Journal of Hydrology: Regional Studies 65 (2026) 103356

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh

Can Sentinel-2-derived spectral indices improve the accuracy of retrieving optically active water quality parameters using machine learning algorithms?

Elizabeth Modjadji Rathupetsane [a], Mahlatse Kganyago [a,*], Sabelo Madonsela [b], Vuyelwa Mvandaba [c]

[a] Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg 2092, South Africa
[b] Precision Agriculture Research Group, Advanced Agriculture and Food, Council for Scientific and Industrial Research (CSIR), Pretoria 0001, South Africa
[c] Smart Water Use Research Group, Water Centre, Council for Scientific and Industrial Research, Pretoria 0001, South Africa

ARTICLE INFO

Keywords:
Remote sensing
Chlorophyll-a
Total Suspended Solids
Sentinel-2
Random Forest
Gaussian Process Regression
Spectral indices
Water quality monitoring

ABSTRACT

Study region: This study was conducted in the Cradle of Humankind World Heritage Site (COHWHS), South Africa, an area characterised by interconnected surface waters and sensitive dolomitic aquifers. The region is subject to increasing pressure from land use change, tourism, and nutrient enrichment, making reliable and spatially explicit water quality monitoring essential for protecting its ecological, cultural, and hydrological integrity.

Study focus: The study aimed to assess whether Sentinel-2-derived spectral indices improve the retrieval accuracy of optically active water quality parameters, namely Chlorophyll-*a* (Chl-*a*) and Total Suspended Solids (TSS). Three input configurations were tested: traditional Landsat-like bands, Sentinel-2 bands, and Sentinel-2 bands combined with spectral indices. These inputs were used within Random Forest and Gaussian Process Regression models to evaluate model performance across wet (summer) and dry (winter) seasons.

New hydrological insights for the region: The results show that integrating Sentinel-2 spectral indices substantially improves Chl-*a* estimation during wet conditions, while TSS retrieval benefits mainly from Sentinel-2 red, red-edge, and SWIR bands. Model performance was strongly seasonal, with reduced accuracy during dry periods due to lower optical variability. The findings provide new insight into how seasonal hydrological conditions and spectral sensitivity influence water quality retrievals in optically complex inland waters of the COHWHS. This approach supports improved regional water quality monitoring and contributes to the protection of connected surface water-groundwater systems in this vulnerable heritage landscape.

1. Introduction

Monitoring water quality is crucial for understanding the variability of aquatic ecosystems, ensuring sustainable management, and providing early warnings of algal blooms and pollution events. This is particularly important in Africa, where many water bodies are severely polluted, leading to the loss of aquatic life, environmental degradation, and increased human health risks. For example,

* Corresponding author.
E-mail address: mahlatsek@uj.ac.za (M. Kganyago).

https://doi.org/10.1016/j.ejrh.2026.103356
Received 15 January 2026; Received in revised form 13 March 2026; Accepted 16 March 2026
Available online 19 March 2026

Africa's largest lake, Lake Victoria, has experienced severe degradation due to eutrophication, leading to disease outbreaks and fish mortality (Bangira et al., 2024). Similarly, Lake Chivero in Zimbabwe has suffered from pollution-induced animal deaths, while in South Africa, over 30 people died in 2023 due to contaminated municipal water (eNCA, 2025; News24 2024). These events highlight the ongoing deterioration of water quality, a problem that has persisted globally since the 1800s and continues to pose serious environmental and public health challenges (Harris et al., 2023). The degradation is primarily driven by anthropogenic activities such as the discharge of untreated sewage and industrial effluents, rapid urbanisation, and unsustainable agricultural practices (Bangira et al., 2024; J. N. Edokpayi et al., 2021). In Southern Africa, additional pressures arise from Acid Mine Drainage (AMD), poor infrastructure, and unregulated development near water sources. Kapalanga et al. (2021) reported that inadequate wastewater treatment in Namibia's Von Bach and Swakoppoort Dams contributes to deteriorating water quality. In South Africa, eutrophication remains a significant concern, intensified by limited maintenance capacity, weak regulation, and inadequate coordination between government departments (Department of Water and Sanitation DWS, 2016). Consequently, there is an urgent need for continuous, spatially comprehensive, and cost-effective monitoring solutions to support practical water management (Kowe et al., 2023; Lima et al., 2023).

Traditional in-situ and laboratory analysis methods of Optically Active Parameters (OAPs) such as Chlorophyll-*a* (Chl-*a*), Total Suspended Solids (TSS), turbidity, and Coloured Dissolved Organic Matter (CDOM) provide accurate data but are time-consuming, labour-intensive, and expensive, often limited in spatial and temporal coverage (Amieva et al., 2023; Maier et al., 2021; Theenathayalan et al., 2022). These limitations necessitate the adoption of remote sensing, enabling synoptic, repeatable, and large-scale water quality observations (Avdan et al., 2019; Chaabane et al., 2024; Taquan. Ma et al., 2023). Remote sensing has revolutionised water quality assessment over the past four decades by offering cost-effective and temporally consistent data for monitoring OAPs. Hyperspectral sensors, such as the PRecursore IperSpettrale della Missione Applicativa (PRISMA) and the Environmental Mapping and Analysis Program (EnMAP), offer high spectral resolution across over 200 narrow bands, enabling the detailed characterisation of bio-optical features (Lima et al., 2023; Maier et al., 2021). For instance, Saberioon et al. (2023) assessed EnMAP's ability to detect Chl-*a* and TSS, achieving higher accuracy than Saberioon et al. (2020), who used multispectral data. However, the challenge limiting its usage lies in selecting the optimal combination of spectral bands from a vast array of closely related bands, as well as the limited number of satellites in orbit, which makes the data costly for continuous monitoring (Yim et al., 2020). Consequently, multispectral sensors such as Landsat-8 Operational Land Imager (OLI) (Rodríguez-López, Usta, et al., 2023; Rubin et al., 2021), Moderate Resolution Imaging Spectroradiometer (MODIS) (Kravitz et al., 2021; Rahat et al., 2023), and Sentinel-2 Multispectral Imager (MSI) (Mpakairi et al., 2024; Saberioon et al., 2020) have become widely used in water quality monitoring due to their free accessibility and improved temporal resolution. Among these, Sentinel-2 has proven particularly effective, offering high spatial (up to 10 m) and temporal (5-day) resolutions ideal for inland water monitoring (Hafeez et al., 2022; Liu et al., 2017; Mpakairi et al., 2024).

Sentinel-2 offers a rich array of bands covering the visible, red-edge, Near-Infrared (NIR), and Shortwave Infrared (SWIR) regions. Still, their relevance varies with the optical properties of the water and the OAPs. For example, Gholizadeh et al. (2016) and Knaeps et al. (2015) observed that Chl-*a* has strong absorption in blue and red wavelengths, while TSS is better captured in red to SWIR regions due to its scattering properties. Moreover, Saberioon et al. (2020) identified the red-edge band as essential for Chl-*a* and TSS detection. To enhance predictive power, researchers have developed spectral indices, which are mathematical combinations of reflectance values across selected bands, to highlight specific optical features associated with water quality parameters. These indices include the Blue/Green ratio (Neil et al., 2019), Normalised Difference Chlorophyll Indices (NDCI) (Mishra and Mishra, 2012), Normalised Difference Suspended Solids Index (NDSSI) (Hossain et al., 2010), Two-Band Algorithm (2-BDA) (Gitelson et al., 2008), and Three-Band Algorithms (3-BDA) (Brivio et al., 2001; Dall'Olmo and Gitelson, 2006), which have been widely tested and applied for Chl-*a* and TSS estimation in inland and coastal water bodies. The effectiveness of these indices is largely attributed to the distinct optical behaviour of the two parameters. Chl-*a* strongly absorbs light in the blue (440 nm) and red (665 nm) regions while reflecting more in the green (560 nm) and red-edge (700–740 nm), whereas TSS increases backscattering across the visible and NIR regions, particularly in the red and NIR wavelengths due to suspended particle scattering. Ali et al. (2022) studied Chl-*a* using the Blue/Green and the Red/NIR ratios based on Sentinel-2, achieving R^2 of 0.86 and an RMSE of 2.56 μg/L, and they noted that these ratios contrast the strong Chl-*a* absorption in the blue and red regions against the high reflectance in the green and NIR. The Normalised Difference Turbidity Index (NDTI) (Lacaux et al., 2007) uses the normalised red and green reflectance, while the NDSSI uses the NIR and green reflectance to estimate TSS. The 3-BDA utilises reflectance combinations from the Visible and NIR (VNIR) bands to reduce background noise and isolate particle-driven scattering, thereby enhancing the accuracy of TSS and Chl-*a* estimation (Dias et al., 2021a; Singh et al., 2024).

While these indices have proven effective, site-specific optical complexity often limits their performance. Many spectral indices perform well in clear, phytoplankton-dominated waters but poorly in turbid, shallow, or CDOM-rich systems, where pigment absorption is masked by scattering (Mishra et al., 2014; Neil et al., 2019). Index performance also varies seasonally due to changes in phytoplankton biomass and the concentration of suspended solids. Moreover, several indices tend to saturate at high concentrations or are dominated by water absorption at longer wavelengths (Dogliotti et al., 2015; Rudorff et al., 2018). Consequently, there is growing recognition that locally optimised indices derived from site-specific spectral responses can outperform well-established indices (Najafzadeh and Basirian, 2023; Van Nguyen et al., 2019). Recent research, including our previous work (Rathupetsane and Kganyago, Under review), has shown that the best-performing spectral indices differ by water composition and optical conditions, suggesting that spectral index performance must be assessed regionally rather than applied universally (Neil et al., 2019; Pahlevan et al., 2020; Prior et al., 2020).

Machine learning algorithms have been applied to complex water bodies to overcome the limitations of individual indices and improve predictive performance. Dias et al. (2021b) evaluated Random Forest (RF), Support Vector Machine Radial Sigma (SVM-RS), Enhanced Adaptive Regression Through Hinges (EARTH), Multiple Linear Regression (MLR), and Cubist algorithms in retrieving TSS,

with SVM-RS achieving the highest accuracy, $R^2 = 0.87$. Maciel et al. (2021) evaluated the performance of Extreme Gradient Boosting (XGB), RF, and Support Vector Regression (SVR) in retrieving Secchi Disk Depth (SDD). The RF model showed better accuracy compared to other algorithms. Similarly, Kupssinskü et al. (2020) compared Linear regression (LR), SVR, K-Nearest Neighbours (KNN), RF, and Artificial Neural Networks (ANN) in retrieving TSS and Chl-*a*. The RF achieved the highest accuracy (R^2) of 0.90 and 0.86 for Chl-*a* and TSS, respectively. Nguyen et al. (2021) evaluated RF, SVR, Gaussian Process Regression (GPR), XGB, and CatBoost (CB) in estimating Chl-*a* using Sentinel-2 and found that GPR (R^2=0.85) and CB (R^2=0.84) yielded better results. On the other hand, Arias-Rodriguez et al. (2020) assessed RF, SVR, GPR, and LR in retrieving SDD and turbidity, and found that GPR exhibited superior performance for both parameters (i.e., $R^2 = 0.76$ for SDD and $R^2 = 0.83$ for turbidity), while the other models achieved R^2 around 0.70 for both parameters. Recent studies (Mpakairi et al., 2024; Ruescas et al., 2018) have compared the performance of spectral bands against a combination of spectral bands and indices as inputs to machine learning algorithms, and noticed improved accuracy in models integrating spectral indices.

This hybrid approach leverages the physical interpretability of indices and the nonlinear learning capability of algorithms. Shen et al. (2022) compared the performance of RF, SVR, XGB, and Deep Neural Network (DNN) models using Sentinel-3 Ocean and Land Colour Instrument (OLCI) bands, spectral indices, and combinations of bands with spectral indices in 24 lakes in China. Nine indices, including the red-edge/red ratio, 3-BDA, 4-BDA, Maximum Chlorophyll Index (MCI), and the Fluorescence Line Height (FLH) were included. Their results show that all models with bands and spectral indices had superiority over the other inputs, with improvements of up to 10%. Similarly, Leggesse et al. (2023) compared XGB, ANN, RF, GBR, ABR, and SVR in estimating Chl-*a*, turbidity, and Total Dissolved Solids (TDS) with different model performance per parameter. These studies emphasise the importance of utilising advanced regression models to capture nonlinear spectral relationships, particularly when combined with multiple input sources, such as spectral bands and derived indices.

Although machine learning algorithms have been increasingly applied in water quality monitoring, studies employing GPR remain limited globally, with no study conducted in South Africa for monitoring inland water quality. Moreover, existing comparisons between RF and GPR remain limited. To our knowledge, no previous work has compared these models for both Chl-*a* and TSS across wet and dry seasons using diverse spectral inputs. The Cradle of Humankind World Heritage Site (COHWHS) in South Africa has been facing severe water pollution from AMD and sewage effluent in recent years, posing a significant threat to its World Heritage status. Building upon this context, this study aimed to develop a robust machine learning model for retrieving OAPs using Sentinel-2 data over

Fig. 1. Locality map of the Cradle of Humankind World Heritage Site (COHWHS) in South Africa and the sampling sites.

inland surface water under various seasonal conditions. The objectives were: (1) to assess the impact of spectral indices on the performance of robust machine learning algorithms for retrieving inland OAPs under various seasonal conditions, and (2) to compare the performance of various machine learning algorithms in retrieving key optically active water quality parameters under various seasonal conditions and different input configuration, i.e., Traditional Landsat-like Bands (*TB*), comprising of visible and NIR bands without red-edge; Sentinel-2 Bands (*S2*), excluding B9 and B10; and Sentinel-2 combined with spectral indices (*S2 +Indices*). Given the ecological sensitivity of the COHWHS, its economic contribution and cultural significance, establishing an effective approach for monitoring the increasing pollution levels in the area is essential. Ultimately, the research advances remote sensing-based water quality monitoring by identifying the most effective input configurations and the impact of spectral indices and algorithms on Chl-*a* and TSS retrieval in optically complex South African waters.

