MONITORING SURFACE WATER QUALITY IN THE CRADLE OF HUMANKIND WORLD HERITAGE SITE USING REMOTE SENSING TECHNOLOGY

Report

to the Water Research Commission

by

Kganyago M[1], Mvandaba V[2], Madonsela S[3]

[1] Department of Geography, Environmental Management and Energy Studies, University of Johannesburg

[2] Smart Places Cluster, Water Research Centre, Council for Scientific and Industrial Research

[3] Advanced Agriculture and Food Cluster, Council for Scientific and Industrial Research

WRC report no. 3260/1/26
ISBN 978-0-6392-0799-5

June 2026

This is the final report of WRC project No. C2023/2024-01241

EXECUTIVE SUMMARY

This is the Final Report for the project *Monitoring Surface Water Quality in the Cradle of Humankind World Heritage Site using Remote Sensing Technology*, Project No: C2023/2024-01241. The report details all work done for the project.

BACKGROUND

South Africa faces significant challenges regarding its surface water resources in terms of quantity and quality. Water quality in the Cradle of Humankind World Heritage Site (COHWHS) is severely compromised by factors including acid mine drainage (AMD) emanating from the Western Basin gold mining, as well as bacterial contamination and elevated nutrient levels originating from municipal sewage effluent and agricultural runoff. These pressures raise concerns for stakeholders in the area and pose severe risks to downstream water users, including the Hartbeespoort Dam, which battles water hyacinth infestation. Traditional water quality monitoring methods, relying on laboratory and in situ testing, are spatially and temporally limited, making them inadequate for comprehensive and regular monitoring. Therefore, integrating conventional methods with advanced satellite-based Earth Observation (EO) technologies and machine learning (ML) algorithms is proposed in this project as a viable and cost-effective solution to monitor water quality trends over large geographical areas.

AIMS

The objectives of the present project are as follows:

1. Determine the sensitivity of Sentinel-2 data to spatial and seasonal trends of surface water quality in the rivers and dams in and around COHWHS.
2. Use existing Sentinel-2 and field measurements to establish a five-year historical spatial profile of surface water quality with a focus on optically active water quality parameters.
3. Test the transferability of Sentinel-2 derived models to dammed waterbodies across different seasons.

METHODOLOGY

The surface water quality monitoring initiative under the COHWHS involved a structured campaign with several objectives. This included synchronous surface water sampling and data collection during satellite overpasses, purposive sampling across various affected water bodies, and gathering a minimum of 20 samples to ensure comprehensive data coverage. The inaugural wet season sampling occurred on 14 and 16 March 2024, while the dry season data collection happened on 28 and 31 August 2024, with a focus on both optically active parameters (OAPs) and non-optically active water quality parameters (NOAPs). The dates were carefully chosen to coincide with Sentinel-2 satellite overpasses and the water samples were collected following standard protocols and analysed in the laboratory over 72 hours. Several ML algorithms with varying sophistications were exploited, including linear regression, random forest, and Gaussian process regression (GPR). Moreover, multi-output versions of these algorithms were also explored and tested for their transferability to dammed water bodies.

DATA ANALYSIS

Sensitivity analysis using simple linear regression (SLR) and random forest regression (RFR) indicated that dissolved oxygen (DO) had the most substantial influence on Sentinel-2 band sensitivity during the high-flow period (i.e., wet season), while this relationship declined during the low-flow (i.e., dry) season, especially for parameters such as DO and pH. The next challenge we tackled was to generate a five-year profile of water quality trends in the study area. However, the lack of adequate training data limits the extrapolation of the ML algorithms to unknown periods, characterised by a varying range of water quality concentrations. To overcome this, synthetic data was generated using multivariate empirical distribution modelling to simulate statistical properties and inter-variable dependencies required for robust ML model training. Hence, GPR models developed for OAPs achieved a 75% accuracy for chlorophyll-*a* (Chl-*a*) and 73% accuracy for total suspended solids (TSS). These models demonstrated adaptability in creating a five-year historical spatial profile, successfully adjusting to historical changes in OAP concentrations, albeit results were not validated owing to limited data availability. The project then explored the multi-output algorithms (namely, multi-output Gaussian process regression [MGPR] and random forest for survival, regression and classification [RFSRC]) based on the recognition that water constituents do not exist in isolation; they rather interact and are interdependent. The MGPR model proved to be the superior predictive tool, yielding exceptional coefficients of determination (R^2) of 0.95 for Chl-*a* and 0.89 for TSS in the wet season, while retaining high accuracy for TSS ($R^2 = 0.94$) in the dry season. This performance highlights the robustness of MGPR in modelling interdependent water quality variables within heterogeneous aquatic systems. The RFSRC model also proved valuable and exhibited high transferability, achieving over 94% accuracy when tested on external dammed waterbodies for TSS estimation.

