalance PM2.5 forecasting data using cluster-based under sampling with different partial sampling ratios of minority to majority classes to find the most optimal ratio.
3. Build and train a transformer model to leverage multi-head attention for time series forecasting
4. Design and implement a website where users can explore the effects of partial sampling ratio and cutoff thresholds
    a. Input their own data to experiment with to determine optimal data augmentation procedure


Distribution of PM2.5 after Random Sampling
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
120
140
PM2.5 Concentration (µg/m3)


Distribution of PM2.5 after Random Sampling
Frequency
8000
6000
4000
2000
0
0
20
40
60
80
100
120
140
PM2.5 Concentration (µg/m3)


Distribution of PM2.5 after Random Sampling
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
120
140
PM2.5 Concentration (µg/m3)

Objectives

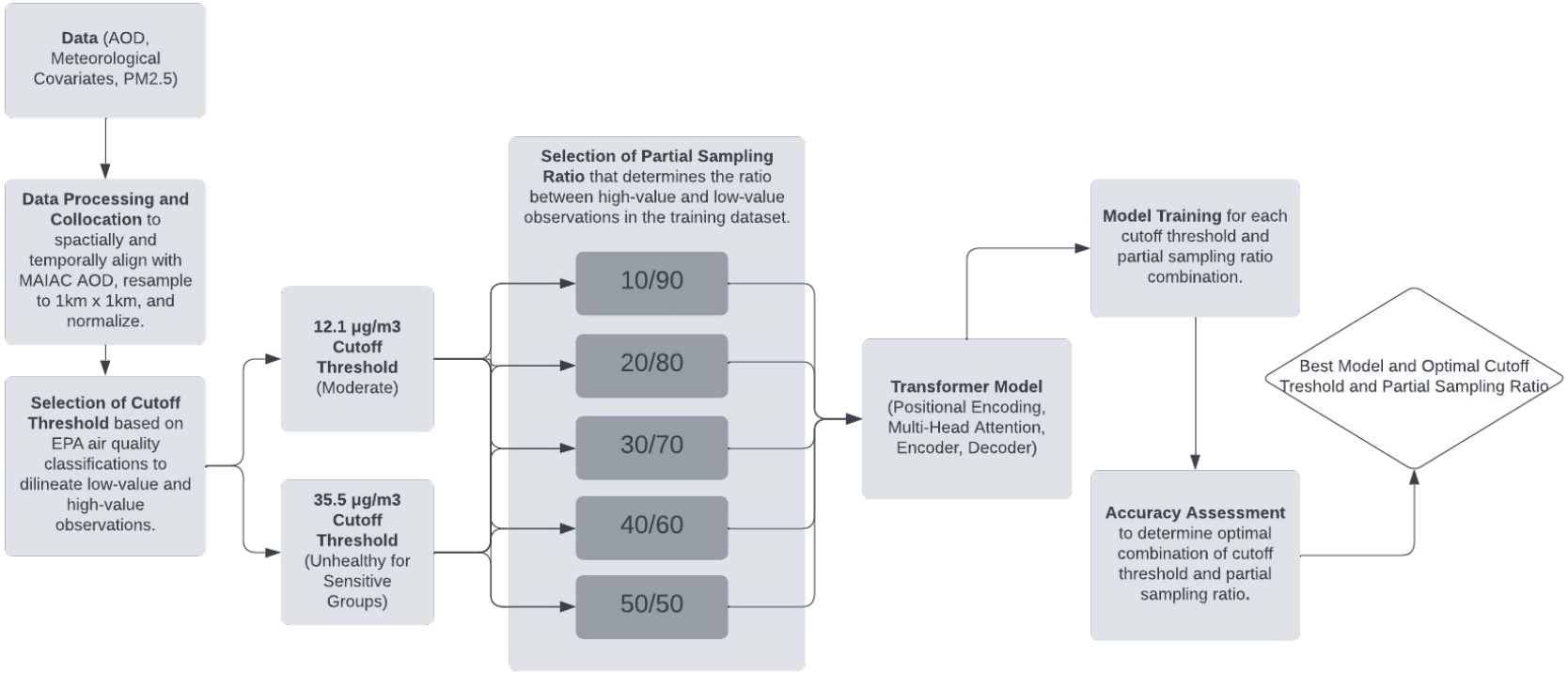
1. Investigate the impact of cutoff thresholds based on limits set by the Environmental Protection Agency (EPA) on model performance
2. Augment imbalance PM2.5 forecasting data using cluster-based under sampling with different partial sampling ratios of minority to majority classes to find the most optimal ratio.
3. Build and train a transformer model to leverage multi-head attention for time series forecasting
4. Design and implement a website where users can explore the effects of partial sampling ratio and cutoff thresholds
    a. Input their own data to experiment with to determine optimal data augmentation procedure