2. Study area

South Africa hosts one of the largest and oldest fossil sites (i.e., The Cradle of Humankind World Heritage Site [COHWHS]), with more than 500 *Australopithecus* finds, including "Mrs Ples" and "Little Foot", and hosts about 40% of the world's known hominin fossils (Du Preez, 2019; Mugova and Wolkersdorfer, 2022). Due to its paleoanthropological significance, the COHWHS was declared a World Heritage Site in 1999 by the United Nations Educational, Scientific and Cultural Organisation (UNESCO) (Makhubela et al., 2019). This area comprises 13 major fossil sites, including the Bolts Farm, Swartkrans, Sterkfontein, Coopers, Kromdraai, Minnars, Gladysvale, Malapa, Haasgat, Gondolin, and the Maropeng Visitor Centre. The COHWHS spans approximately 47,000 ha across Gauteng and North West provinces, located approximately 50 km northwest of Johannesburg and 10 km north of Krugersdorp (Matyukira and Mhangara, 2023). Within this area, key river systems such as the Skeerpoort River, Bloubank River, Crocodile River, and Magaliesburg River exist. These rivers are under increasing anthropogenic pressure and contamination from toxic substances, primarily from Acid Mine Drainage (AMD) and wastewater (https://cohwhs.com/page7.html, accessed March 16, 2025) (Lukhele and Msagati, 2024). The COHWHS is characterised by karst landscapes, formed from chemically weathered dolomitic rock that supports diverse vegetation and complex hydrogeological systems connected to the polluted waters (Holland and Witthüser, 2009). Therefore, monitoring water quality in this area is crucial to preserve its ecological, cultural, and economic significance. Fig. 1 represents the extent of the study area in the context of South Africa.

3. Methods

3.1. Data

3.1.1. In-situ water sampling and laboratory analysis

Two field campaigns were held in wet (summer) and dry (winter) seasons for Chlorophyll-*a* (Chl-*a*) and Total Suspended Solids (TSS) sampling. The wet season fieldwork was conducted on the 14th and 16th of March 2024, characterised by high-flow waters. On the other hand, the dry season fieldwork occurred on August 28th and 31st 2024, characterised by low water flow. In both campaigns, twenty sites were visited, each selected based on accessibility and minimal obstruction from overhead features such as trees (see Fig. 1). At every sampling location, GPS coordinates were recorded using a handheld Garmin GPSMAP 65S device with an accuracy of 3 m (Lee et al., 2023). Before collection, water bottles were rinsed with water from the source to reduce the risk of contamination, and the collected samples were stored in a cooler box filled with ice cubes to maintain low temperatures (Gaur et al., 2022). This cooling step was essential to slow bacterial growth and preserve the samples in conditions as close as possible to those at the time of collection, ensuring more accurate analytical results (Gaur et al., 2022). The samples were transported to the Hydrology and Water Resources Laboratory at the Council for Scientific and Industrial Research (CSIR) in Pretoria, where they were stored in a cold room at 4°C (J. Edokpayi et al., 2016), before being transferred to the CSIR Chemistry Laboratory in Stellenbosch for analysis of optically active water quality parameters (OAPs) (Table 1).

3.1.2. Remotely sensed satellite data

This study employed the Sentinel-2 data provided by the European Space Agency (ESA). Sentinel-2 consists of three identical optical sensors, namely the Sentinel-2A which was launched in 2015, Sentinel-2B launched in March 2017 and Sentinel-2C, which was recently launched in September 2024 (Konapala et al., 2021). Each satellite has a Multi-Spectral Instrument (MSI) sensor with 13 spectral bands and a 290 km swath. Four spectral bands are provided at 10 m (B2, B3, B4, B8), six at 20 m (B5, B6, B7, B8A, B11, B12), and three bands at 60 m (B1, B9, B10) spatial resolution (Abazaj, 2020). The bands at 10 m present the Visible and NIR (VNIR) region with B2, B3, B4, and B8 representing the blue, green, red, and NIR bands, respectively. The bands at 20 m are designated for red-edge,

Table 1
Summary statistics of optically active parameters (AOPs), i.e., *Chl-a* and TSS, collected in the wet (i.e., summer) and dry (i.e., winter) seasons.

Season	OAP	Min.	Max.	Median	Range	Mean	Std. Dev.	CV (%)
Wet	Chl-*a* (μg/L)	5	653	24	648	92.8	174.29	187.81
	TSS (mg/L)	10	1543	53.5	1533	222.25	357.39	160.80
Dry	Chl-*a* (μg/L)	5	382	6	377	45	101.03	224.51
	TSS (mg/L)	8	549	39	510	124.65	158.05	126.8

narrow NIR and Shortwave Infrared (SWIR) bands, specifically red-edge 1, red-edge 2, red-edge 3, NIR-2, SWIR 1, and SWIR 2, corresponding to B5, B6, B7, B8A, B11, and B12, respectively. Lastly, the 60 m bands are for coastal, water vapour and cirrus clouds detection. The data is offered in three levels, i.e., "Top-Of-Atmosphere radiances" in sensor geometry, "Top-of-Atmosphere Reflectance" in cartographic geometry, and "Atmospherically corrected Surface Reflectance" in cartographic geometry as Level 1B, Level 1 C, and Level 2 A, respectively (Konapala et al., 2021). Sentinel-2 offers a high revisit time of 10 days at the equator with one satellite in operation and five days with two operational satellites. It has 2–3 days of revisiting time in other regions, such as COHWHS, where two swaths overlap (Yang and Jin, 2023). The high resolution makes it ideal for inland water quality monitoring, particularly in small and narrow water bodies. While Sentinel-2 Level-2A (L2A) imagery corrected using the Sentinel 2 Correction (Sen2Cor) algorithm is readily available, it was primarily designed for land applications and relies on the Dark Dense Vegetation (DDV) method to estimate aerosol optical thickness. This assumption makes it less suitable for water surfaces, particularly small or narrow inland water bodies, as it often fails to correct adjacency effects and may overcorrect aerosol contributions, leading to errors in water-leaving reflectance (Bui et al., 2022; Caballero et al., 2022). Sentinel-2 Level 1 C images of consistent dates with field campaigns mentioned in 3.1.1, representing the wet season (i.e., high-flow conditions) and dry season (i.e., low-flow conditions), were retrieved from the Copernicus Open Access Hub (https://dataspace.copernicus.eu/, accessed on 15 May 2024).

3.2. Data analysis

3.2.1. Data preprocessing

The Image Correction for Atmospheric Effects (*i*COR) algorithm (https://remotesensing.vito.be/services/icor, accessed 16 May 2025), developed by the Vlaamse Instelling Voor Technologisch Onderzoek (VITO), was used to correct atmospheric effects on Sentinel-2 Level 1 C data using the Sentinel Application Platform (SNAP) (De Keukelaere et al., 2018). *i*COR is a scene- and sensor-specific Atmospheric Correction (AC) algorithm compatible with Sentinel-2 MSI and Landsat-8 OLI data. It has been found to be effective when compared to other AC algorithms in previous studies (Ahmadi et al., 2025; Zolfaghari et al., 2023). Importantly, it includes the SIMilarity Environmental Correction (SIMEC) module to correct adjacency effects, a key advantage in heterogeneous environments like inland waters (Wolters et al., 2021). Its surface-adaptive design, minimal user interaction, and reliable performance in both land and water targets made *i*COR the most suitable option for this study. AC is necessary, especially in water quality monitoring, as atmospheric effects caused by aerosols, water vapour, and gases can contribute up to 90% of the signal received by optical sensors, affecting retrieval accuracy (Ansper and Alikas, 2019). The atmospherically corrected images were resampled to 10 m by 10 m spatial resolution using the nearest resampling method, which retains the original values of the imagery.

3.2.2. Synthetic data and experimental scenarios

Due to the lack of extensive data to develop robust machine learning models, we applied Multivariate Empirical Distribution Modelling (MEDM) to simulate the statistical properties and inter-variable dependencies observed in laboratory and satellite measurements (Smania and Jonsson, 2021). This approach treats inputs as correlated variables following a probability distribution derived from observed data. The method ensures that the generated data retains the statistical structure of in-situ and laboratory measurements, hence enabling reliable estimations in machine learning model training and validation (Teutonico et al., 2015). Approximately 1000 data points were simulated, resulting in some negative values for Chl-*a* and TSS that do not exist in reality. Therefore, these were removed from the data frame, and the remaining data points were used as inputs for machine learning regression analysis. Table 2 presents the summary statistics of data generated by MEDM in wet (i.e., summer season, high-flow conditions) and dry (i.e., winter season, low-flow conditions). The difference between the synthetic data and laboratory analysis is not major, with a notable difference of 167 mg/L for TSS and 55 μg/L for Chl-*a*, occurring under wet conditions, and no difference for both OAPs under dry conditions. The remaining data was divided into 80% training and 20% testing data for modelling OAP using machine learning algorithms.

Prior to modelling using machine learning algorithms, we divided the input data into three experimental scenarios, i.e., Traditional bands (*TB*) denoting the Landsat-like spectral bands covering the visible, NIR and SWIR regions without the red-edge bands, Sentinel-2 bands (*S2*) which included all Sentinel-2 bands except bands 9 and 10 and Sentinel-2 bands combined with selected spectral indices (*S2 +Indices*). The selection of indices was based on our previous work (Rathupetsane and Kganyago, *under review*). The indices chosen are presented in Table 3.

3.2.3. Machine learning algorithms

Machine Learning is a subset of Artificial Intelligence that focuses on constructing computer systems capable of learning from data (Nevo et al., 2022). Due to advancements in computation and algorithms, machine learning has become a valuable tool for

Table 2
Summary statistics of simulated data using the Multivariate Empirical Distribution Modelling (MEDM) approach under wet (summer) and dry (winter) seasons.

Season	OAP	*n*	Min	Max	Range	Median	Mean	Std. Dev	CV (%)
Wet	Chl-*a* (μg/L)	542	0.67	613.75	613.08	166.92	179.8	126.22	70.20
	TSS (mg/L)	542	1.56	1375.99	1374.42	371.95	404.85	257.56	63.62
Dry	Chl-*a* (μg/L)	521	0.3	327.96	327.66	81.189	92.35	64.45	69.79
	TSS (mg/L)	521	0.07	605.07	605	158.12	177.07	119.47	67.47

understanding non-linear relationships and generating automatic predictions (Zehra, 2021). To model Chl-*a* and TSS, two machine learning algorithms were implemented in the R Statistical software, i.e., Random Forest (RF) and Gaussian Process Regression (GPR). Each model was evaluated independently for each season and OAP, using the same experimental scenario (including training and testing data) for consistency.

RF is a supervised machine learning algorithm for classification and regression analysis (Breiman, 2001) that uses decision trees, where each tree is trained on a random subset of the input features. The model generates a diverse set of uncorrelated trees by introducing this randomness at each node. For regression tasks, the RF model combines the predictions of all individual trees by averaging their outputs (Arias-Rodriguez et al., 2020). This ensemble approach reduces model variance and minimises the risk of overfitting by leveraging the collective strength of multiple independent learners (Joshi et al., 2023; Ngwenya et al., 2025). It is one of the most widely used machine learning algorithms due to its consistent superiority against other machine learning algorithms (Karimi et al., 2023; Ngamile et al., 2025; Saberioon et al., 2023). Moreover, it offers the variable importance of the predictor variables, which helps understand the influence of each predictor on the prediction outcomes. The study utilises the permutation importance, which uses the mean decrease in accuracy (%IncMSE) implemented through the varImp() function in the caret R package. This metric quantifies importance by measuring the increase in prediction error when the values of each predictor are randomly permuted. A larger %IncMSE indicates a more influential variable. Permutation importance is considered more reliable than impurity-based metrics because it directly evaluates a variable's contribution to predictive performance on out-of-bag samples. The hyperparameters required in RF include the number of trees in the forest (i.e., *ntree*) and the number of samples needed at each node (i.e., *mtry*). The RF models were executed using the 'randomForest' R package. The variable importance was scaled from 0 to 100, allowing comparison of predictor influence on the modelled AOP.

On the other hand, GPR is a versatile, non-linear, kernel-based method grounded in a Bayesian framework. As a non-parametric probabilistic model, GPR captures complex relationships between inputs and outputs without assuming a predefined functional form. Unlike traditional regression techniques that aim to determine a single best-fit model, GPR estimates a distribution over possible functions and computes posterior predictive distributions for new test inputs (Das, 2025). This approach enhances predictive robustness and enables explicit quantification of uncertainty in model predictions. The performance of a GPR model is heavily influenced by the choice of a kernel function, which defines the degree of similarity between data points (Varvia et al., 2023). Typically, an empirical Bayes strategy is employed, wherein the hyperparameters are optimised by maximising the log marginal likelihood (García-Nieto et al., 2020). The hyperparameter for the Radial Basis kernel function (rbfdot), used here, includes the signal variance, which controls the vertical scaling of the kernel function and the noise variance, representing the level of Gaussian white noise assumed to be independently and identically distributed across observations. The GPR model was executed using the 'gausspr' R package.

The tuning parameters for both RF and GPR were optimised using the GridSearch strategy. The *ntree* values from 50 to 500 with an interval of 50 and the *mtry* values from 1 to 8 were evaluated using a 5-fold cross-validation. For GPR, the signal variance and noise variance were also tuned using 5-fold cross-validation, with no specified ranges, allowing the model to select the kernel hyper-parameter that minimises the prediction error.

3.2.4. Validation metrics

We used various statistical metrics, i.e., coefficient of determination (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Bias, which are commonly used to assess the accuracy and reliability of regression models in literature (Rodríguez-López et al., 2023). The R^2 (Eq. 1) measures the relationship between actual and predicted water quality OAPs. The RMSE (Eq. 2) and MAE (Eq. 3) measure the prediction error of a model. RMSE evaluates the square root of the average squared differences between predicted and actual values, while MAE calculates the average absolute differences, offering a more direct measure of prediction accuracy (García-Nieto et al., 2020). Bias (Eq. 4) measures the average difference between predicted and actual values, providing insight into

Table 3
Optimal spectral indices used as input with spectral bands for estimation of Chl-*a* and TSS under wet (i.e., summer) and dry (i.e., winter) conditions.