GENERAL

Based on the achievements of this research project, the following key take-home messages:

- The integration of high-resolution Sentinel-2 Multispectral Imagery with advanced ML techniques offers an effective and scalable solution for water quality monitoring, addressing limitations inherent in traditional monitoring systems.
- The analysis confirmed that Sentinel-2 spectral bands exhibit sensitivity to water quality parameters, although this relationship was observed to decline during the low-flow season compared to the higher relationships found during high-flow conditions for parameters like DO and pH.
- The MGPR model proved to be the superior predictive tool, explaining the greatest variability in OAPs (R^2) in both seasons.
- Given the demonstrated robustness of the multi-output ML algorithms (MGPR and RFSRC), it is essential to conduct extensive testing of their transferability across a broader range of water bodies, both within other areas in South Africa, to fully evaluate their generalisation capabilities.
- This evaluation should include assessing their performance on other water quality parameters beyond Chl-*a* and TSS, extending the analysis over multi-year and seasonal scales.
- Furthermore, the testing of model transferability should be supported by a larger validation dataset derived from field measurements to ensure reliability for operational deployment. This can be achieved through the alignment of field monitoring efforts.

ACKNOWLEDGEMENTS

The project team wishes to thank the following people for their contributions to the project:

Reference Group	Affiliation
Dr Eunice Ubomba-Jaswa	Research Manager, Water Research Commission (WRC)
Ms Penny Jaca	Project Coordinator, WRC
Prof. Mohamed Ahmed	Reference Group, Sol Plaatjie University (SPU)
Prof. Timothy Dube	Reference Group, University of the Western Cape (UWC)
Dr Torsten Bondo	Reference Group, DHI Group
Mr Simon Moloele	Reference Group, Department of Water and Sanitation (DWS)
Ms Ledile Jeanette Nyama	Reference Group, DWS
Others	
Prof Abel Ramoelo	Project Team, South African National Space Agency (SANSA)
Dr Oupa Malahlela	Project Team, University of Venda (UNIVEN)
Ms Sinesipho Ngamile	Project Team, University of Johannesburg (UJ)
Ms Elizabeth Modjadji Rathupetsane	Project Team, UJ

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

AC	atmospheric correction
AMD	acid mine drainage
C2RCC	Case 2 Regional CoastColour
Chl-*a*	chlorophyll-*a*
COHWHS	Cradle of Humankind World Heritage Site
CSIR	Council for Scientific and Industrial Research
DO	dissolved oxygen
DWS	Department of Water and Sanitation
EC	electrical conductivity
EO	Earth Observation
ETM+	Enhanced Thematic Mapper Plus
GERS	Gauteng Environmental Research Symposium
GPR	Gaussian process regression
ICEES	International Conference on Earth and Environmental Sciences
IoT	Internet of Things
IQR	interquartile range
MAE	mean absolute error
MAPE	mean absolute percentage error
MCLM	Mogale City Local Municipality
MEDM	multivariate empirical distribution modelling
MERIS	Medium Resolution Imaging Spectrometer
MGPR	multi-output Gaussian process regression
ML	machine learning
MMC	Member of the Mayoral Committee
MoA	Memorandum of Agreement

