Ensemble Machine Learning Approach for $PM_{2.5}$ Reconstruction using MERRA-2 and Long-term Analysis (1980-2021) for India

Vikas Kumar, Vasudev Malyan, Manoranjan Sahu, Basudev Biswal

Indian Institute of Technology Bombay, Powai, Mumbai India-400076

INTRODUCTION

- Air pollution is a global concern causing 7 million premature deaths and costing the global economy ~$2.9 trillion every year (WHO, 2021; Health Effects Institute, 2019; IQAir, 2020).
- According to the report by CPCB, India requires ~ 4,000 air quality monitoring stations due to its size, population, and worsening air pollution, but has only ~ 300 continuous monitoring stations (Sengupta, 2003).
- In the recent years, artificial intelligence and machine learning, have been applied to construct datasets of near-surface $PM_{2.5}$, such as neural network (Di et al., 2017), random forest (RF) (Li et al., 2020; Wei et al., 2019; Wei et al., 2020; Wei et al., 2022), the fast space-time Light Gradient Boosting Machine (LGBM) (Wei et al., 2021a; Zhong et al., 2021), and Extreme Gradient Boosting (XGBoost) (Li et al., 2020; Chen et al., 2019).
- The MERRA-2 products developed by the Global Modeling Assimilation Office (GMAO) (Randles et al., 2017) provide relative long-term surface $PM_{2.5}$ mass concentration since 1980s.

LITERATURE REVIEW

Reference	Region	Methodology	Temporal Resolution	Performance (R^2)
Provençal et al. (2017)	Israel, Taiwan	Formula	Hourly	0.07-0.31
Huang et al. (2018)	China	ML (RF)	Daily	0.88
Song et al. (2018)	China	Formula	Hourly	0.35
Xiao et al. (2018)	China	ML (Ensemble)	Daily	0.85
He et al. (2019)	China	Formula	Daily	0.1-0.34
Carmona et al. (2020)	Mexico	NN	Monthly	0.53-0.81
Dey et al. (2020)	India	Formula	Daily	0.8
Ma et al. (2020)	China	Formula	Daily	0.09-0.66
Navinya et al. (2020)	India	Formula	Daily	0.36
Bali et al. (2021)	India	Formula	Hourly	NR
Gupta et al. (2021)	Thailand	ML (RF)	Hourly	0.9
Yin (2021)	China, Japan, Korea	Formula	Daily	0.69
Jin et al. (2022)	Global	Formula	Daily	0.1
Ma et al. (2022)	China	ML (LightGBM)	Daily	0.26 – 0.61

Formula: $PM_{2.5} = [DUST] + [SS] + [BC] + 1.4 \times [OC] + 1.375 \times [SO_4]$
ML: Machine Learning, RF: Random Forest, NN: Neural Network

OBJECTIVES

This study aims to address the following objectives:

- A machine learning model to estimate ground level $PM_{2.5}$ concentration using MERRA-2 satellite data
- Comparison with empirical formula $PM_{2.5}$ concentration estimates
- A long-term state wise trend analysis (1980-2021) for India
- Identification of major aerosol component sources in India