Impact

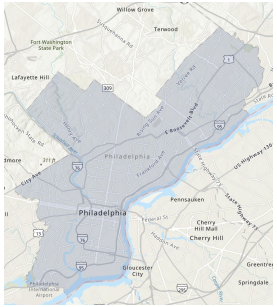
- Everybody needs to breathe air
- More accurate forecasting helps people with underlying conditions
- Helps prevent serious conditions in the whole population
- Wildfires are only going to become more frequent due to climate change
- Introduce data augmentation to the field of air quality prediction
- Better environmental and health policies

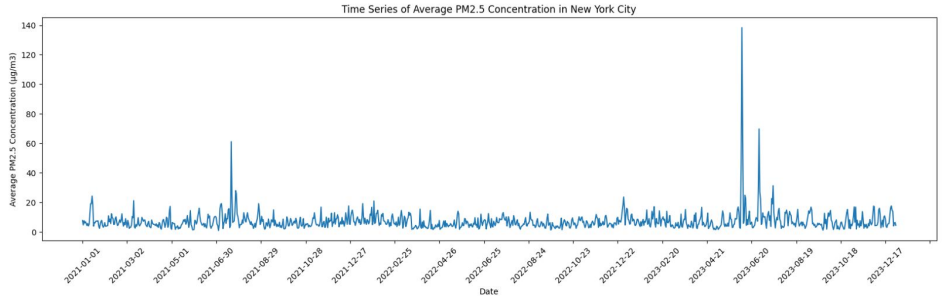


Methodology: Workflow Diagram



Methodology: Study Area





- New York City, New York
  - Population: 8.336 million
  - Land Area: 790 square km (302.6 square miles)
  - Coordinates: 40.4774° N, -74.2591° W (southwest) to 40.9176° N, -73.7004° W (northeast)
- Philadelphia, Pennsylvania
  - Population: 1.567 million
  - Land Area: 347.52 square km (134.18 square miles)
  - Coordinates: 39.8670° N, -75.2803° W (southwest) to 40.1379° N, -74.9558° W (northeast)
- Washington DC
  - Population: 671,803
  - Land Area: 176 square km (68 square miles)
  - Coordinates: 38.7916° N, -77.1198° W (southwest) to 38.9955° N, -76.9094° W (northeast)
- Correlation between smoke particles and number of patients showing up with asthma-related symptoms in New York City (Chen et al., 2023)

Methodology: Data

Variable	Source	Unit
AOD	MODIS MAIAC (Terra and Aqua)	
Boundary Layer Height (BLH)	ECMWF ERA5-hourly	meter
Relative Humidity	ECMWF ERA5-hourly	%
Temperature (at 2m)	ECMWF ERA5-hourly	K
Surface Pressure	ECMWF ERA5-hourly	Pa
Wind Speed	ECMWF ERA5-hourly	m/s
Elevation	USGS	meter

Methodology: Cutoff Threshold

- Yin et al., 2022 defined 75 ug/m3 are their cutoff to determine low and high value samples
- Study area in China with historically higher PM2.5 levels
- The Environmental Protection Agency defines limits for PM2.5 pollution categories in the US
- Does the cutoff threshold change results and which is optimal?

$PM_{2.5}$	Air Quality Index	$PM_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	Good 0 to 50	Little to no risk.	None.
12.1 to 35.4	Moderate 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.

Methodology: Partial Sampling Ratio

- Partial sampling ratio refers to a minority/majority or in our case low/high ratio between the original dataset and 50/50
- Limited research into most optimal partial sampling ratio
- Kamalov et al., 2022 found 0.75 (minority/majority, ~44/56) to be most optimal after a systematic study
- Yin et al., 2022 found 30/70 to be most optimal

Methodology: Cluster Based Under Sampling

- Data points are grouped into clusters using k-means algorithm
- Within each cluster, same number of samples are randomly selected
- Control ratio of low-value to high-value samples
- Ensure sampling is more representative