MPN	most probable number
MSc	Master of Science
MSE	mean squared error
MSI	MultiSpectral Instrument
MSS	Multispectral Scanner
MUG	4-methylumbelliferyl-β-D-glucuronide
NDI	normalised difference index
NGO	non-governmental organisation
NIR	near-infrared
NOAP	non-optically active water quality parameter
NWU	North-West University
OAP	optically active parameter
OC-SMART	Ocean Color Simultaneous Marine and Aerosol Retrieval Tool
OLI	Operational Land Imager
ONPG	o-nitrophenyl-β-D-galactopyranoside
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RBF	radial basis function
RFR	random forest regression
RFSRC	random forest survival, regression, and classification
RI	ratio index
RMSE	root mean squared error
RPD	residual prediction deviation
SASAS	South African Society for Atmospheric Sciences
SDG	Sustainable Development Goal
SLR	simple linear regression
SNAP	Sentinel Application Platform
SSAG	Society of South African Geographers

SWIR	shortwave infrared
TM	Thematic Mapper
TP	total phosphorus
TSS	total suspended solids
UJ	University of Johannesburg
WoS	Web of Science
WRC	Water Research Commission
WWTW	Waste Water Treatment Works

CHAPTER 1: BACKGROUND

1.1 INTRODUCTION

South Africa is a water-scarce country with an annual average rainfall of 490 mm. Despite this, the country has a poor record of water conservation. Moreover, the World Wide Fund – South Africa 2016 Water Report also revealed that the water quality in the country's rivers has deteriorated significantly over the past 20 years (Colvin et al., 2016). It is the quality of water that determines its utility for human consumption, agricultural use, or supporting aquatic life. The mismanagement of the water resources around the Cradle of Humankind World Heritage Site (COHWHS) has raised concerns among stakeholders and communities that depend on groundwater for drinking and irrigation purposes (Durand et al., 2010). These concerns arise owing to the harmful impacts of acid mine drainage (AMD) emanating from the Western Basin gold mining void on the Witwatersrand; this poses severe risks to the health and well-being of humans, animals, the environment, and the overall status of the COHWHS (Durand et al., 2010; Van Deventer et al., 2014).

Moreover, surface water quality in the Bloubankspruit water streams is further impacted negatively by bacterial contamination and associated elevated nitrate and phosphate levels coming from the municipal sewage effluent and agricultural activities. This high nutrient load delivered by the Bloubankspruit and Crocodile River is a cause for concern for the downstream water users, including the Hartbeespoort Dam (Hobbs, 2011). Bacterial contamination and the presence of mine water present a fundamental threat to the fitness for use of water resources in the COHWHS region. This makes water quality monitoring in key waterbodies such as lakes, dams, and rivers (among others) a necessary activity to ensure that available water meets the set standards for various uses. As the custodian of South Africa's water resources, the Department of Water and Sanitation (DWS) has water monitoring stations along the Magalies River, Crocodile River, and Bloubankspruit to measure the quality and quantity of surface water resources (Holland et al. in Buchanan, 2010). Water quality monitoring focuses on biological, chemical, and physical parameters of water and has been completed through laboratory tests, and in situ testing (Ahmed et al., 2020). Laboratory and in situ tests involve collection of water samples at key, accessible locations for analysing physical, chemical, and biological indicators of water quality (Jaji et al., 2007).

Although in situ sampling enables testing of a variety of water quality indicators, it is not ideal for regular monitoring of water resources over large areas since it is point-based and therefore limited spatially and temporally. This means that in situ monitoring cannot provide detailed spatial or temporal trends of water quality in waterbodies (Gholizadeh et al., 2016). Furthermore, Hobbs (2011) criticised the DWS's monitoring efforts as being biased towards surface water resources and restricted to a small geographic area. This emphasises the need for catchment-level river health monitoring programmes that carry out monitoring activities on a regular basis to determine the trends in the surface water quality and quantity (Holland et al., in Buchanan, 2010). The availability of satellite sensors with high temporal and spatial resolution presents an opportunity to monitor water quality regularly and over large geographic areas using spectral data onboard satellites (Dube et al., 2015; Pizani et al., 2020).