MODELLING FRAMEWORK

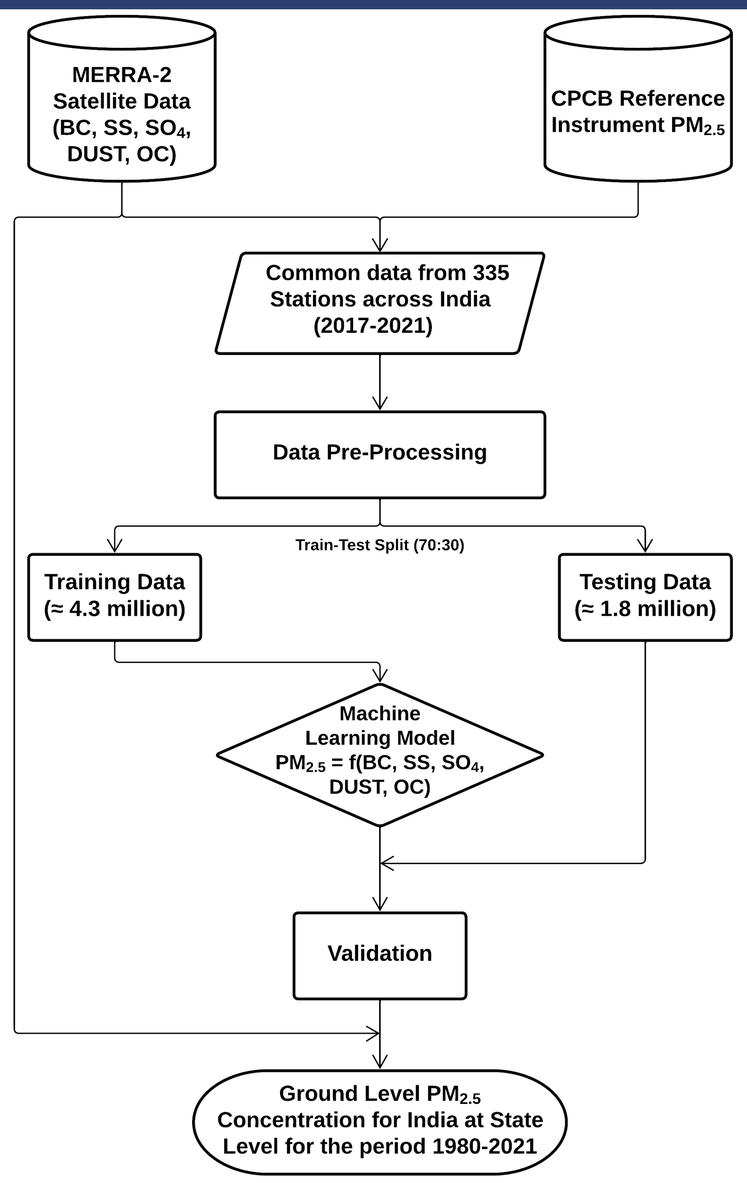

Figure 1. Modelling Framework for Estimation of Ground $PM_{2.5}$ from MERRA-2 Satellite Data

MODELLING RESULTS

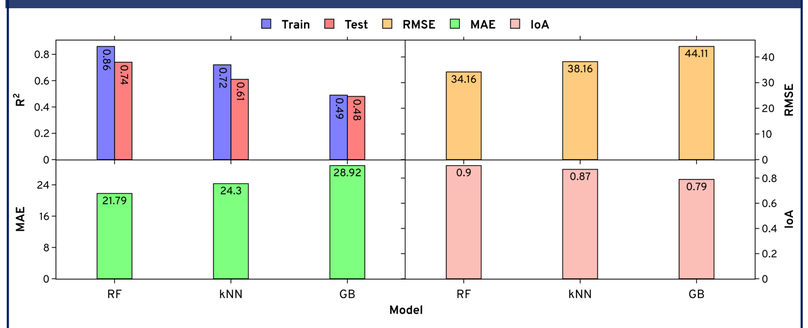

Figure 2. Performance of ML prediction models

Table 1. Comparison of Measured $PM_{2.5}$ with calculated and predicted

Metrics	$PM_{2.5\_Calculated}$	$PM_{2.5\_Predicted}$
n	1846469	1846469
R^2	0.27	0.74
RMSE	52.13	34.16
MAE	32.63	21.79
IoA	0.69	0.90

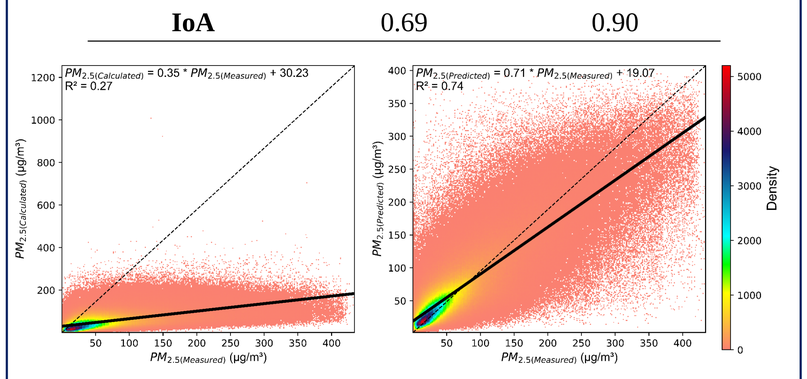

Figure 3. Comparison of $PM_{2.5}$ measured with predicted and calculated

Table 2. Summary of Measured, Calculated and Predicted $PM_{2.5}$

Metric	Calculated	Measured	Predicted
Count	1846469	1846469	1846469
Mean	52.66	63.31	64.30
SD	39.04	60.89	52.50
Min	0.91	0.07	1.79
25%	26.76	23.50	29.33
50%	43.51	44.00	48.68
75%	67.33	79.99	79.07
Max	1255.33	434.00	407.36

