K-MEANS CLUSTERING FOR ENHANCED SATELLITE AEROSOL RETRIEVAL

Author's: B. Hemamalini[1], R. Surendiran[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

Satellite-based aerosol retrieval plays a critical role in climate monitoring, air quality assessment, and environmental modeling. Aerosol Optical Depth (AOD), a key atmospheric parameter, is commonly derived from top-of-atmosphere (TOA) reflectance measurements using physics-based inversion algorithms. However, these traditional methods often struggle over heterogeneous surfaces such as bright urban regions, deserts, and mixed land–vegetation areas, where surface reflectance introduces significant uncertainty. To address these limitations, this study proposes a data-driven aerosol retrieval framework enhanced by K-Means clustering to improve prediction robustness and accuracy across varying atmospheric and surface regimes.The proposed system integrates synthetic satellite-like observations including multi-spectral reflectance bands (R470, R550, R670), vegetation proxy (NDVI), observation geometry (Solar Zenith Angle, View Zenith Angle, Relative Azimuth Angle), and meteorological parameters (Relative Humidity and Wind Speed). A supervised regression model is first developed as a baseline aerosol retrieval approach using all available training samples. Subsequently, K-Means clustering is applied to the standardized feature space to identify distinct atmospheric-surface regimes without prior labeling. These clusters represent implicit aerosol patterns such as urban-industrial, dust-dominant, or marine-clean conditions.For each identified cluster, a dedicated regression model is trained to learn localized nonlinear relationships between spectral features and AOD. During inference, new observations are assigned to the nearest cluster centroid, and the corresponding cluster-specific model is used for AOD prediction. This hybrid unsupervised–supervised architecture enables regime-aware modeling, allowing the system to adapt to heterogeneity in surface reflectance and aerosol type. Performance is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2).