The application of remote sensing relies on the spectral characteristics of water quality parameters, which are detectable in certain spectral regions of the electromagnetic spectrum present in satellite sensors. The presence of parameters (such as total suspended solids [TSS], chlorophyll-*a* [Chl-*a*] concentration, turbidity, salinity, total phosphorus [TP], temperature, pH, and dissolved organic carbon) makes it possible to use remote sensing technology directly or indirectly for monitoring water quality (Gholizadeh et al., 2016). Clear water has a unique spectral signature defined largely by its high reflectance in the visible region and high absorption in the near shortwave infrared regions. However, optically active parameters (OAPs) interact with solar radiation and alter the spectral signature of water, making it possible to use the spectral signal as a proxy for water quality. Importantly, the application of remote sensing technology in water quality monitoring still needs to be complemented by in situ sampling, computing power capable of handling a large dataset and efficient modelling techniques (Gholizadeh et al., 2016). The nature of remote sensing data itself, in terms of spatial and spectral settings, can determine the level of success or accuracy achievable. In this report, a review of remote sensing and ML techniques application in water quality assessment is presented. The review provides a summary of the current research, satellite technologies, and approaches; it also identifies gaps and limitations and identifies emerging trends and prospects for future research.

These satellite sensors use remote sensing techniques that have developed rapidly, and these techniques have been used widely for monitoring large-scale variations in water quality (Li et al., 2022). As compared to traditional methods for water quality monitoring, satellite remote sensing is more cost-effective, less time-consuming and has been effective in detecting optically active water quality parameters (Li et al., 2022). Remote sensing applications provide synoptic and temporal coverage of water bodies that traditional methods are not able to attain and, in turn, fill the global gap of spatiotemporal data, making them ideal for the cost-effective monitoring of water quality (Adjovu et al., 2023). Therefore, there is a need to integrate traditional, field-based methods with satellite-based Earth Observation (EO) to explore the prospects of upscaling field measurements and reducing the burden and cost of frequent fieldwork. This also increases the accuracy in predicting impending water quality outbreaks (Sagan et al., 2019). Water managers and socio-economic policymakers deal with risk and uncertainty related to water quality. As decision-makers, they need to plan, design, and manage water systems to reduce risks associated with water quality, and the more these risks and uncertainties are understood, the easier this task becomes for them (Agnoli et al., 2023).

This is where satellite-based EO technologies play a role, and when they are linked with in situ data and algorithmic solutions, they can provide reliable and useful data on extensive temporal, spatial, and spectral scales; thus, also enabling proper water resources management, planning, design, and operation (IOCCG, 2018; Agnoli et al., 2023). Researchers have used multi- and hyperspectral sensors for several studies that include water and land observation monitoring. Multispectral sensors collect data in multiple, distinct spectral bands within the electromagnetic spectrum. This typically covers both the visible and near-infrared (NIR) regions, and their spectral bands range from 400 nm to 700 nm for red-green-blue, and infrared wavelengths within the range of 700 nm-2500 nm (Kurihara et al., 2018; Ji et al., 2020). Studies related to the estimation of several water quality parameters have used multispectral sensors, such as Landsat-8 Operational Land Imager (OLI) and Sentinel-2 MultiSpectral Instrument (MSI) sensors in monitoring water quality parameters in sub-Saharan Africa, and this is gaining momentum, as compared to the use of hyperspectral sensors, which are still evolving because of their high costs and limited availability in developing countries (Dube et al., 2015).

For these remote sensors to quantitively retrieve various water quality parameters, the water-leaving reflectance is calculated, and this may be complicated due to the radiation at various wavelengths transferred among the atmosphere, water surface, and water body (Yang et al., 2022). The retrieval of water quality parameters from these sensors is then calculated through empirical, analytical, semi-empirical, and ML retrieval methods (Yang et al., 2022). The ML algorithms such as neural networks, support vector machines, or random forests are used to learn complex relationships between input variables and water quality parameters derived from large datasets (Leggesse et al., 2023). The ML models have been applied in various studies related to water quality monitoring and water quality retrieval and have achieved relatively satisfactory results, as they are useful in complex and dynamic aquatic environments (Zhao et al., 2022). The ML models can be utilised in capturing both linear and non-linear data relationships when applied to hyperspectral data of water surfaces with complex backgrounds and multiple variables (such as water quality parameters and sediment deposits) (Zhao et al., 2022).

