Ensemble Machine Learning Approach for PM2.5 Reconstruction using MERRA-2 and Long-term Analysis for India (1980-2021)

Vikas Kumar1, Vasudev Malyan2, Manoranjan Sahu2,1,3\*, Basudev Biswal4

1Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Mumbai 400076, India

2Aerosol and Nanoparticle Technology Laboratory, Environmental Science and Engineering Department, Indian Institute of Technology Bombay, Mumbai 400076, India

3Centre for Machine Intelligence and Data Science, Indian Institute of Technology Bombay, Mumbai 400076, India

4Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India

**Abstract**

Particle exposure affects more humans globally than any other air pollutant. However, due to expensive instruments and infrastructural deficiency, a high spatiotemporal network of monitoring stations is not possible, leading to data-scarce regions. Satellite and reanalysis datasets can be implemented to estimate particulate matter, but they do not provide surface concentration and needs to be reconstructed from the components. In this study, a machine learning (ML) framework is implemented to reconstruct PM2.5 from MERRA-2 data components, namely black carbon (BC), organic carbon (OC), dust (DUST), sea salt (SS), and sulfate (SO4). The ground level and respective MERRA-2 data were collected from India's 335 continuous ambient air quality monitoring stations (CAAQMS) for 2017-2021 at hourly resolution. Random forest (RF) performs better with train and test scores of 0.86 and 0.74, respectively, while the empirical equation provides an R2 of only 0.27 on test data. The estimated PM2.5 for Indian states from 1980-2021 indicates a significant increase in most cases. However, states in the Indo Gangetic plain such as Delhi, Punjab, Haryana, and Uttar Pradesh are the most polluted regions of India. The major shift in concentration is from 2000 onwards, which can be seen as a direct result of the economic liberalization policies implemented in 1991. The results provide evidence for the limitations of the broad application of the empirical equation and the feasibility of ML algorithms as a potential reconstruction technique for developing robust and accurate region-specific models from MERRA-2 data.

Keywords: PM2.5, MERRA-2, Machine Learning, India, Trend Analysis