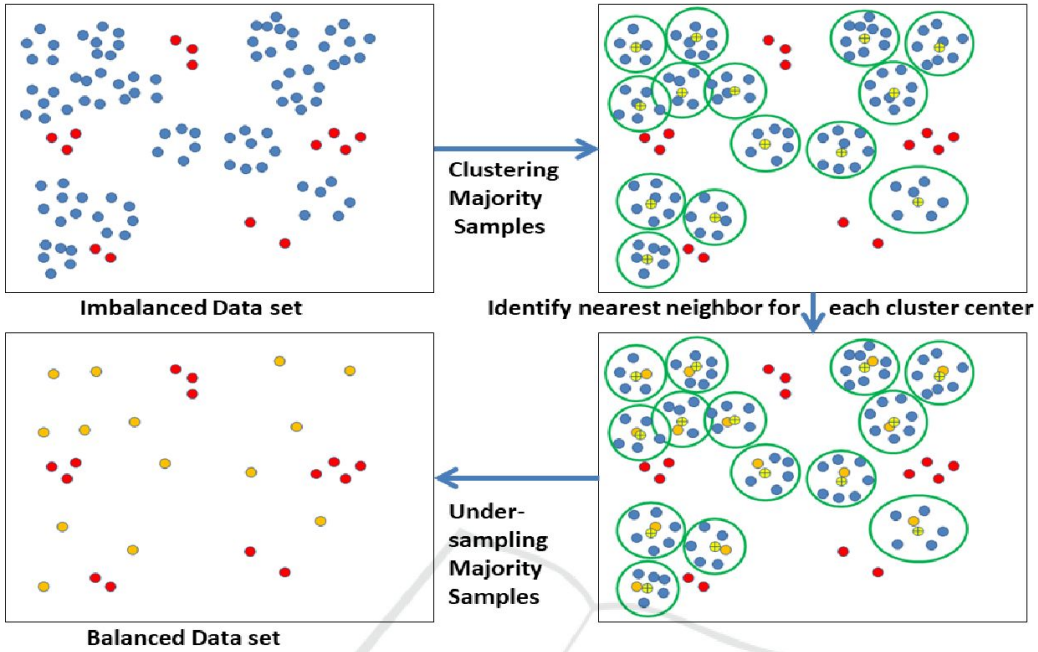
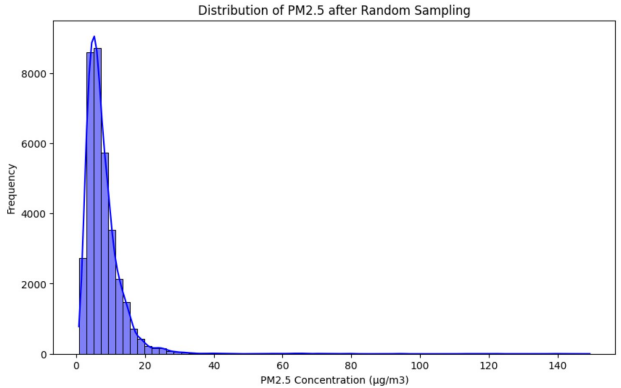
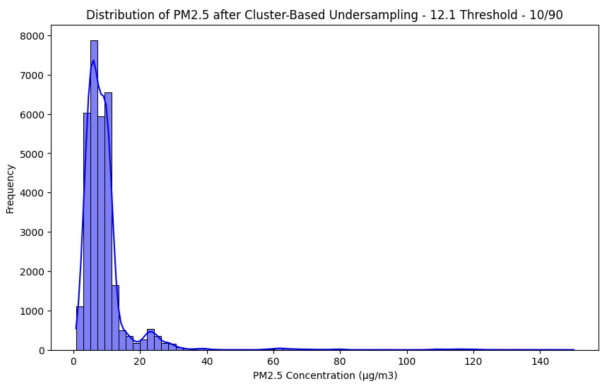


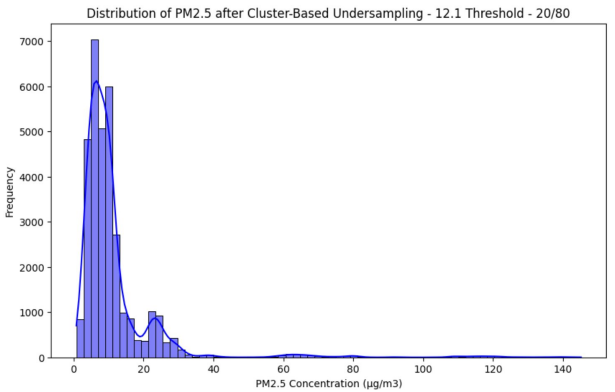
Figure 1: Clustering-based Under-sampling majority samples.

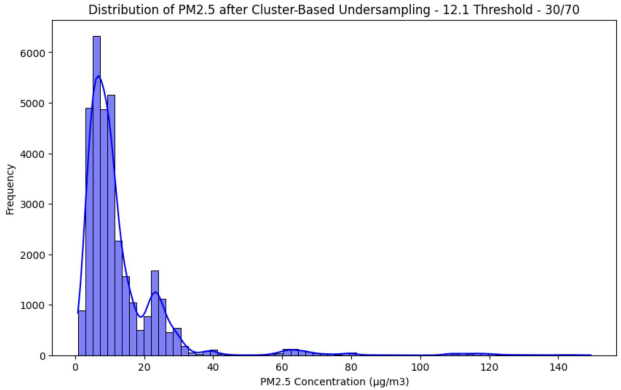
Source: [Henein et al., 2018]

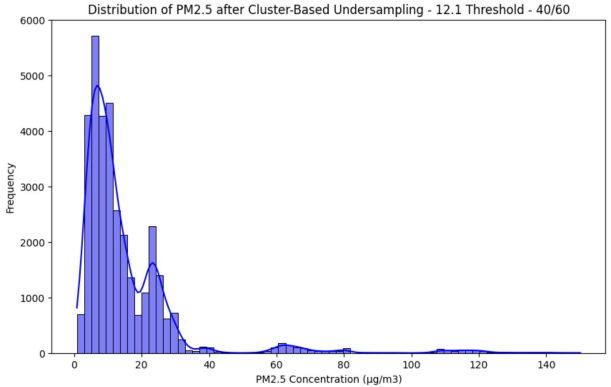
Methodology: Partial Sampling Ratio - 12.1 ug/m3