Season	OAP	Spectral index	R^2
Wet	Chl-a	$\frac{B8-B11}{B12}$	0.25
		$\frac{B8A-B11}{B12}$	0.25
	TSS	$\frac{B7-B8}{B12}$	0.74
		$\frac{B6-B8}{B12}$	0.63
Dry	Chl-a	$\frac{B7-B6}{B7+B6}$	0.12
		$\frac{B7}{B6}$	0.12
	TSS	$\frac{B4-B5}{B12}$	0.64
		$\frac{B6-B7}{B12}$	0.55

the model's ability to reduce under- and over-estimations (Tian et al., 2022). The model is said to perform well if it has a higher R^2 (i.e., closer to 1) and lower RMSE and MAE values. In contrast, the model performs well when the Bias is close to 0, showing less under- and over-estimation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(X_i - Y_i)^2}{n}} \quad (2)$$

$$MAE = \sum_{i=1}^{n}\frac{|X_i - Y_i|}{n} \quad (3)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\frac{[X_i - Y_i]}{X_i} \quad (4)$$

X_i is the actual value, Y_i is the predicted value, and n is the number of samples.

4. Results

4.1. Hyperparameter tuning of machine learning algorithms

The optimal hyperparameters for Random Forest (RF) and Gaussian Process Regression (GPR) under various experimental scenarios and seasonal conditions are presented in Table 4. These hyperparameters were used to train the RF and GPR models for Chl-*a* and TSS.

4.2. Performance of machine learning algorithms under various experimental scenarios

Different experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*), were tested to retrieve Chl-*a* and TSS under wet (i.e., summer) and dry (i.e., winter) seasonal conditions using Random Forest (RF) and Gaussian Process Regression (GPR). The prediction accuracy results for Chl-*a* are presented in Figs. 2 and 3 for RF and GPR, respectively. The RF model, comprising Traditional Landsat-like bands (i.e., *RF-TB*), performed poorly with very low R^2 values (0.004 in both wet and dry seasons) and high levels of errors (RMSE > 135 μg/L and > 72 μg/L in wet and dry seasons) (Fig. 2[a] and [b]), respectively. On the other hand, when using S2 spectral bands (i.e., *RF-S2*), the model improved substantially under wet conditions (Figure [c]), achieving R^2 of about 0.44 and RMSE = 98 μg/L, but provided only marginal gains in accuracy under dry conditions (Fig. 2 [d], i.e., R^2 = 0.01. The addition of S2-derived spectral indices (*S2 +Indices*) further improved RF performance, reaching R^2 = 0.47 under wet conditions (Fig. 2[e]) and reduced errors (RMSE = 96 μg/L), though the performance remained weak (R^2 = 0.025) under the dry season (Fig. 3[f]). The results for Chl-*a* retrieval using GPR models for the wet and dry seasons are illustrated in Fig. 3. Using the *TB* inputs, the performance was low under wet conditions (Fig. 3[a]), with R^2 = 0.044. However, when using *S2* (Fig. 3[c]), the performance increased markedly, to R^2 = 0.57 and RMSE < 87 μg/L. The best results were achieved with *S2 +Indices* model achieving an R^2 of 0.70 with the lowest RMSE < 72.24 μg/L and MAE < 59 μg/L) under wet conditions. However, all models failed to reliably predict Chl-*a* under dry conditions, where R^2 values were below 0.03 for RF and below 0.02 for GPR. Overall, the GPR models showed consistently higher accuracies than RF for both seasons, except for the *GPR-S2-Indices* in dry conditions.

The results for TSS in both wet and dry seasons, using RF and GPR, are shown in Figs. 4 and 5, respectively. RF comprising the Traditional Landsat-like bands (i.e., *RF-TB*) achieved the lowest accuracy during the wet (summer) season, achieving an R^2 of 0.33 and RMSE of 211 mg/L. During the dry (winter) season, the R^2 improved markedly by about 17% and the RMSE reduced by 136 mg/L. Expanding the input set to include red-edge bands (i.e., *S2*) and indices (i.e., *S2 +Indices*) yielded incremental gains in RF models, with

Table 4
Optimal hyperparameters of the Random Forest (RF) and Gaussian Process Regression (GPR) models under different experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*), and seasons, i.e., wet (dry).

		Chl-*a*			TSS		
		TB	*S2*	*S2 +Indices*	*TB*	*S2*	*S2 +Indices*
RF	*ntree*	500	500	500	500	500	500
	mtry	2 (1)	8 (1)	8 (1)	1 (6)	7(8)	8(8)
GPR	Noise variance	0.65 (0.79)	0.29 (0.76)	0.27 (0.73)	0.49 (0.31)	0.35 (0.33)	0.35 (0.36)
	Signal variance	0.47 (0.62)	0.19 (0.21)	0.10 (0.11)	0.50 (0.57)	0.24 (0.23)	0.12 (0.11)

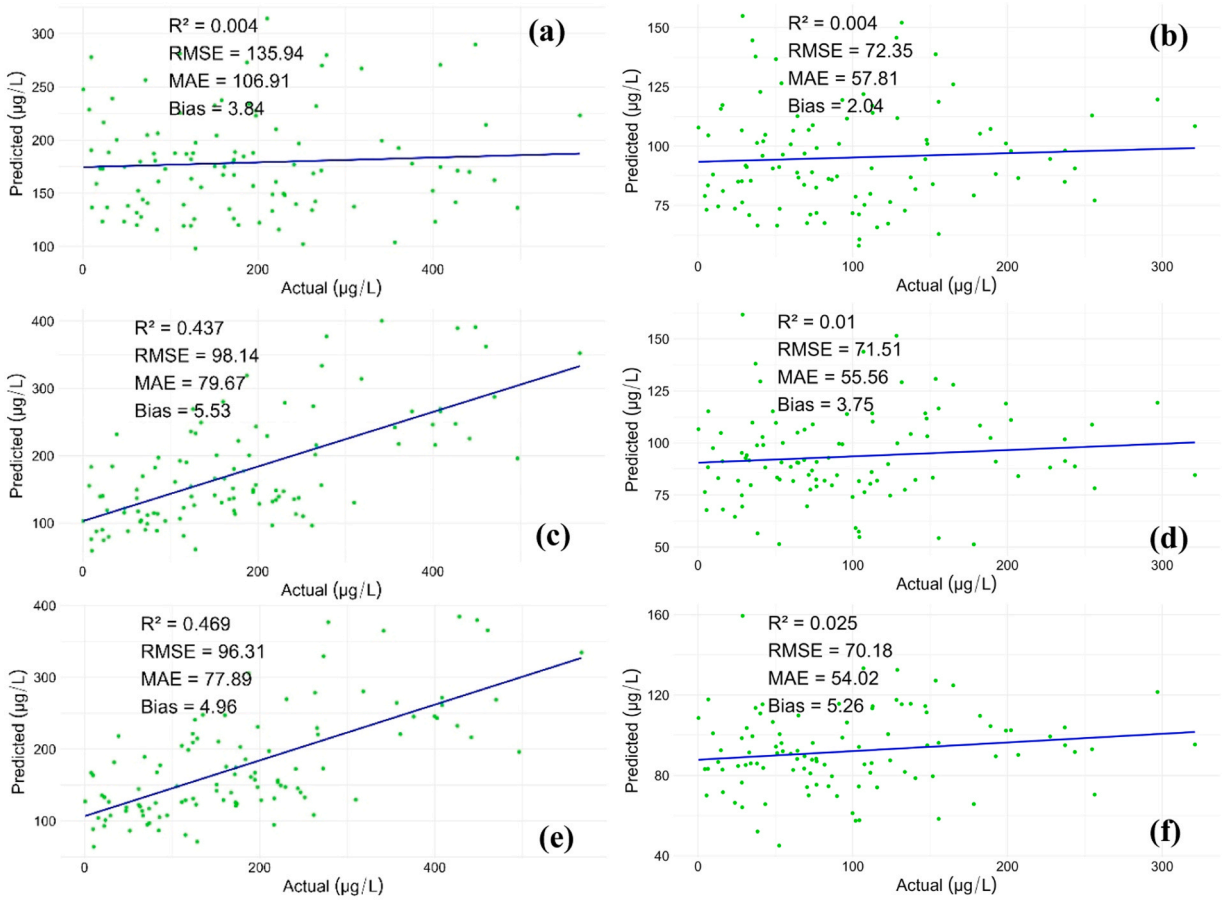

Fig. 2. Scatter plots for Chl-*a* predictions in the wet (i.e., summer) season [(a), (c), (e)] and in the dry (i.e., winter) season [(b), (d), (f)] using Random Forest (RF) under various experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*). Panels (a) and (b) are for *RF-TB*, while (c) and (d) are for *RF-S2*, and (e) and (f) are for *RF-S2 +Indices*.

the *RF-S2-Indices* model achieving the highest R^2 value of 0.44 during the wet conditions. However, the adding red-edge bands and indices did not necessarily improve the performance during the dry season, with *RF-TB* achieving an R^2 equivalent to *RF-S2*, while *S2 +Indices* resulted in a slightly worse R^2 of about 0.49. Using the *TB* inputs, it achieved an R^2 that is better than *RF-TB* by 10% and 7% during wet and dry conditions, respectively. Using *S2* further improved the results, achieving better R^2 than *RF-S2* by 16% and 9% during wet and dry seasons, respectively. The *GPR-S2-Indices* provided the most robust results during the wet season, achieving the lowest RMSE of approximately 159 mg/L, though slightly lower R^2 than *GPR-S2*. In the dry season, the performance of *GPR-S2-Indices* was slightly lower ($R^2 = 0.537$) than that of *GPR-TB* and *GPR-S2*. Overall, GPR consistently showed lower RMSE and MAE compared to RF, especially with the inclusion of additional spectral bands and indices across both seasons. Both RF and GPR models achieved better performance for TSS than for Chl-*a*, with generally higher R^2 values and lower errors than those achieved for Chl-*a*.

4.3. Variable importance

The results for RF variable importance for each model are presented in Fig. 6 for the wet (summer) and dry (winter) seasons. The performance of the *RF-S2* and *RF-S2-Indices* models for Chl-*a* was mainly influenced by B1, exhibiting maximum importance in both seasons. For the wet season (summer), B5, B8, and B12 were the second most influential features, followed by the B8A-B11-B12 spectral index and B11. B6 did not influence both models, while B8A did not contribute to the RF-SB model. The *RF-TB* model identified B8 as the most influential, reaching maximum importance, followed by B2 and B3. B4 made no contribution to the model's performance. During the dry (winter) season, the other influential features after B1 were B6, B4, B2, B5, and B7 in sequential order of importance for the *RF-S2* model. In contrast, the index: B6-B7-B12, and the spectral bands: B12, B3, B5, and B2 (in order of importance), were the second most influential features in the *RF-S2-Indices* model. B8A exhibited no contribution to the *RF-S2* model performance while showing the least contribution to the *RF-S2-Indices* model. In the *RF-TB* model, the most influential features were B3, followed by B8, B2, and B4, B12 had no contribution to model performance.

The performance of *RF-S2* and *RF-S2-Indices* TSS models during the wet season was mostly influenced by B5 and B3. The *RF-S2-*

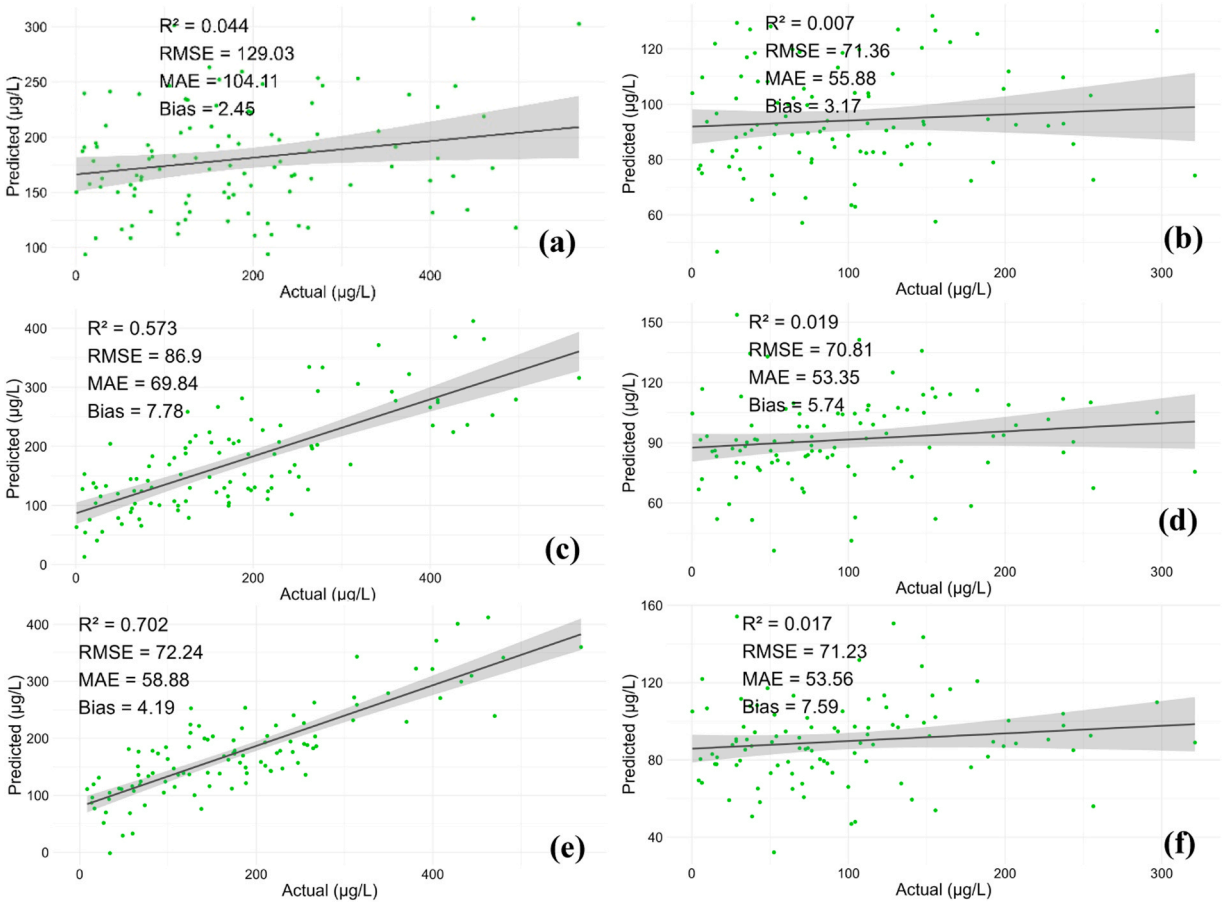

Fig. 3. Scatter plots for Chl-*a* predictions in the wet (i.e., summer) season [(a), (c), (e)] and in the dry (i.e., winter) season [(b), (d), (f)] using Gaussian Process Regression (GPR) under various experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*). Panels (a) and (b) are for *GPR-TB*, while (c) and (d) are for *GPR-S2*, and (e) and (f) are for *GPR-S2 +Indices*.