LONG-TERM VARIATION OF $PM_{2.5}$

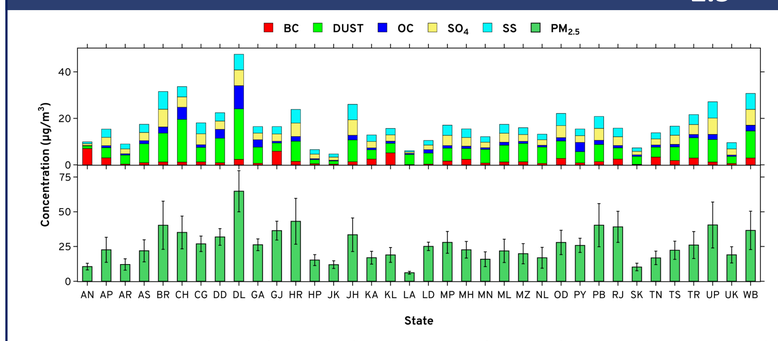

Figure 4. Average Elemental Concentration and $PM_{2.5}$ (1980-2021) across states

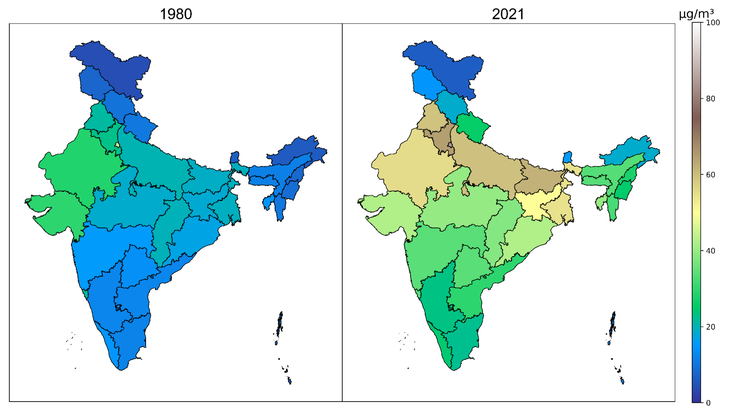

Figure 5. Average Annual $PM_{2.5}$ Concentration Comparison between 1980 and 2021 for Indian States

APPLICATIONS AND FUTURE WORK

- Back tracing to identify region-specific emission extensive activities through critical analysis of historical data.
- Dense spatial monitoring network to establish regulations and policy frameworks.
- Study of pollution episodes and seasonal variation of $PM_{2.5}$ ground level concentration.
- Exposure assessment across Indian states to assess air pollution exposures without extensive $PM_{2.5}$ monitoring locations, to adopt preventive public health measures.

REFERENCES

- Provençal S., Buchard, V., Silva, A. M. da, Leduc, R., Barrette, N., Elhacham, E., & Wang, S.-H. (2017). Evaluation of PM2.5 Surface Concentrations Simulated by Version 1 of NASA's MERRA Aerosol Reanalysis over Israel and Taiwan. Aerosol and Air Quality Research, 17(1), 253–261. https://doi.org/10.4209/aaqr.2016.04.0145
- Song, Z., Fu, D., Zhang, X., Wu, Y., Xia, X., He, J., Han, X., Zhang, R., & Che, H. (2018). Diurnal and seasonal variability of PM2.5 and AOD in North China plain: Comparison of MERRA-2 products and ground measurements. Atmospheric Environment, 191, 70–78. https://doi.org/10.1016/j.atmosenv.2018.08.012
- Ma, J., Zhang, R., Xu, J., & Yu, Z. (2022). MERRA-2 PM2.5 mass concentration reconstruction in China mainland based on LightGBM machine learning. Science of the Total Environment, 827, 154363. https://doi.org/10.1016/j.scitotenv.2022.154363.

PUBLICATIONS

- Kumar, V., Malyan, V., & Sahu, M. (2022). Significance of Meteorological Feature Selection and Seasonal Variation on Performance and Calibration of a Low-Cost Particle Sensor. Atmosphere, 13(4), 587. https://doi.org/10.3390/atmos13040587
- Kumar, V., Sahu, M., & Biswas, P. (2022). Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms. Aerosol and Air Quality Research, 22, 210240. https://doi.org/10.4209/aaqr.210240
- Kumar, V., & Sahu, M. (2021). Evaluation of nine machine learning regression algorithms for calibration of low-cost PM2.5 sensor. Journal of Aerosol Science, 157, 105809. https://doi.org/10.1016/j.jaerosci.2021.105809
- Sahu, M., Malyan, V., & Mayya, Y. S. (2021). Technologies for Controlling Particulate Matter Emissions from Industries. In S. P. Singh, K. Rathinam, T. Gupta, & A. K. Agarwal (Eds.), Pollution Control Technologies: Current Status and Future Prospects (pp. 253–290). Springer. https://doi.org/10.1007/978-981-16-0858-2_12