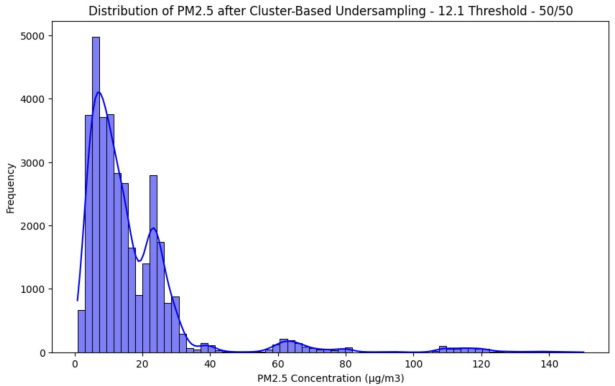




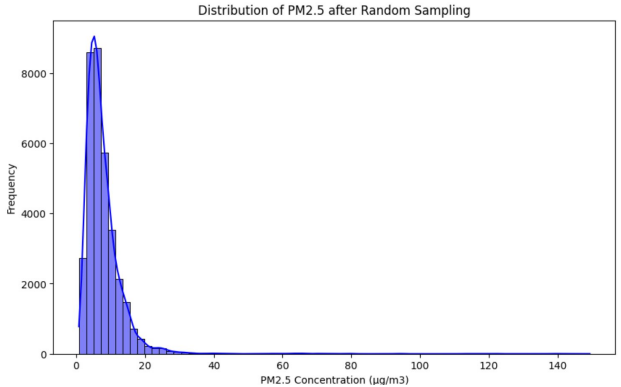


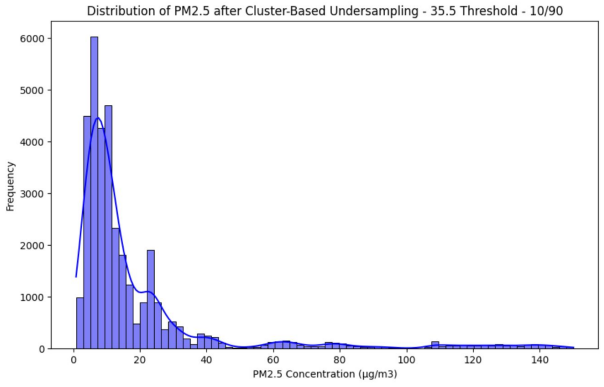


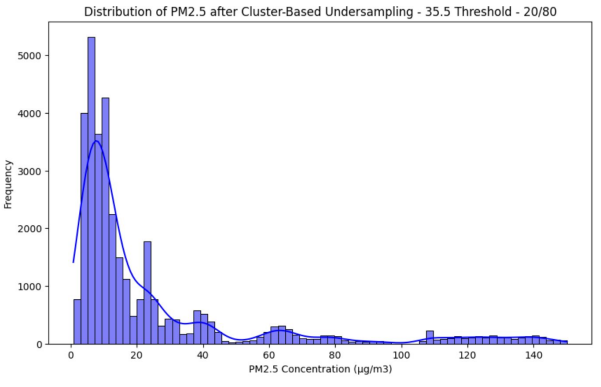


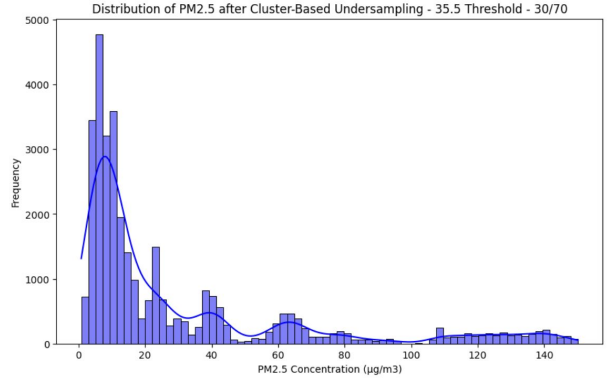


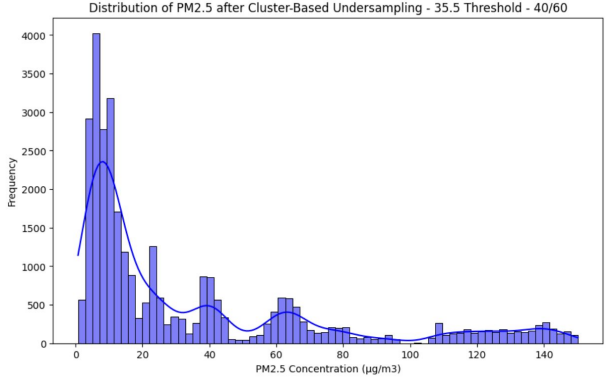
Methodology: Partial Sampling Ratio - 35.5 ug/m3