Indices identified the B7-B8-B12 as the third most influential and B4 as the fourth, while *RF-S2* had B4 as the third most influential feature. In the model with traditional Landsat-like bands, i.e., *RF-TB*, the performance was mostly influenced by B4, followed by B3 and B12, while B6 exhibited no contribution. During the dry (winter) season, B4 had the highest influence on performance of all models, B2 was the second most important feature, and B11 was third. Overall, during the dry season, B6, B7, and B8A exhibited marginal contributions to the performance of *RF-S2,* and *RF-S2-Indides,* while B3 and B12 had the least contributions for the *RF-TB model*.

4.4. Spatial distribution of optically active water quality parameters

The best models, i.e., *GPR-S2-Indices*, were utilised to predict the seasonal spatial distribution of Chl-*a* and TSS using Sentinel-2 imagery for Cradle Moon Lake, i.e., the largest dam in the study area (see Fig. 1). As shown in Fig. 7, the spatial distribution of Chl-*a* in the Cradle Moon Lake ranges between 0 and 271.9 µg/L and 45–151.3 µg/L during the wet (summer) and dry (winter) seasons, respectively, indicating higher wet season concentrations than dry season. Moreover, higher concentrations of Chl-*a* are found towards the edges of the lake, with low-median concentrations (blue-dark green) in the middle of the lake. In contrast, the TSS concentration ranged between 86.9 and 605.3 mg/L and 24.2–350 mg/L under wet and dry conditions, showing decreased concentrations in the dry season. The highest concentrations (orange-yellow) are found at the edges of the lake, while the lowest concentrations (blue-red) are found in the middle in wet conditions. However, in dry conditions, the distribution shows high TSS concentration on the edges except in the western part of the lake, while the middle is dominated by the median concentrations.

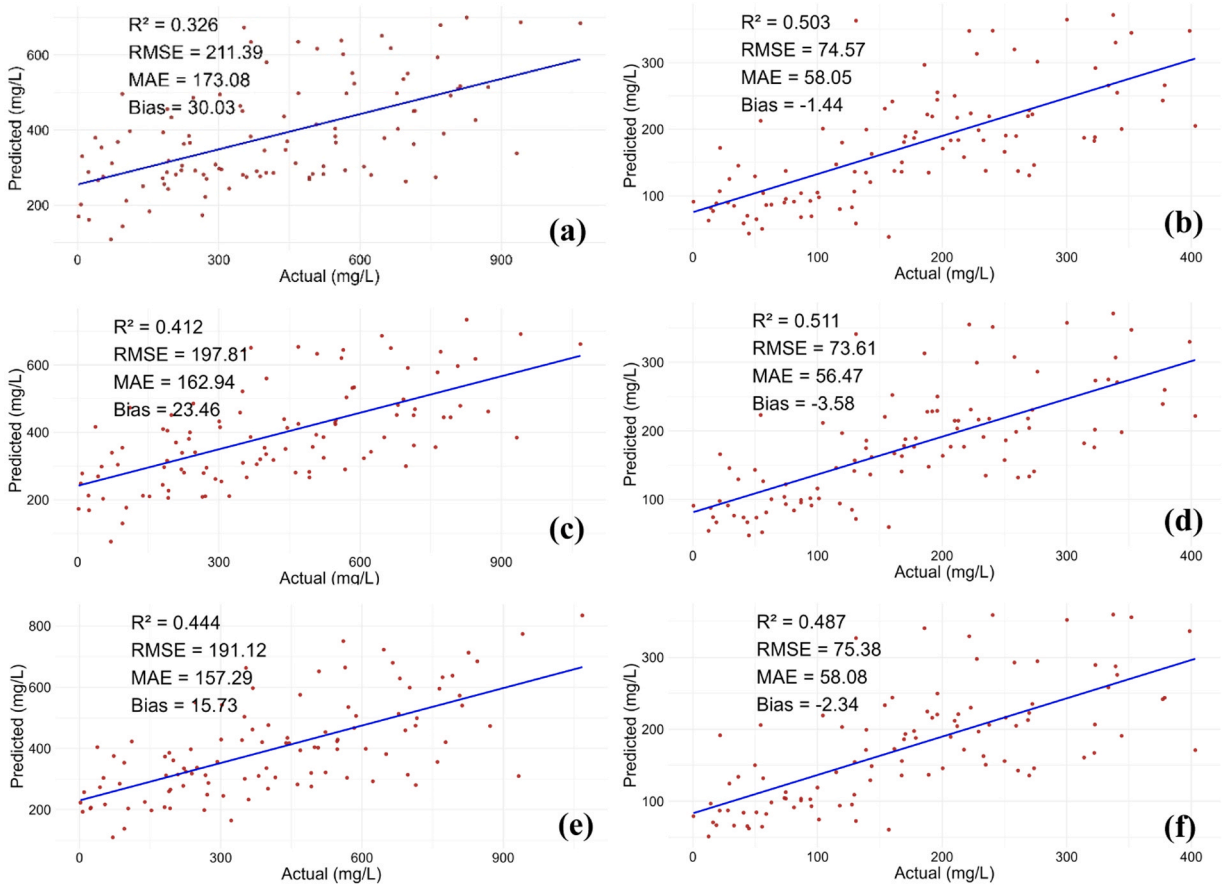

Fig. 4. Scatter plots for TSS predictions in the wet (i.e., summer) season [(a), (c), (e)] and in the dry (i.e., winter) season [(b), (d), (f)] using Random Forest (RF) under various experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*). Panels (a) and (b) are for *RF-TB*, while (c) and (d) are for *RF-S2*, and (e) and (f) are for *RF-S2 +Indices*.

5. Discussion

5.1. Influence of spectral indices on the performance machine learning algorithms

This study examined the impact of three input variables on model performance using the Random Forest (RF) and Gaussian Process Regression (GPR). Three input configurations were tested. Traditional Bands (*TB*), comprising visible and near-infrared bands without red-edge information; Sentinel-2 Bands (*S2*), which included all Sentinel-2 bands except Bands 9 and 10; and Sentinel-2 with derived spectral indices (*S2 +Indices*), which combined full-band reflectance with optimised indices derived from spectral analysis. Comparing *TB*, *S2*, and *S2 +Indices* inputs demonstrates that model performance systematically improved as more spectral information was included. For Chl-*a* estimation, the *TB* models captured only basic reflectance-OAP relationships, explaining only 4% of the variability and reporting the largest errors in both seasons due to limited spectral information (See Table 5). Expanding to full Sentinel-2 bands introduced key red-edge and SWIR features, improving model robustness, especially during the wet season, with a correlation improvement of over 44%. This aligns with findings from our systematic review currently under review (Rathupetsane et al., Under review) and previous studies (Mishra and Mishra, 2012; Mpakairi et al., 2024; Tian et al., 2024), which indicates that the red-edge bands make a significant contribution to the accurate estimation of Chl-*a*. For example, Pahlevan et al. (2022) compared Landsat-8 and Sentinel-2/3 in estimating Chl-*a* in selected global waters, and the accuracy was limited due to the lack of red-edge bands. Moreover, they noted that the bands were more effective on high Chl-*a* concentrations (> 5 mg/m^3). Similarly, higher pixel uncertainty (Saranathan et al., 2023) and underestimations (Ahmadi et al., 2025; Mpakairi et al., 2024) are observed in Landsat-8 data compared to Sentinel-2 data. This is also supported by the inclusion of the red-edge bands (B5, B6, B7) among the most influential bands on model performance (Fig. 6).

Adding derived indices in *S2 +Indices* enhanced the models by incorporating interactions among bands sensitive to scattering and absorption, resulting in the highest R^2 value of 0.70 for Chl-*a*. For TSS, the *S2 +Indices* models improved accuracy by 3% relative to the RF model while reporting the lowest errors on the GPR model during wet conditions. This supports previous research (Leggesse et al., 2023; Mpakairi et al., 2024; Ngamile, Kganyago, et al., 2025), which indicates that the inclusion of spectral indices enhances predictive

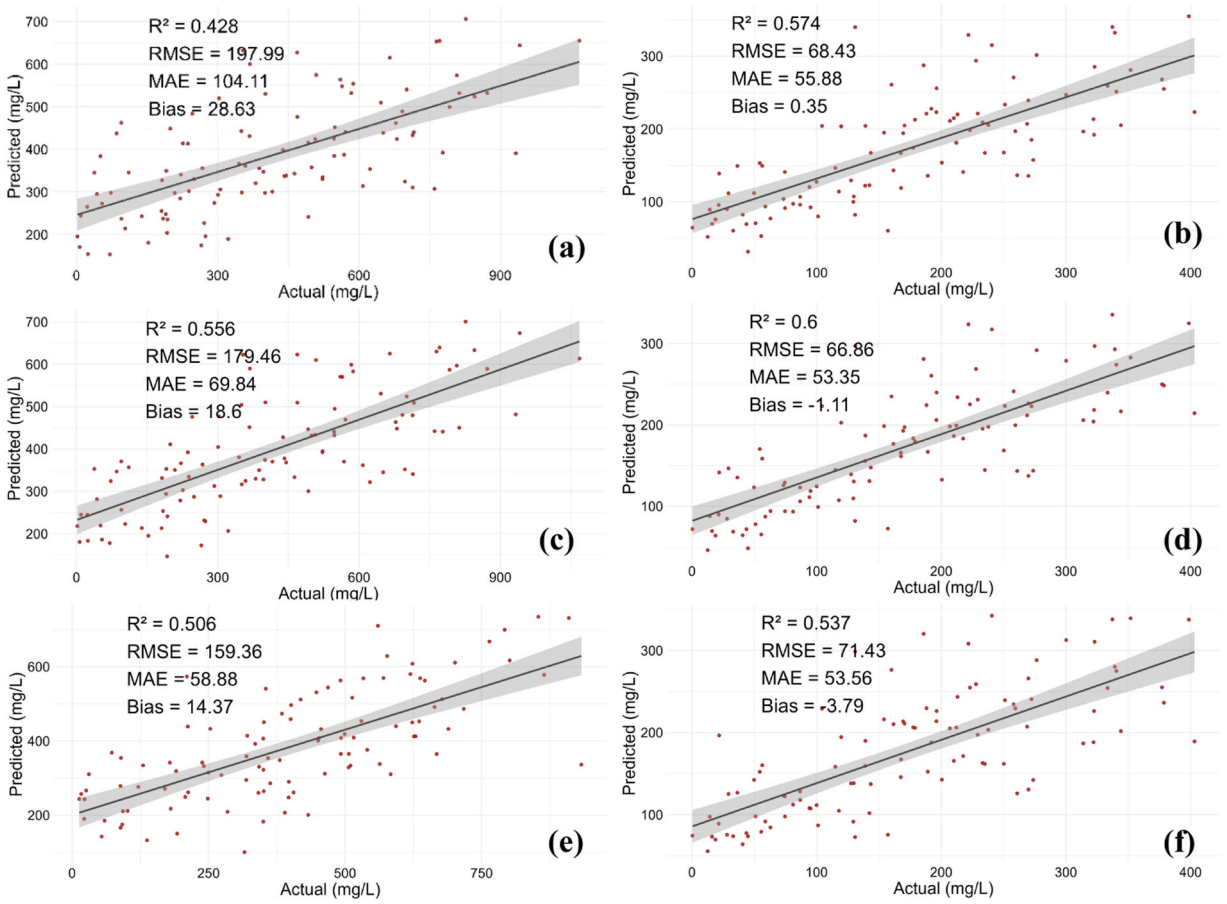

Fig. 5. Scatter plots for TSS predictions in the wet (i.e., summer) season [(a), (c), (e)] and in the dry (i.e., winter) season [(b), (d), (f)] using Gaussian Process Regression (GPR) under various experimental scenarios, i.e., Traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 and spectral indices (*S2 +Indices*). Panels (a) and (b) are for *GPR-TB*, while (c) and (d) are for *GPR-S2*, and (e) and (f) are for *GPR-S2 +Indices*.

performance. For instance, Ngamile, Kganyago, et al. (2025) compared the performance of RF using Sentinel-2 bands against Sentinel-2 bands combined with indices such as Normalised Difference Vegetation Index (NDVI), Normalised Difference Chlorophyll Index (NDCI), Maximum Chlorophyll Index (MCI), Red-Green Index (RGI), and Fluorescence Line Height (FLH) in the Cradle of Humankind World Heritage Site (COHWHS). Their results showed increased performance across OAPs, especially during the dry season. Ruescas et al. (2018) reported models using Sentinel-2 bands and Ratio Index (RI) superior to RI-only or Sentinel-2 bands-only for CDOM, achieving an $R^2 > 0.93$. Similarly, Hafeez et al. (2019) tested TSS, Chl-*a*, and turbidity in the Pearl River using RF, SVR, ANN, and CatBoost with band combinations and single bands in the VNIR region, achieving good accuracy ($R^2 > 0.71$). The effectiveness of spectral indices on model accuracy is more emphasised by their importance level on the variable importance analysis (Fig. 7). For example, $\frac{B8A-B11}{B12}$ had more contribution than B2, B3, B4, B6, B7, B8A, and B11 in the wet condition for estimating Chl-*a*. Similarly, $\frac{B7-B8}{B12}$ was the third most influential variable for TSS during the wet season. On the other hand, $\frac{B6-B7}{B12}$ was the second most influential variable during dry conditions after B1 for Chl-*a* estimations. Additionally, $\frac{B7}{B6}$, $\frac{B7-B6}{B7+B6}$, and $\frac{B4-B5}{B12}$ contributed more than B4, B6, B7, B8, B8A, and B11 on Chl-*a* model's performance. These results align with previous research, which found that the inclusion of spectral indices made a greater contribution to the model's accuracy than some individual bands (Mpakairi et al., 2024; Ngamile, Kganyago, et al., 2025; Zheng et al., 2024).

