ABSTRACT

Satellite-based optical remote sensing (RS) and data-driven models offer new possibilities for monitoring and forecasting river water quality across complex, data-scarce systems. This research developed a unified remote sensing (RS)–machine learning (ML) framework for the lower Ganges, specifically the Hooghly River in India, by integrating Sentinel-2 imagery, ERA5 reanalysis, and C3S seasonal forecast data. A Gaussian Mixture Model (GMM)-based cloud and cloud-shadow masking approach, combined with spatially aggregated reflectance sampling, improved TSS and turbidity retrieval accuracy by 33% and 28%, respectively, relative to retrievals obtained using the conventional Sentinel-2 Qualitative Assessment Band (QA60)/Scene Classification Layer (SCL) masks. Using these optical indicators with reanalysis temperature, four ML algorithms—PLSR, RFR, XGBoost, and SVR—were further employed to estimate non-optically observable parameters (NOOCs) such as dissolved oxygen (DO), electrical conductivity (EC), biochemical oxygen demand (BOD), ions, and pH. Predictor diagnostics using SHAP and VIP analyses identified the dominant spectral and meteorological controls on NOOC variability. For forecasting, a forward-chain training framework integrating ERA5 and C3S inputs enabled monthly-to-seasonal predictions (1–6 months ahead) across multiple parameters. Forecast skill, evaluated using Kling–Gupta Efficiency (KGE) and normalized mean absolute error (MAE%), declined by only 15–30% under ERA5 and 25–45% under C3S with increasing lead time, confirming statistical robustness via bootstrap testing. The forward-chain design consistently outperformed fixed-window training by yielding 10–25% lower MAE across leads. Overall, the proposed framework provides a scalable, interpretable, and operationally realistic approach for satellite-assisted water quality monitoring and forecasting, facilitating near-real-time decision support and adaptive management in dynamic riverine environments.

Keywords: Water quality, Remote sensing, Cloud shadows, Statistical resampling, Gaussian Mixture Model, Spatial aggregations, Machine learning, Non-optically observable parameters, Forward-chain learning, Seasonal prediction