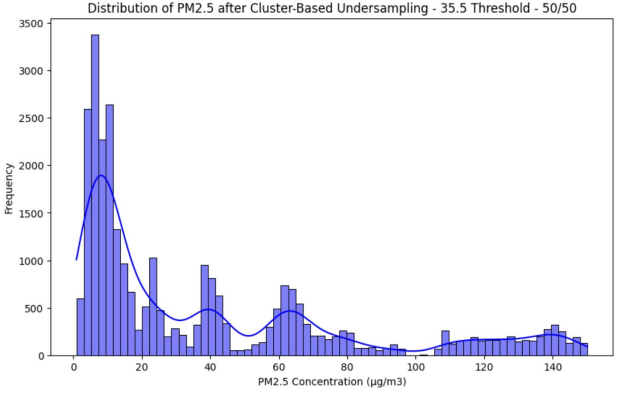






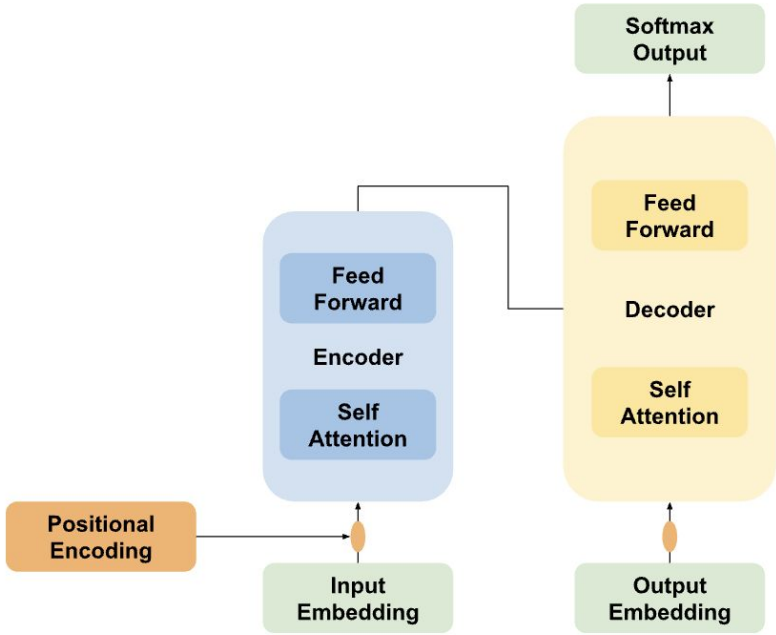






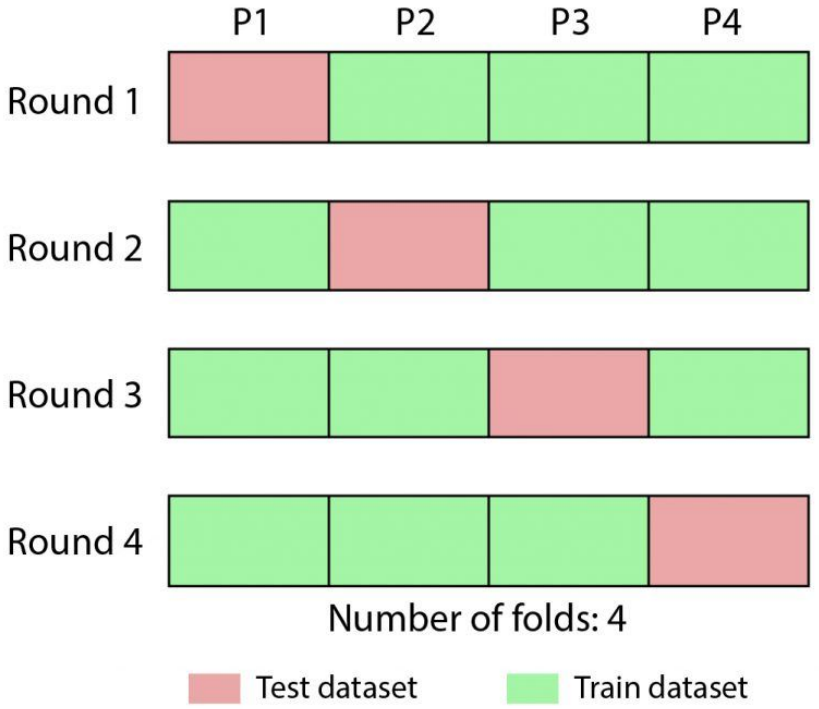
Methodology: Transformer Architecture

- Revolutionized natural language processing and time series predictions
- Positional encoding
- Multi-head attention
- Encoder
- Decoder



k-Fold Cross Validation

- Technique used in machine learning to prevent overfitting
- Increases generalizability of model to new data
- Provides more accurate measurements of model performance
- Variable number of folds