During the dry season, the inclusion of spectral indices did not improve the model's accuracy in estimating TSS, where it was surpassed by the *TB* and *S2* models by up to 6%. This has been reported by Ngamile, Kganyago, et al. (2025) in estimating Chl-*a* and TSS in COHWHS. Their results showed that *S2 +Indices* models improved performance only during the wet season, while underperforming in the dry season for Chl-*a* and reporting the same correlation for TSS with an error improvement of 1.2 mg/L. Similarly, Mpakairi et al. (2024) reported that Landsat-8 bands (similar to *TB*) and Sentinel-2 bands models performed better than *S2 +Indices* model ($R^2 = 0.80$) when estimating Chl-*a* in Nandoni Dam, reporting R^2 values of 0.87 and 0.89, respectively. Furthermore, Adjovu et al. (2024) reported the best model with only spectral bands ($R^2 = 0.88$), while models with spectral indices and spectral bands underperformed ($R^2 = 0.45$–0.86). This suggests that including spectral indices improves the model's performance, although not in all

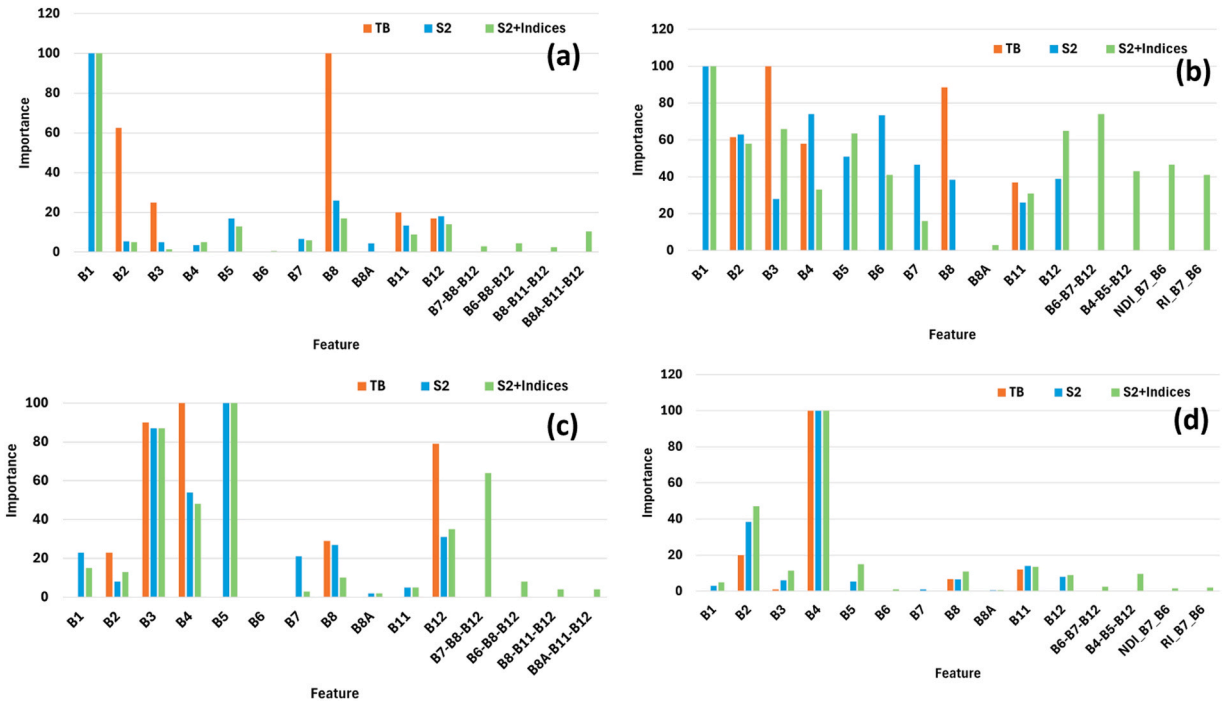

Fig. 6. The variable importance of Random Forest (RF) models for Chl-*a* in wet (summer [a]) and dry (winter [b]), and TSS in wet (summer [c]) and dry (winter [d]) seasons.

cases.

5.2. *Sensitivity of machine learning algorithms to seasonal OAPs distributions*

The RF model has been recorded to perform well in water quality monitoring, particularly due to its ensemble structure and resistance to multicollinearity compared to other machine learning algorithms (Khan et al., 2024; Kim et al., 2022; Ma et al., 2021; Saberioon et al., 2023). For example, Leggesse et al. (2023) and Kupssinskü et al. (2020) achieved R^2 values above 0.7 to predict Chl-*a* and TSS in various aquatic systems. In this study, the RF performed moderately in summer with R^2 values of 0.47 for Chl-*a* and 0.51 for TSS, in line with results from Joshi et al. (2024) and Arias-Rodriguez et al. (2023), who reported comparable R^2 values of 0.55 and 0.41, respectively. However, previous research (Arias-Rodriguez et al., 2020; Nguyen et al., 2021) has demonstrated that RF underperforms GPR. GPR's superiority is attributed to its kernel-based flexibility and probabilistic framework, which effectively capture complex spectra-Chl-*a* and spectra-TSS relationships, especially with small or noisy datasets. The results of this study further support the effectiveness of GPR, as models demonstrated superior accuracy across all configurations, seasons, and parameters compared to the RF models, except for Chl-*a* during the dry season, with a 1% difference. This comparison also allowed for assessing how band diversity and feature integration affect model sensitivity to the optical characteristics of Chl-*a* and TSS. The performance of RF and GPR models showed that model accuracy varied substantially by parameter, season, and input type, highlighting the interplay between spectral input diversity and algorithmic flexibility (See Table 5).

The seasonal analysis reveals that Chl-*a* was better predicted in wet conditions, whereas TSS was well predicted in dry conditions by both models. This is due to Chl-*a* peaks in late summer, where higher temperatures and light availability favour algal proliferation. This allows for a strong and distinct spectral signature, with the highest absorption around 440 nm and the highest reflectance around NIR, making it more detectable (Dalu et al., 2015; Gitelson et al., 1993). This is supported by the clear importance of B1, B5, B8, and B12) (Fig. 6) on Chl-*a* estimations during wet conditions and aligns with Joshi et al. (2024), who also found the B03 (442.5 nm), B04 (490 nm), B11 (708.75 nm), B16 (778.75 nm), and B21 (1020 nm) majorly influenced the predictiveness of the RF model using Sentinel-3 OLCI. These findings indicate that while Chl-*a* absorption dominates in the blue region, combinations integrating longer wavelengths helps to stabilise retrievals under optically complex conditions (Neil et al., 2019; Toming et al., 2016). In dry conditions, Chl-*a* is poorly predicted because lower concentrations exhibit a weaker spectral signal, which is weakened by scattering. Consequently, models depend on a wider range of variables, with no single wavelength dominating pigment retrieval. Moreover, the collection date (i.e., August) in this study was characterised by windy weather, which might have contributed to the suspension of sediments, masking Chl-*a* absorption (Leggesse et al., 2023). Consequently, this improved TSS retrieval during dry conditions. Although TSS concentrations were high during the wet season, introduced by rainfall, they have been found to promote algal bloom conditions assisted by higher temperatures, which might have overpowered their reflectance, leading to reduced retrieval accuracy. This spatial pattern corresponds with the spatial distribution (Fig. 7), and high concentrations of Chl-*a* and TSS are found on the edges

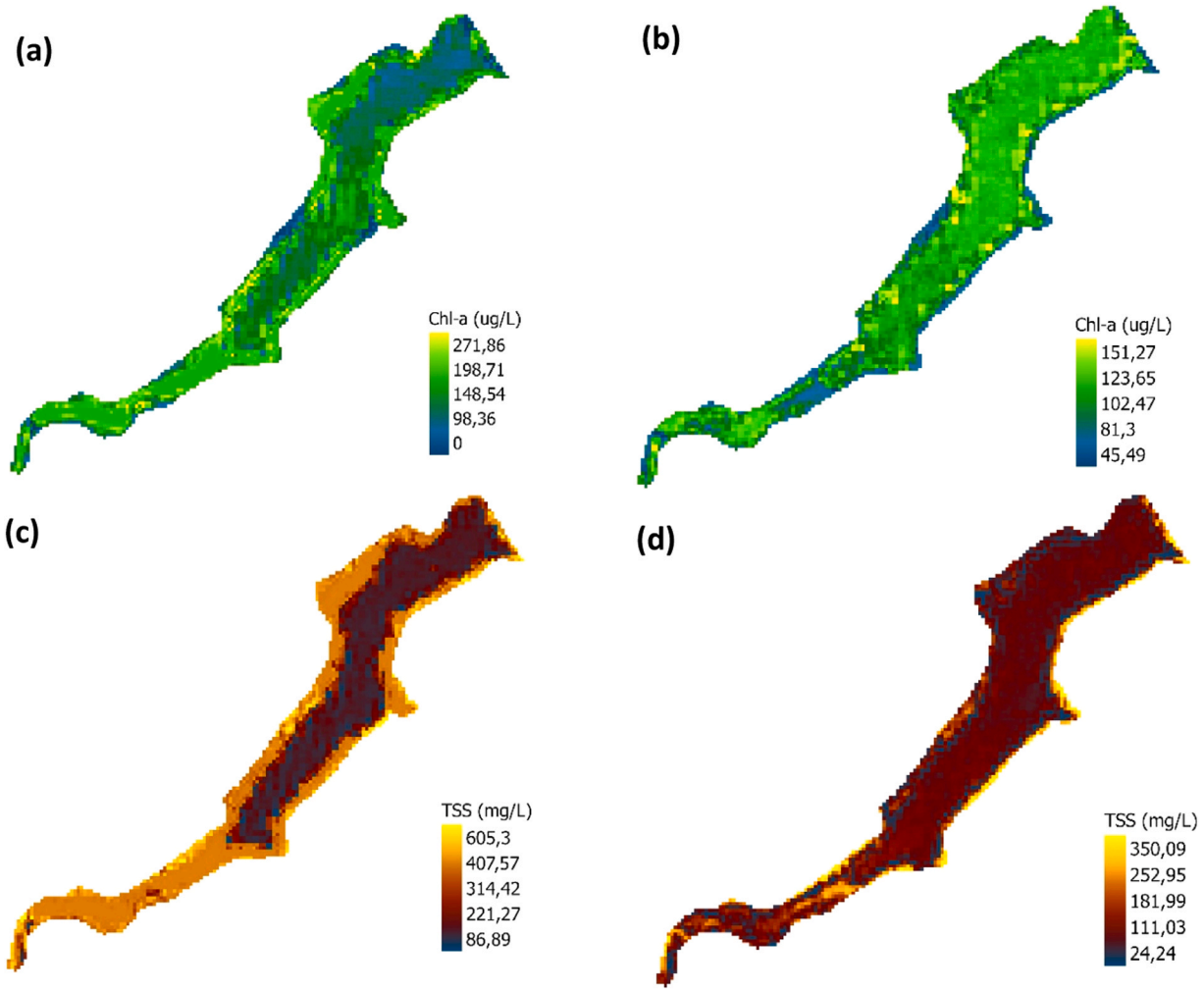

Fig. 7. Spatial distribution of Chl-*a* [(a) and (b)] and TSS ([c) and (d)] in wet season (summer; left) and dry season (winter; right) [c) and d)] using the Gaussian Process Regression *S2 + Indices* model (*GPR-S2-Indices*).

Table 5
Accuracy Assessment of Chl-*a* and TSS in wet (summer) and dry (winter) using traditional bands, Sentinel-2 bands, and spectral indices plus Sentinel-2 bands for Random Forest (RF) and Gaussian Process Regression (GPR).

OAP	Model	Wet (summer)				Dry (winter)			
		R^2	RMSE	MAE	Bias	R^2	RMSE	MAE	Bias
Chl-*a*	RF-TB	0.004	135.94	106.91	**3.84**	0.004	72.35	57.81	**2.04**
	RF-S2	0.44	98.14	79.67	5.53	0.01	71.51	55.56	3.75
	RF-S2-Indices	**0.47**	**96.31**	**77.89**	4.96	**0.03**	**70.18**	**54.02**	5.26
	GPR-TB	0.04	129.03	104.11	**2.45**	0.01	71.36	55.88	**3.17**
	GPR-S2	0.57	86.9	69.84	7.78	**0.02**	**70.81**	**53.35**	5.74
	GPR-S2-Indices	**0.70**	**72.24**	**58.88**	4.19	0.02	71.23	53.56	7.59
TSS	RF-TB	0.33	211.39	173.08	30.03	0.50	74.57	58.05	**-1.44**
	RF-S2	0.41	197.81	162.94	23.46	**0.51**	**73.61**	**56.47**	-3.58
	RF-S2-Indices	**0.44**	**191.12**	**157.29**	**15.73**	0.49	75.38	58.08	-2.34
	GPR-TB	0.43	197.99	104.11	28.63	0.57	68.43	55.88	**0.35**
	GPR-S2	**0.56**	179.46	69.84	18.6	**0.60**	**66.86**	**53.35**	-1.11
	GPR-S2-Indices	0.51	**159.36**	**58.88**	**14.37**	0.54	71.43	53.56	-3.79

of the lake, more specifically, on the southern part of the lake. This is where the inflow from the contaminated Bloubankspruit and Crocodile Rivers (Lukhele and Msagati, 2024) joins the lake and decreases towards the northern side due to vertical mixing. In contrast, during the dry season, both Chl-*a* and TSS concentrations were substantially reduced and more evenly distributed across the lake,

suggesting effects from internal processes, such as phytoplankton settling and reduced vertical mixing, dominate water quality variability when hydrological inputs are low (Sommer et al., 1986; Toming et al., 2024). The concentrations reported in the COHWHS indicate highly polluted waters, which align with the findings of Lukhele and Msagati (2024), who reported that waterbodies across the area and the country are highly polluted. Increased Chl-*a* concentrations and TSS concentrations hinder light penetration and deplete oxygen for the organisms present in the lake (Ngamile, Kganyago, et al., 2025). This emphasises the need for more effective monitoring of water quality in the area to maintain the ecological and cultural status of COHWHS, economic contribution, and access to clean water for humans and animals.