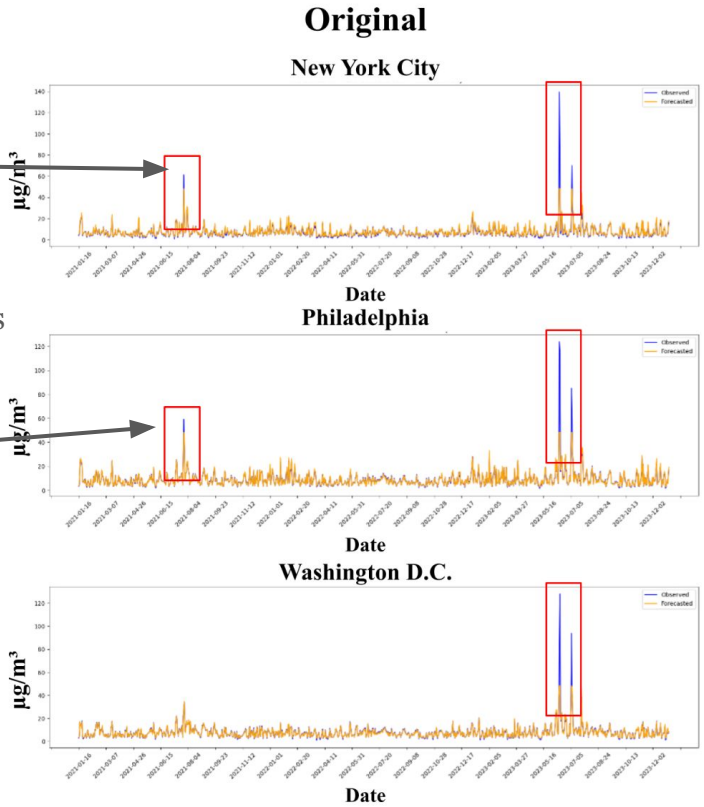
k-Fold (k = 5) Cross Validation Results - 12.1 ug/m3

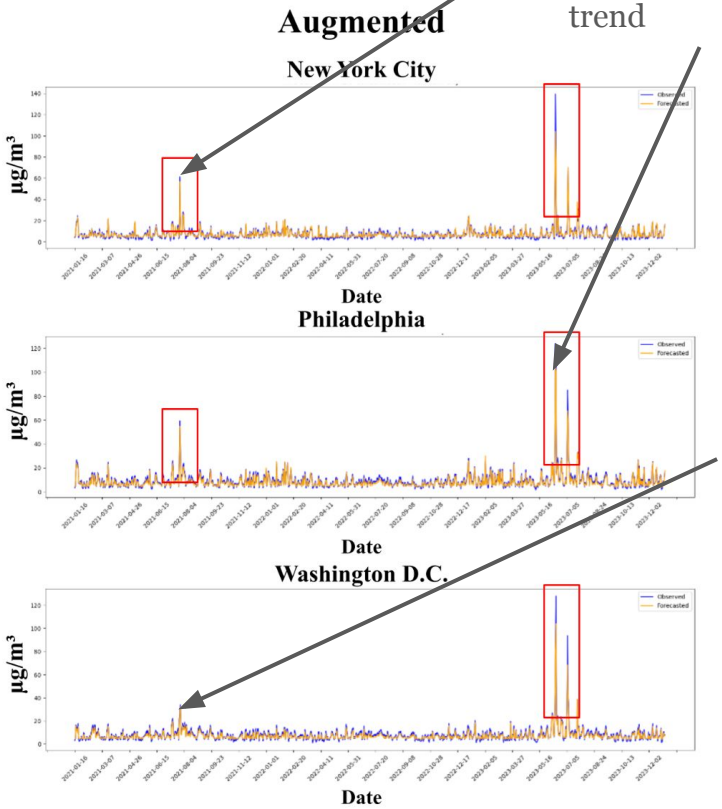
12.1 ug/m3 Cutoff Threshold						
Partial Sampling Ratio	Whole Dataset			High Values		
	RMSE	MAE	R-squared	RMSE	MAE	R-squared
Original	3.247 ± 0.101	0.676 ± 0.021	0.819 ± 0.025	32.748 ± 1.014	27.318 ± 0.846	0.037 ± 0.001
10/90	3.291 ± 0.102	0.743 ± 0.023	0.743 ± 0.023	30.040 ± 0.930	20.750 ± 0.642	0.192 ± 0.006
20/80	3.161 ± 0.098	1.171 ± 0.036	0.831 ± 0.026	26.544 ± 0.822	19.481 ± 0.603	0.374 ± 0.012
30/70	2.888 ± 0.089	1.570 ± 0.049	0.862 ± 0.027	25.822 ± 0.800	19.259 ± 0.596	0.409 ± 0.013
40/60	2.881 ± 0.089	1.355 ± 0.042	0.862 ± 0.027	23.818 ± 0.737	17.782 ± 0.551	0.501 ± 0.016
50/50	2.820 ± 0.087	1.068 ± 0.033	0.870 ± 0.027	21.776 ± 0.674	14.438 ± 0.447	0.587 ± 0.018