Overall, our study demonstrates the effectiveness of machine learning algorithms in estimating OAPs. Moreover, the inclusion of spectral indices as input with spectral bands improving the prediction accuracy of models. However, the accuracy reported in this study is lower than that of previous research, which used spectral indices and spectral bands as inputs for machine learning algorithms. Leggesse et al. (2023) and Zhang et al. (2024) reported $R^2 > 0.70$ for Chl-*a* in Hulun Lake and Lake Tana using the RF model. This difference can be attributed to the water conditions exhibited in the studies, which show lower concentrations (<192 μg/L) compared to those reported in the COHWHS (Table 1 and Fig. 7). Moreover, these studies utilised the RI, DI, SI, NDI, and 3-BDA (Gitelson et al., 2008), with spectral bands that could have enhanced the accuracy of the models. Our results showed higher performance on Chl-*a* ($R^2 = 0.47$) during the wet season than those of Ngamile, Kganyago, et al. (2025) in the COHWHS using RF (0.23). This can be attributed to differences in the spectral indices used: they used established indices, while we used optimal indices tailored for COHWHS. These indices have demonstrated higher performance in COHWHS than well-established indices, achieving a correlation of up to 74%, whereas established indices have reached a correlation of up to 68% (Rathupetsane and Kganyago n.d.). However, our methods underperformed in estimating TSS accuracy, achieving less than 50%, while Ngamile, Kganyago, et al. (2025) reported greater than 50%. This can be attributed to the synthetic data used here, whereas they used real data, and lower concentrations than we did. While the removing outliers has been recommended to improve model accuracy, this approach limits the model's application for operational purposes, as these outliers exist in reality. Therefore, their exclusion may lead to inaccurate reporting, resulting in the misleading formulation of policies and water management decisions.

5.3. Limitations of the study

This study demonstrated the utility of Sentinel-2 spectral information and spectral indices in improving the performance of machine learning algorithms in estimating OAPs in the COHWHS. However, several limitations are observed, resulting in lower accuracies than in previous research. The reduced accuracy can be attributed to the wide range of data, which spans both high and low concentrations. This is similar to Blix and Eltoft (2018), who reported reduced accuracy when predicting combined datasets from stations exhibiting both low and high concentrations than when estimating the stations separately. This highlights the limitation of single-output machine learning algorithms when dealing with diverse data for predicting OAPs in highly productive waters. Their failure to utilise the relationship between parameters where multiple OAPs influence the same spectral regions and exist in the same water body limits them. This might be enhanced by the use of multi-output machine learning algorithms that utilise the relationship between OAPs as recommended by Rathupetsane and Kganyago (Under review) and Rathupetsane et al. (Under review). These have been reported to exhibit excellent accuracy in forestry (Sahin et al., 2019; Varvia et al., 2023) and social affairs (Buebos-Esteve and Dagamac, 2024), but their transferability has not been explored. Moreover, the study utilised synthetic data generated by simulating satellite and laboratory data using Multivariate Empirical Distribution Modelling (MEDM). This nonparametric method offers several advantages in water quality monitoring. First, it preserves the complex interdependencies among water quality variables such as Chl-*a* and TSS, ensuring that the simulated data accurately reflects real-world conditions (Smania and Jonsson, 2021). Second, it is particularly useful in water quality analysis due to its flexibility, as it does not assume any predefined statistical distribution, making it robust against skewed data, outliers, or multimodal distributions common in water quality datasets. However, this method is limited in remote sensing applications. This study focused solely on broad seasonal trends, overlooking the specific conditions of autumn and spring, which might provide transitional conditions between the wet and dry seasons or between the dry and wet seasons. Future research should extend the temporal scope of observations to include multiple hydrological cycles and extreme events, thereby improving model calibration and transferability. Moreover, collaboration between researchers, environmental managers, and local authorities is essential to integrate these models into operational water quality monitoring frameworks and to align sampling dates with satellite passes, particularly in clear locations with no tree coverage for satellite observation. Addressing these limitations, the remote sensing-based monitoring of water quality can be enhanced, allowing cost-effective while achieving high accuracy for decision making and policy formulation across the world.

6. Conclusion

6.1. Main findings

This study demonstrated the potential of integrating Sentinel-2 imagery with robust machine learning algorithms to estimate Chlorophyll-*a* (Chl-*a*) and Total Suspended Solids (TSS) across diverse aquatic systems in the Cradle of Humankind World Heritage Site (COHWHS). By comparing traditional Landsat-like bands (*TB*), Sentinel-2 bands (*S2*), and Sentinel-2 bands combined with spectral indices (*S2 +Indices*), this study showed that incorporating S2 red-edge bands and carefully selected spectral indices significantly enhanced Random Forest (RF) and Gaussian Process Regression (GPR) performance. GPR consistently outperformed RF across Optically Active water quality Parameters (OAPs) considered (i.e., Chl-*a* and TSS), all experimental scenarios (i.e., designed to assess

the impact of various input configurations), and all seasons. For Chl-*a*, *GPR-S2-Indices* achieved the highest accuracy ($R^2 = 0.702$, RMSE = 72.24 μg/L), surpassing *GPR-TB* ($R^2 = 0.04$) and *GPR-S2* ($R^2 = 0.57$) during wet conditions. Similarly, the *RF-S2-Indices* (0.47) outperformed the *RF-S2* (0.44) and *RF-TB* (0.004). In contrast, dry season predictions were weak ($R^2 < 0.03$), reflecting reduced Chl-*a* concentrations and lower optical variability. For TSS, the *GPR-S2* and *RF-S2* models outperformed the *S2 +Indices* models, achieving up to $R^2 = 0.60$ and RMSE = 66.86 mg/L during the dry season. However, during wet conditions, the *S2 +Indices* models maintained their superior position, outperforming *S2* and *TB*. The variable importance analysis revealed that B1, B5, and B12 are important in estimating Chl-*a* and TSS. Their prominence reflects the complex and highly turbid nature of COHWHS waterbodies, where elevated concentrations of OAPs shift reflectance signals toward broader wavelengths from the VNIR to red-edge and SWIR regions, while the coastal band maintains the absorption peak of OAPs.

Overall, the methodology demonstrated an effective integration of remote sensing, machine learning algorithms, and spectral indices for monitoring water quality in the COHWHS. Accurate retrieval of Chl-*a* and TSS also indicates a strong potential to improve optically inactive parameters in the area, given their dependence on OAP information. At a policy level, the adoption of remote sensing-based approaches can support the Department of Water and Sanitation's efforts to expand spatial and temporal coverage of water monitoring, ultimately aiding in the protection of the regions' sensitive dolomitic aquifers and ensuring the long-term ecological sustainability of the COHWHS.

6.2. Limitations

The dataset included a wide range of low and high concentrations of water quality parameters, which may have reduced model accuracy when using single-output machine learning algorithms. These models do not account for the relationships between optically active parameters that often coexist and influence similar spectral regions within a water body. Additionally, the study relied on synthetic data generated using Multivariate Empirical Distribution Modelling (MEDM) to simulate satellite and laboratory observations. Although this approach preserves relationships between variables and is robust to skewed data and outliers, simulated datasets may not fully capture the variability present in real environmental conditions. The study also focused mainly on broad seasonal patterns, excluding transitional seasons such as autumn and spring, which may influence water quality dynamics.

6.3. Recommendations for future research

Future research should investigate the utility of multi-output machine learning algorithms. These algorithms can simultaneously predict multiple water quality parameters and exploit their interrelationships to improve prediction accuracy. Expanding the temporal coverage to include multiple hydrological cycles, extreme events, and transitional seasons would also improve model robustness and generalisation. Furthermore, future studies should prioritise the collection of extensive in-situ datasets and strengthen collaboration between researchers and environmental management authorities. Aligning field sampling with satellite overpass times would improve the reliability of satellite-derived water quality estimates and support the operational use of remote sensing technologies for long-term water quality monitoring.

Funding

This work is part of a project funded by the Water Research Commission (WRC), titled: Monitoring Surface Water Quality Using Remote Sensing Technology (Project number: C2023–2024–01241). The Article Processing Charge was funded by the University of Johannesburg.

CRediT authorship contribution statement

Vuyelwa Mvandaba: Writing – review & editing, Supervision, Funding acquisition, Data curation. **Elizabeth Modjadji Rathupetsane:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mahlatse Kganyago:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Funding acquisition, Conceptualization. **Sabelo Madonsela:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare ChatGPT was used to generate a graphical abstract. All the components of the generated graphic were carefully assessed for accuracy and confirmed by the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors appreciate the support provided by Dr Eunice Ubomba-Jaswa (Water Research Commission) and the University of Johannesburg (UJ) for providing the resources for this study. Ms S. Ngamile assisted with fieldwork. Finally, we acknowledge the contribution of the anonymous reviewers and the editorial team for guiding the crafting of this manuscript.

Data availability

Data will be made available on request.

References

Abazaj, F., 2020. SENTINEL-2 imagery for mapping and monitoring flooding in Buna River Area. J. Int. Environ. Appl. & Sci. 15 (Number 2). ⟨https://scihub.copernicus.eu/⟩.

Adjovu, G.E., Stephen, H., Ahmad, S., 2024. Application of machine learning algorithms for the estimation of the concentration of total suspended solids in the Colorado River Using Landsat 8 operational land imager data. World Environ. Water Resour. Congr. 2024 1424–1442. https://doi.org/10.1061/9780784485477.127.

Ahmadi, B., Gholamalifard, M., Ghasempouri, S.M., Kutser, T., 2025. Comparative analysis of k-nearest neighbors distance metrics for retrieving coastal water quality based on concurrent in situ and satellite observations. Mar. Pollut. Bull. 214, 117816. https://doi.org/10.1016/j.marpolbul.2025.117816.

Ali, K., Abiye, T., Adam, E., 2022. Integrating in situ and current generation satellite data for temporal and spatial analysis of harmful algal blooms in the Hartbeespoort Dam, Crocodile River Basin, South Africa. REMOTE Sens. 14 (17). https://doi.org/10.3390/rs14174277.

Amieva, J.F., Oxoli, D., Brovelli, M.A., 2023. Machine and deep learning regression of chlorophyll-a concentrations in lakes using PRISMA satellite hyperspectral imagery. Remote Sens. 15 (22). https://doi.org/10.3390/rs15225385.

Ansper, A., Alikas, K., 2019. Retrieval of chlorophyll a from Sentinel-2 MSI data for the European Union water framework directive reporting purposes. Remote Sens. 11 (1). https://doi.org/10.3390/rs11010064.

Arias-Rodriguez, L.F., Duan, Z., Sepúlveda, R., Martinez-Martinez, S.I., Disse, M., 2020. Monitoring water quality of Valle de Bravo Reservoir, Mexico, using entire lifespan of MERIS data and machine learning approaches. Remote Sens. 12 (10), 1586. https://doi.org/10.3390/rs12101586.

Arias-Rodriguez, L.F., Tüzün, U.F., Duan, Z., Huang, J., Tuo, Y., Disse, M., 2023. Global water quality of inland waters with harmonized Landsat-8 and Sentinel-2 using cloud-computed machine learning. Remote Sens. 15 (5). https://doi.org/10.3390/rs15051390.

Avdan, Z.Y., Kaplan, G., Goncu, S., Avdan, U., 2019. Monitoring the water quality of small water bodies using high-resolution remote sensing data. ISPRS Int. J. GeoInf. 8 (12). https://doi.org/10.3390/ijgi8120553.

Bangira, T., Matongera, T.N., Mabhaudhi, T., Mutanga, O., 2024. Remote sensing-based water quality monitoring in African reservoirs, potential and limitations of sensors and algorithms: a systematic review. Phys. Chem. Earth Parts A/B/C. 134, 103536. https://doi.org/10.1016/j.pce.2023.103536.

Blix, K., Eltoft, T., 2018. Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval. Remote Sens. 10 (5), 775. https://doi.org/10.3390/rs10050775.

Breiman, L., 2001. Random For. 45.

Brivio, P.A., Giardino, C., Zilioli, E., 2001. Determination of chlorophyll concentration changes in Lake Garda using an image-based radiative transfer code for Landsat TM images. Int. J. Remote Sens. 22 (2–3), 487–502. https://doi.org/10.1080/014311601450059.

Buebos-Esteve, D.E., Dagamac, N.H.A., 2024. Spatiotemporal models of dengue epidemiology in the Philippines: Integrating remote sensing and interpretable machine learning. Acta Trop. 255, 107225. https://doi.org/10.1016/j.actatropica.2024.107225.

Bui, Q.-T., Jamet, C., Vantrepotte, V., Mériaux, X., Cauvin, A., Mograne, M.A., 2022. Evaluation of sentinel-2/MSI atmospheric correction algorithms over two contrasted french coastal waters. Remote Sens. 14 (5), 1099. https://doi.org/10.3390/rs14051099.

Caballero, I., Román, A., Tovar-Sánchez, A., Navarro, G., 2022. Water quality monitoring with Sentinel-2 and Landsat-8 satellites during the 2021 volcanic eruption in La Palma (Canary Islands). Sci. Total Environ. 822. https://doi.org/10.1016/j.scitotenv.2022.153433.

Chaabane, S., Riahi, K., Khlifi, S., Slama, E., Vanclooster, M., 2024. Assessing the performance of a citizen science based water quality monitoring program for nitrates using test strips implemented in the medjerda hydrosystem in Northern Tunisia. Hydrology 11 (1). https://doi.org/10.3390/hydrology11010006.

Dall'Olmo, G., Gitelson, A.A., 2006. Effect of bio-optical parameter variability and uncertainties in reflectance measurements on the remote estimation of chlorophyll-a concentration in turbid productive waters: modeling results. Appl. Opt. 45 (15), 3577. https://doi.org/10.1364/AO.45.003577.

Dalu, T., Dube, T., Froneman, P.W., Sachikonye, M.T.B., Clegg, B.W., Nhiwatiwa, T., 2015. An assessment of chlorophyll- *a* concentration spatio-temporal variation using Landsat satellite data, in a small tropical reservoir. Geocarto Int. 30 (10), 1130–1143. https://doi.org/10.1080/10106049.2015.1027292.

Das, A., 2025. Surface water quality evaluation impacting drinking water sources and sanitation using water quality index, multivariate techniques, and interpretable machine learning models in Mahanadi River, Odisha (India). Environ. Geochem. Health 47 (11), 497. https://doi.org/10.1007/s10653-025-02806-0.

De Keukelaere, L., Sterckx, S., Adriaensen, S., Knaeps, E., Reusen, I., Giardino, C., Bresciani, M., Hunter, P., Neil, C., Van der Zande, D., Vaiciute, D., 2018. Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. Eur. J. Remote Sens. 51 (1), 525–542. https://doi.org/10.1080/22797254.2018.1457937.

Department of Water and Sanitation (DWS), 2016. Water Qual. Manag. Policies Strateg. South Afr. Rep. No. 1. 3 Water Qual. Water Qual. Manag. Chall. South Afr. ⟨https://www.dws.gov.za/iwrp/iwqms/Documents/Report%201.3_WQ%20and%20WQM%20Challenges%20in%20SA.pdf⟩.

Dias, R.L.S., da Silva, D.D., Fernandes-Filho, E.I., do Amaral, C.H., dos Santos, E.P., Marques, J.F., Veloso, G.V., 2021b. Machine learning models applied to TSS estimation in a reservoir using multispectral sensor onboard to RPA. Ecol. Inform. 65. https://doi.org/10.1016/j.ecoinf.2021.101414.

Dias, R.L.S., da Silva, D.D., Fernandes-Filho, E.I., do Amaral, C.H., dos Santos, E.P., Marques, J.F., Veloso, G.V., 2021a. Machine learning models applied to TSS estimation in a reservoir using multispectral sensor onboard to RPA. Ecol. Inform. 65. https://doi.org/10.1016/j.ecoinf.2021.101414.

Dogliotti, A.I., Ruddick, K.G., Nechad, B., Doxaran, D., Knaeps, E., 2015. A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters. Remote Sens. Environ. 156, 157–168. https://doi.org/10.1016/j.rse.2014.09.020.

Du Preez, E.A., 2019. The contribution of geological features to visitor experiences: comparison between two geotourism attractions in South Africa. Geoj. Tour. Geosites 26 (3), 1006–1020. https://doi.org/10.30892/gtg.2.

Edokpayi, J., Odiyo, J., Popoola, O., Msagati, T., 2016. Assessment of trace metals contamination of surface water and sediment: a case study of Mvudi River, South Africa. Sustainability 8 (2), 135. https://doi.org/10.3390/su8020135.

Edokpayi, J.N., Odiyo, J.O., Popoola, O.E., Msagati, T.A.M., 2021. Evaluation of contaminants removal by waste stabilization ponds: a case study of Siloam WSPs in Vhembe District, South Africa. Heliyon 7 (2), e06207. https://doi.org/10.1016/j.heliyon.2021.e06207.

eNCA, 2025, July 23. Resid. Hammanskraal Demand Answ. [Broadcast].

García-Nieto, P.J., García-Gonzalo, E., Alonso Fernández, J.R., Díaz Muñiz, C., 2020. A new predictive model for evaluating chlorophyll-a concentration in tanes reservoir by using a Gaussian process regression. Water Resour. Manag. 34 (15), 4921–4941. https://doi.org/10.1007/s11269-020-02699-x.

Gaur, N., Sarkar, A., Dutta, D., Gogoi, B.J., Dubey, R., Dwivedi, S.K., 2022. Evaluation of water quality index and geochemical characteristics of surfacewater from Tawang India. Sci. Rep. 12 (1), 11698. https://doi.org/10.1038/s41598-022-14760-3.

Gholizadeh, M., Melesse, A., Reddi, L., 2016. A comprehensive review on water quality parameters estimation using remote sensing techniques. Sensors 16 (8), 1298. https://doi.org/10.3390/s16081298.

Gitelson, A., Garbuzov, G., Szilagyi, F., Mittenzwey, K.-H., Karnieli, A., Kaiser, A., 1993. Quantitative remote sensing methods for real-time monitoring of inland waters quality. Int. J. Remote Sens. 14 (7), 1269–1295. https://doi.org/10.1080/01431169308953956.
Gitelson, A.A., Dall'Olmo, G., Moses, W., Rundquist, D.C., Barrow, T., Fisher, T.R., Gurlin, D., Holz, J., 2008. A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. Remote Sens. Environ. 112 (9), 3582–3593. https://doi.org/10.1016/j.rse.2008.04.015.
Hafeez, S., Wong, M.S., Ho, H.C., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K.H., Pun, L., 2019. Comparison of machine learning algorithms for retrieval of water quality indicators in case-ii waters: a case study of Hong Kong. Remote Sens. 11 (6), 617. https://doi.org/10.3390/rs11060617.
Hafeez, S., Wong, M.S., Abbas, S., Asim, M., 2022. Evaluating landsat-8 and sentinel-2 data consistency for high spatiotemporal inland and coastal water quality monitoring. Remote Sens. 14 (13). https://doi.org/10.3390/rs14133155.
Harris, A.R., Daly, S.W., Pickering, A.J., Mrisho, M., Harris, M., Davis, J., 2023. Safe today, unsafe tomorrow: Tanzanian households experience variability in drinking water quality. Environ. Sci. Technol. 57 (45), 17481–17489. https://doi.org/10.1021/acs.est.3c05275.
Holland, M., Witthüser, K.T., 2009. Geochemical characterization of karst groundwater in the cradle of humankind world heritage site, South Africa. Environ. Geol. 57 (3), 513–524. https://doi.org/10.1007/s00254-008-1320-2.
Hossain, A., Jia, Y., Chao, X., 2010. Dev. Remote Sens. Based Index Estim. /Mapp. Suspended Sediment Conc. River Lake Environ. ⟨https://www.researchgate.net/publication/251236287⟩.
Joshi, A., Pradhan, B., Gite, S., Chakraborty, S., 2023. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: a systematic review (MDPI). Remote Sens. 15 (Number 8). https://doi.org/10.3390/rs15082014.
Joshi, N., Park, J., Zhao, K., Londo, A., Khanal, S., 2024. Monitoring harmful algal blooms and water quality using sentinel-3 OLCI satellite imagery with machine learning. Remote Sens. 16 (13), 2444. https://doi.org/10.3390/rs16132444.
Kapalanga, T.S., Hoko, Z., Gumindoga, W., Chikwiramakomo, L., 2021. Remote-sensing-based algorithms for water quality monitoring in olushandja dam, north-central namibia. Water Supply 21 (5), 1878–1894. https://doi.org/10.2166/ws.2020.290.
Karimi, B., Hashemi, S.H., Aghighi, H., 2023. Development of the best retrieval models of non-optically active parameters for an artificial shallow lake by random forest algorithm. Remote Sens. Appl. Soc. Environ. 29. https://doi.org/10.1016/j.rsase.2023.100926.
Khan, R.M., Salehi, B., Niroumand-Jadidi, M., Mahdianpari, M., 2024. Mapping water clarity in small oligotrophic lakes using sentinel-2 imagery and machine learning methods: a case study of Canandaigua Lake in Finger Lakes, New York. IEEE J. Sel. Top. Appl. EARTH OBSERVATIONS REMOTE Sens. 17, 4674–4688. https://doi.org/10.1109/JSTARS.2024.3359648.
Kim, J., Jang, W., Hwi Kim, J., Lee, J., Hwa Cho, K., Lee, Y.-G., Chon, K., Park, S., Pyo, J., Park, Y., Park, Y., Kim, S., 2022. Application of airborne hyperspectral imagery to retrieve spatiotemporal CDOM distribution using machine learning in a reservoir. Int. J. Appl. Earth Obs. Geoinf. 114. https://doi.org/10.1016/j.jag.2022.103053.
Knaeps, E., Ruddick, K.G., Doxaran, D., Dogliotti, A.I., Nechad, B., Raymaekers, D., Sterckx, S., 2015. A SWIR based algorithm to retrieve total suspended matter in extremely turbid waters. Remote Sens. Environ. 168, 66–79. https://doi.org/10.1016/j.rse.2015.06.022.
Konapala, G., Kumar, S.V., Khalique Ahmad, S., 2021. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. ISPRS J. Photogramm. Remote Sens. 180, 163–173. https://doi.org/10.1016/j.isprsjprs.2021.08.016.
Kowe, P., Ncube, E., Magidi, J., Ndambuki, J.M., Rwasoka, D.T., Gumindoga, W., Maviza, A., de jesus Paulo Mavaringana, M., Kakanda, E.T., 2023. Spatial-temporal variability analysis of water quality using remote sensing data: A case study of Lake Manyame. Sci. Afr. 21. https://doi.org/10.1016/j.sciaf.2023.e01877.
Kravitz, J., Matthews, M., Lain, L., Fawcett, S., Bernard, S., 2021. Potential for High Fidelity Global Mapping of Common Inland Water Quality Products at High Spatial and Temporal Resolutions Based on a Synthetic Data and Machine Learning Approach. Front. Environ. Sci. 9. https://doi.org/10.3389/fenvs.2021.587660.
Kupssinskü, L.S., Guimarães, T.T., De Souza, E.M., Zanotta, D.C., Veronez, M.R., Gonzaga, L., Mauad, F.F., 2020. A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. Sens. (Switz.) 20 (7). https://doi.org/10.3390/s20072125.
Lacaux, J.P., Tourre, Y.M., Vignolles, C., Ndione, J.A., Lafaye, M., 2007. Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal. Remote Sens. Environ. 106 (1), 66–74. https://doi.org/10.1016/j.rse.2006.07.012.
Lee, T., Bettinger, P., Merry, K., Cieszewski, C., 2023. The effects of nearby trees on the positional accuracy of GNSS receivers in a forest environment. PLOS ONE 18 (3), e0283090. https://doi.org/10.1371/journal.pone.0283090.
Leggesse, E.S., Zimale, F.A., Sultan, D., Enku, T., Srinivasan, R., Tilahun, S.A., 2023. Predicting Optical Water Quality Indicators from Remote Sensing Using Machine Learning Algorithms in Tropical Highlands of Ethiopia. Hydrology 10 (5). https://doi.org/10.3390/hydrology10050110.
Lima, T.M.A., de, Giardino, C., Bresciani, M., Barbosa, C.C.F., Fabbretto, A., Pellegrino, A., Begliomini, F.N., 2023. Assessment of Estimated Phycocyanin and Chlorophyll-a Concentration from PRISMA and OLCI in Brazilian Inland Waters: A Comparison between Semi-Analytical and Machine Learning Algorithms. Remote Sens. 15 (5). https://doi.org/10.3390/rs15051299.
Liu, H., Li, Q., Shi, T., Hu, S., Wu, G., Zhou, Q., 2017. Application of Sentinel 2 MSI Images to Retrieve Suspended Particulate Matter Concentrations in Poyang Lake. Remote Sens. 9 (7), 761. https://doi.org/10.3390/rs9070761.
Lukhele, T., Msagati, T.A.M., 2024. Eutrophication of Inland Surface Waters in South Africa: An Overview. Int. J. Environ. Res. 18 (2), 27. https://doi.org/10.1007/s41742-024-00568-8.
Ma, Taquan, Zhang, Donghui, Li, Xusheng, Huang, Yao, Zhang, Lifu, Zhu, Zhenchang, Sun, Xuejian, Lan, Ziyue, Guo, Wei, 2023. Hyperspectral remote sensing technology for water quality monitoring: knowledge graph analysis and Frontier trend (Frontiers Media SA). Front. Environ. Sci. 11. https://doi.org/10.3389/fenvs.2023.1133325.
Ma, Y., Song, K.S., Wen, Z.D., Liu, G., Shang, Y.X., Lyu, L.L., Du, J., Yang, Q., Li, S.J., Tao, H., Hou, J.B., 2021. Remote Sensing of Turbidity for Lakes in Northeast China Using Sentinel-2 Images With Machine Learning Algorithms. IEEE J. Sel. Top. Appl. EARTH OBSERVATIONS REMOTE Sens. 14, 9132–9146. https://doi.org/10.1109/JSTARS.2021.3109292.
Maciel, D.A., Barbosa, C.C.F., Novo, E.M.L.D.M., Flores Júnior, R., Begliomini, F.N., 2021. Water clarity in Brazilian water assessed using Sentinel-2 and machine learning methods. ISPRS J. Photogramm. Remote Sens. 182, 134–152. https://doi.org/10.1016/j.isprsjprs.2021.10.009.
Maier, P.M., Keller, S., Hinz, S., 2021. Deep learning with wasi simulation data for estimating chlorophyll a concentration of inland water bodies. Remote Sens. 13 (4), 1–27. https://doi.org/10.3390/rs13040718.
Makhubela, T.V., Kramers, J.D., Scherler, D., Wittmann, H., Dirks, P.H.G.M., Winkler, S.R., 2019. Effects of long soil surface residence times on apparent cosmogenic nuclide denudation rates and burial ages in the Cradle of Humankind, South Africa. Earth Surf. Process. Landf. 44 (15), 2968–2981. https://doi.org/10.1002/esp.4723.
Matyukira, C., Mhangara, P., 2023. Land Cover and Landscape Structural Changes Using Extreme Gradient Boosting Random Forest and Fragmentation Analysis. Remote Sens. 15 (23). https://doi.org/10.3390/rs15235520.
Mishra, D., Schaeffer, B.A., Keith, D., 2014. Performance evaluation of normalized difference chlorophyll index in northern Gulf of Mexico estuaries using the Hyperspectral Imager for the Coastal Ocean. GIScience & Remote Sens. 51 (2), 175–198. https://doi.org/10.1080/15481603.2014.895581.
Mishra, S., Mishra, D.R., 2012. Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sens. Environ. 117, 394–406. https://doi.org/10.1016/j.rse.2011.10.016.
Mpakairi, K.S., Muthivhi, F.F., Dondofema, F., Munyai, L.F., Dalu, T., 2024. Chlorophyll-a unveiled: unlocking reservoir insights through remote sensing in a subtropical reservoir. Environ. Monit. Assess. 196 (4). https://doi.org/10.1007/s10661-024-12554-w.
Mugova, E., Wolkersdorfer, C., 2022. Identifying potential groundwater contamination by mining influenced water (MIW) using flow measurements in a sub-catchment of the "Cradle of Humankind" Unesco World Heritage Site, South Africa. Environ. Earth Sci. 81 (3). https://doi.org/10.1007/s12665-022-10224-z.
Najafzadeh, M., Basirian, S., 2023. Evaluation of River Water Quality Index Using Remote Sensing and Artificial Intelligence Models. Remote Sens. 15 (9), 2359. https://doi.org/10.3390/rs15092359.
Neil, C., Spyrakos, E., Hunter, P.D., Tyler, A.N., 2019. A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sens. Environ. 229, 159–178. https://doi.org/10.1016/j.rse.2019.04.027.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Matias, Y., 2022. Flood forecasting with machine learning models in an operational framework. Hydrol. Earth Syst. Sci. 26 (15), 4013–4032. https://doi.org/10.5194/hess-26-4013-2022.