k-Fold (k = 5) Cross Validation Results - 35.5 ug/m3

35.5 ug/m3 Cutoff Threshold						
Partial Sampling Ratio	Whole Dataset			High Values		
	RMSE	MAE	R-squared	RMSE	MAE	R-squared
Original	3.247 ± 0.103	1.629 ± 0.050	0.819 ± 0.025	32.748 ± 1.014	27.318 ± 0.846	0.037 ± 0.001
10/90	2.334 ± 0.072	1.629 ± 0.050	0.918 ± 0.028	20.200 ± 0.625	14.127 ± 0.437	0.648 ± 0.020
20/80	2.128 ± 0.066	1.418 ± 0.044	0.935 ± 0.029	15.705 ± 0.486	10.308 ± 0.319	0.796 ± 0.025
30/70	2.359 ± 0.073	1.709 ± 0.053	0.916 ± 0.028	16.464 ± 0.510	12.484 ± 0.387	0.773 ± 0.024
40/60	2.479 ± 0.077	1.766 ± 0.055	0.904 ± 0.028	16.936 ± 0.524	13.213 ± 0.409	0.758 ± 0.023
50/50	2.738 ± 0.085	1.918 ± 0.059	0.878 ± 0.027	19.555 ± 0.605	14.650 ± 0.454	0.671 ± 0.021

Time Series Analysis



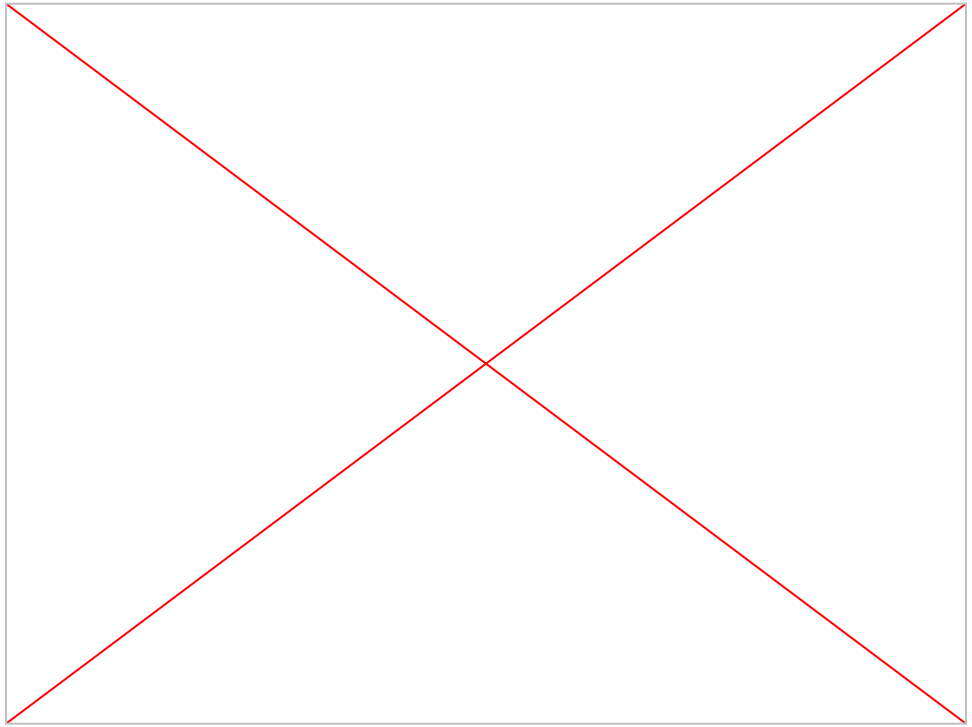


Does not capture peak magnitudes during periods of severe pollution.

Much better at reaching the high values and capturing the trend

Lose a bit of accuracy on the lower values but it still predicts pretty well

Website Demo



Limitations

- Generalizability: since data from only three cities was used, the model cannot be applied to other locations without further training
- Bias: while data augmentation improves forecasting of high-sample values, because it is not an exact representation of the original dataset, new biases can be introduced
  - Tradeoff between accuracy and performance on high-values
- Multistep prediction: because forecasts build on previous forecasts, inaccuracies can be propagated and amplified

Conclusion and Future Work

- Increase the size and scope of the dataset to improve generalizability
- Transfer learning or creating a base model that can be tuned with small datasets
- Uncertainty quantification to measure bias from data augmentation and multistep prediction

References

Abedi, A., Baygi, M. M., Poursafa, P., Mehrara, M., Amin, M. M., Hemami, F., & Zarean, M. (2020). Air pollution and hospitalization: an autoregressive distributed lag (ARDL) approach. Environmental Science and Pollution Research, 27(24), 30673–30680. https://doi.org/10.1007/s11356-020-09152-x