News24, 2024. Four Zimbabwe rhinos die after drinking polluted water (December). News 24.

Ngamile, S., Kganyago, M., Madonsela, S., Mvandaba, V., 2025. Characterising the spatio-temporal patterns of water quality parameters in the cradle of humankind world heritage site using Sentinel-2 and random forest regressor. Front. Remote Sens. 6. https://doi.org/10.3389/frsen.2025.1631403.

Ngamile, S., Madonsela, S., Kganyago, M., 2025. Trends in remote sensing of water quality parameters in inland water bodies: a systematic review. Front. Environ. Sci. 13. https://doi.org/10.3389/fenvs.2025.1549301.

Nguyen, H.Q., Ha, N.T., Nguyen-Ngoc, L., Pham, T.L., 2021. Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: a case study in the Tri An Reservoir, Vietnam. Water Environ. Res. 93 (12), 2941–2957. https://doi.org/10.1002/wer.1643.

Ngwenya, N., Bangira, T., Sibanda, M., Kebede Gurmessa, S., Mabhaudhi, T., 2025. UAV-based remote sensing of chlorophyll-a concentrations in inland water bodies: a systematic review. Geocarto Int. 40 (1). https://doi.org/10.1080/10106049.2025.2452246.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., Alikas, K., Kangro, K., Gurlin, D., Hà, N., Oppelt, N., Stumpf, R., 2020. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. Remote Sens. Environ. 240. https://doi.org/10.1016/j.rse.2019.111604.

Pahlevan, N., Smith, B., Alikas, K., Anstee, J., Barbosa, C., Binding, C., Bresciani, M., Cremella, B., Giardino, C., Gurlin, D., Fernandez, V., Jamet, C., Kangro, K., Lehmann, M.K., Loisel, H., Matsushita, B., Hà, N., Olmanson, L., Potvin, G., Ruiz-Verdù, A., 2022. Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. REMOTE Sens. Environ. *270*. https://doi.org/10.1016/j.rse.2021.112860.

Prior, E.M., O'Donnell, F.C., Brodbeck, C., Donald, W.N., Runion, G.B., Shepherd, S.L., 2020. Measuring high levels of total suspended solids and turbidity using small unoccupied aerial systems (sUAS) multispectral imagery. Drones 4 (3), 54. https://doi.org/10.3390/drones4030054.

Rahat, S.H., Steissberg, T., Chang, W., Chen, X., Mandavya, G., Tracy, J., Wasti, A., Atreya, G., Saki, S., Bhuiyan, M.A., Ray, P., 2023. Remote sensing-enabled machine learning for river water quality modeling under multidimensional uncertainty. Sci. TOTAL Environ. 898. https://doi.org/10.1016/j.scitotenv.2023.165504.

Rathupetsane, E.M., & Kganyago, M. (n.d.). *Optimising Sentinel-2 Spectral Indices for Seasonal Estimation of Chlorophyll-a and Total Suspended Solids in Optically Complex Inland*.

Rathupetsane, E.M., Kganyago, M., Madonsela, S., & Mvandaba, V. (n.d.). *Performance of Machine Learning Algorithms and Multi-sensor Data 1 in the Retrieval of Optically-active Water Quality Parameters in 2 Inland Water Bodies. A Systematic Review*.

Rodríguez-López, L., Usta, D.B., Duran-Llacer, I., Alvarez, L.B., Yépez, S., Bourrel, L., Frappart, F., Urrutia, R., 2023. Estimation of water quality parameters through a combination of deep learning and remote sensing techniques in a Lake in Southern Chile. REMOTE Sens. 15 (17). https://doi.org/10.3390/rs15174157.

Rodríguez-López, L., Duran-Llacer, I., Bravo Alvarez, L., Lami, A., Urrutia, R., 2023. Recovery of water quality and detection of algal blooms in Lake Villarrica through landsat satellite images and monitoring data. Remote Sens. 15 (7). https://doi.org/10.3390/rs15071929.

Rubin, H.J., Lutz, D.A., Steele, B.G., Cottingham, K.L., Weathers, K.C., Ducey, M.J., Palace, M., Johnson, K.M., Chipman, J.W., 2021. Remote sensing of lake water clarity: performance and transferability of both historical algorithms and machine learning. Remote Sens. 13 (8). https://doi.org/10.3390/rs13081434.

Rudorff, N., Rudorff, C.M., Kampel, M., Ortiz, G., 2018. Remote sensing monitoring of the impact of a major mining wastewater disaster on the turbidity of the Doce River plume off the eastern Brazilian coast. ISPRS J. Photogramm. Remote Sens. 145, 349–361. https://doi.org/10.1016/j.isprsjprs.2018.02.013.

Ruescas, A.B., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K., Camps-Valls, G., 2018. Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI Simulated Data. Remote Sens. 10 (5), 786. https://doi.org/10.3390/rs10050786.

Saberioon, M., Brom, J., Nedbal, V., Souček, P., Císař, P., 2020. Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters. Ecol. Indic. 113. https://doi.org/10.1016/j.ecolind.2020.106236.

Saberioon, M., Khosravi, V., Brom, J., Gholizadeh, A., Segl, K., 2023. Examining the sensitivity of simulated EnMAP data for estimating chlorophyll-a and total suspended solids in inland waters. Ecol. Inform. 75. https://doi.org/10.1016/j.ecoinf.2023.102058.

Sahin, Z.M., Erten, E., Kaya, G.T., 2019. Multi-output regressions for estimating canola biophysical parameters from PolSAR data. 8th Int. Conf. AgroGeoinformatics (AgroGeoinformatics 2019, 1–4. https://doi.org/10.1109/Agro-Geoinformatics.2019.8820646.

Saranathan, A.M., Smith, B., Pahlevan, N., 2023. Per-pixel uncertainty quantification and reporting for satellite-derived chlorophyll-a estimates via mixture density networks. IEEE Trans. Geosci. REMOTE Sens. 61. https://doi.org/10.1109/TGRS.2023.3234465.

Shen, M., Luo, J., Cao, Z., Xue, K., Qi, T., Ma, J., Liu, D., Song, K., Feng, L., Duan, H., 2022. Random forest: An optimal chlorophyll-a algorithm for optically complex inland water suffering atmospheric correction uncertainties. J. Hydrol. 615. https://doi.org/10.1016/j.jhydrol.2022.128685.

Singh, R., Saritha, V., Pande, C.B., 2024. Monitoring of wetland turbidity using multi-temporal Landsat-8 and Landsat-9 satellite imagery in the Bisalpur wetland, Rajasthan, India. Environ. Res. 241, 117638. https://doi.org/10.1016/j.envres.2023.117638.

Smania, G., Jonsson, E.N., 2021. Conditional distribution modeling as an alternative method for covariates simulation: Comparison with joint multivariate normal and bootstrap techniques. CPT Pharmacomet. & Syst. Pharmacol. 10 (4), 330–339. https://doi.org/10.1002/psp4.12613.

Sommer, U., Gliwicz, Z.M., Lampert, W., Duncan, A., 1986. The PEG-model of seasonal succession of planktonic events in fresh waters. Arch. F. üR. Hydrobiol. 106 (4), 433–471. https://doi.org/10.1127/archiv-hydrobiol/106/1986/433.

Teutonico, D., Musuamba, F., Maas, H.J., Facius, A., Yang, S., Danhof, M., Della Pasqua, O., 2015. Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques. Pharm. Res. 32 (10), 3228–3237. https://doi.org/10.1007/s11095-015-1699-x.

Theenathayalan, V., Sathyendranath, S., Kulk, G., Menon, N., George, G., Abdulaziz, A., Selmes, N., Brewin, R.J.W., Rajendran, A., Xavier, S., Platt, T., 2022. Regional Satellite Algorithms to Estimate Chlorophyll-a and Total Suspended Matter Concentrations in Vembanad Lake. Remote Sens. 14 (24). https://doi.org/10.3390/rs14246404.

Tian, D., Zhao, X., Gao, L., Liang, Z., Yang, Z., Zhang, P., Wu, Q., Ren, K., Li, R., Yang, C., Zhang, Z., Chen, J., 2024. Estimation of water quality variables based on machine learning model and cluster analysis-based empirical model using multi-source remote sensing data in inland reservoirs, South China. Environ. Pollut. 342. https://doi.org/10.1016/j.envpol.2023.123104.

Tian, S., Guo, H., Huang, J.J., Zhu, X., Zhang, Z., 2022. Comprehensive comparison performances of Landsat-8 atmospheric correction methods for inland and coastal waters. Geocarto Int. 37 (27), 15302–15323. https://doi.org/10.1080/10106049.2022.2097320.

Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., Nõges, T., 2016. First experiences in Mapping Lake water quality parameters with sentinel-2 MSI imagery. Remote Sens. 8 (8), 640. https://doi.org/10.3390/rs8080640.

Toming, K., Liu, H., Soomets, T., Uuemaa, E., Nõges, T., Kutser, T., 2024. Estimation of the biogeochemical and physical properties of lakes based on remote sensing and artificial intelligence applications. Remote Sens. 16 (3). https://doi.org/10.3390/rs16030464.

Van Nguyen, M., Lin, C.-H., Chu, H.-J., Muhamad Jaelani, L., Aldila Syariz, M., 2019. Spectral feature selection optimization for water quality estimation. Int. J. Environ. Res. Public Health 17 (1), 272. https://doi.org/10.3390/ijerph17010272.

Varvia, P., Räty, J., Packalen, P., 2023. mgpr: An R package for multivariate Gaussian process regression. SoftwareX 24, 101563. https://doi.org/10.1016/j.softx.2023.101563.

Wolters, E., Toté, C., Sterckx, S., Adriaensen, S., Henocq, C., Bruniquel, J., Scifoni, S., Dransfeld, S., 2021. iCOR Atmospheric Correction on Sentinel-3/OLCI over Land: Intercomparison with AERONET, RadCalNet, and SYN Level-2. Remote Sens. 13 (4), 654. https://doi.org/10.3390/rs13040654.

Yang, Y., Jin, S., 2023. Long-Time Water Quality Variations in the Yangtze River from Landsat-8 and Sentinel-2 Images Based on Neural Networks. Water (Switz.) 15 (21). https://doi.org/10.3390/w15213802.

Yim, I., Shin, J., Lee, H., Park, S., Nam, G., Kang, T., Cho, K.H., Cha, Y.K., 2020. Deep learning-based retrieval of cyanobacteria pigment in inland water for in-situ and airborne hyperspectral data. Ecol. Indic. 110. https://doi.org/10.1016/j.ecolind.2019.105879.

Zehra, N., 2021. Prediction analysis of floods using machine learning algorithms (NARX & SVM). Int. J. Sci. Basic Appl. Res. ⟨http://gssrr.org/index.php?journal=JournalOfBasicAndApplied⟩.

Zhang, S.Y., Yinglan, A., Wang, L.B., Wang, Y.T., Zhang, X.J., Zhu, Y., Ma, G.W., 2024. Monitoring of low Chl-a concentration in hulun lake based on fusion of remote sensing satellite and ground observation data. REMOTE Sens. 16 (10). https://doi.org/10.3390/rs16101811.

Zheng, Z.J., Jiang, Y.Z., Zhang, Q.T., Zhong, Y.L., Wang, L.Z., 2024. A Feature selection method based on relief feature ranking with recursive feature elimination for the inversion of urban river water quality parameters using multispectral imagery from an unmanned aerial vehicle. Water 16 (7). https://doi.org/10.3390/w16071029.

Zolfaghari, K., Pahlevan, N., Simis, S.G.H., O'Shea, R.E., Duguay, C.R., 2023. Sensitivity of remotely sensed pigment concentration via Mixture Density Networks (MDNs) to uncertainties from atmospheric correction. J. Gt. LAKES Res. 49 (2), 341–356. https://doi.org/10.1016/j.jglr.2022.12.010.