Agarwal, S., Sharma, S., R, S., Rahman, M. H., Vranckx, S., Maiheu, B., Blyth, L., Janssen, S., Gargava, P., Shukla, V., & Batra, S. (2020). Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. The Science of the Total Environment, 735, 139454. https://doi.org/10.1016/j.scitotenv.2020.139454

Cekim, H. O. (2020). Forecasting PM10 concentrations using time series models: a case of the most polluted cities in Turkey. Environmental Science and Pollution Research, 27(20), 25612–25624. https://doi.org/10.1007/s11356-020-08164-x

Chen K, Ma Y, Bell ML, Yang W. Canadian Wildfire Smoke and Asthma Syndrome Emergency Department Visits in New York City. JAMA. 2023 Oct 10;330(14):1385-1387. doi: 10.1001/jama.2023.18768. PMID: 37733685; PMCID: PMC10514869.

Chu, J., Dong, Y., Han, X. et al. Short-term prediction of urban PM2.5 based on a hybrid modified variational mode decomposition and support vector regression model. Environ Sci Pollut Res 28, 56–72 (2021). https://doi.org/10.1007/s11356-020-11065-8

Graupe, D., Krause, D., & Moore, J. (1975). Identification of autoregressive moving-average parameters of time series. IEEE Transactions on Automatic Control, 20(1), 104–107. https://doi.org/10.1109/tac.1975.1100855

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786), 504–507. https://doi.org/10.1126/science.1127647

References

Hystad, Perry et al. Associations of outdoor fine particulate air pollution and cardiovascular disease in 157 436 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *The Lancet Planetary Health*, Volume 4, Issue 6, e235 - e245

Jian, L., Zhao, Y., Zhu, Y., Zhang, M., & Bertolatti, D. (2012). An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. The Science of the Total Environment, 426, 336–345. https://doi.org/10.1016/j.scitotenv.2012.03.025

Kamalov, Firuz, Amir F. Atiya, Dina Elreedy. Partial Resampling of Imbalanced Data. 2022. arXiv:2207.04631

Lao, X.Q., Guo, C., Chang, Ly. et al. Long-term exposure to ambient fine particulate matter (PM2.5) and incident type 2 diabetes: a longitudinal cohort study. Diabetologia 62, 759–769 (2019). https://doi.org/10.1007/s00125-019-4825-1

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

Linjiong Liu, Yuanyuan Zhang, Zhiming Yang, Siqi Luo, Yunquan Zhang. Long-term exposure to fine particulate constituents and cardiovascular diseases in Chinese adults. *Journal of Hazardous Materials*, Volume 416, 2021, https://doi.org/10.1016/j.jhazmat.2021.126051.

Sharma, A., Valdes, A. C. F., & Lee, Y. (n.d.). Impact of Wildfires on Meteorology and Air Quality (PM2.5 and O3) over Western United States during September 2017. Atmosphere, 13(2), 262. https://doi.org/10.3390/atmos13020262

References

Spracklen, D. V., Mickley, L. J., Logan, J. A., Hudman, R. C., Yevich, R., Flannigan, M. D., & Westerling, A. L. (2009). Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. Journal of Geophysical Research Atmospheres, 114(D20). https://doi.org/10.1029/2008jd010966

Suji Lee, Whanhee Lee, Dahye Kim, Ejin Kim, Woojae Myung, Sun-Young Kim, Ho Kim. Short-term PM2.5 exposure and emergency hospital admissions for mental disease, Environmental Research, Volume 171, 2019, Pages 313-320, https://doi.org/10.1016/j.envres.2019.01.036.

Thangavel, P.; Park, D.; Lee, Y.-C. Recent Insights into Particulate Matter ($PM_{2.5}$)-Mediated Toxicity in Humans: An Overview. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7511. https://doi.org/10.3390/ijerph19127511

Xu Gao, Petros Koutrakis, Brent Coull, Xihong Lin, Pantel Vokonas, Joel Schwartz, Andrea A. Baccarelli. Short-term exposure to PM2.5 components and renal health: Findings from the Veterans Affairs Normative Aging Study, *Journal of Hazardous Materials,* Volume 420, 2021, https://doi.org/10.1016/j.jhazmat.2021.126557.

Yang, W., Deng, M., Xu, F., & Wang, H. (2018). Prediction of hourly PM2.5 using a space-time support vector regression model. Atmospheric Environment, 181, 12–19. https://doi.org/10.1016/j.atmosenv.2018.03.015

Yin, S., Li, T., Cheng, X., & Wu, J. (2022). Remote sensing estimation of surface PM2.5 concentrations using a deep learning model improved by data augmentation and a particle size constraint. Atmospheric Environment, 287, 119282. https://doi.org/10.1016/j.atmosenv.2022.119282

Thanks!
Questions?