Models are better than physical and traditional statistical models as they provide flexible and versatile estimations (Fu et al., 2022). The ML algorithms can be utilised to monitor the concentration of various optically and NOAPs, and this is done by linking these algorithms with satellite sensor data and water quality sampling. Non-optically active parameters (NOAPs) are those that are less likely to have an impact on the optical properties measured by the remote sensing sensors, while OAPs are those that are likely to affect those characteristics (Yang et al., 2022). While NOAPs do not directly affect the spectral properties of water, they can indirectly influence the spectral signatures of water owing to their impact on the aquatic ecosystem (e.g., TN and TP are indicators of nutrient levels in the water, leading to an increase in algae) (Leggesse et al., 2023). Using ML algorithms has assisted in estimating water quality parameters in complex situations where a reflected signal contains information from more than one constituent and is spatially and temporally variable (Leggesse at el., 2023).

The estimation of these various water quality parameters using ML algorithms can assist in exploring the spatiotemporal distribution and patterns of the various optically and non-optically active physicochemical parameters. Various satellite sensors have been used for estimating water quality parameters, but the Sentinel-2 MSI sensor may prove to be more suitable for water quality monitoring because of the presence of a red-edge band, which enhances its sensitivity to Chl-*a* content in vegetation and water (Ndou, 2022). The objective of this report was to provide an overview of the current state of the art regarding water quality assessment using remote sensing and ML techniques. The report provides a summary of the current research, satellite technologies, and approaches, identifies gaps and limitations, and identifies emerging trends and prospects for future research.

1.2 PROJECT AIMS

The aims of the project were as follows:

1. Determine the sensitivity of Sentinel-2 data to spatial and seasonal trends of surface water quality in the rivers and dams in and around COHWHS.
2. Use existing Sentinel-2 data and field measurements to establish a spatial five-year historical profile of surface water quality with a focus on optically active water quality parameters.
3. Test the transferability of Sentinel-2 derived models to dammed waterbodies across different seasons, i.e., winter and summer.

1.3 SCOPE AND LIMITATIONS

The scope of this report is limited to the technical work related to the project aims, capacity building and community engagement activities.

CHAPTER 2: PROGRESS IN REMOTE SENSING OF WATER QUALITY USING MACHINE LEARNING ALGORITHMS

2.1 INTRODUCTION

This chapter provides an overview of the state of the art regarding water quality assessment using remote sensing and ML techniques. We synthesise the current research, satellite technologies, and approaches, identify gaps and limitations, and identify emerging trends and prospects for future research. Among the many water quality parameters which exist, we could only cover a few due to the scope of this project.

2.2 LITERATURE SEARCH AND DATA EXTRACTION

The literature reviewed in this study was selected based on *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) guidelines (Page et al., 2021). These guidelines were designed to facilitate transparency in the reporting of literature reviews in terms of methods for identifying, selecting, appraising, and synthesising various studies. This approach is relevant to the current report, which seeks to provide state- of-the-art remote sensing of water quality parameters.

2.2.1 Literature search

To search for relevant literature, the first step was to identify keywords, terms, and phrases used in search strings. The following keywords were identified: “water quality”, “water pollution”, “water quality parameters”, “remote sensing”, “inland water bodies”, “machine learning”, “vegetation indices”, “red-edge”, “Sentinel-2”, and “Landsat”. The SCOPUS Abstract and Citation Database (https://www.scopus.com) and Web of Science Core Collection Database (https://www.webofscience.com/) were utilised to access literature based on specified keywords combined using the Boolean operators “AND” and “OR”. These two databases were used because they constitute the most robust, reliable, and frequently used research engines, allowing the discovery of present and past indexed research and patented materials. Other databases, such as Google Scholar, were not used during the search owing to irregular updates and erroneous information, as anyone can upload articles. Literature searches from SCOPUS and Web of Science (WoS) databases returned 3 232 and 2 170 documents, respectively. The search string codes were then refined by filtering articles by date range, language, document type, and publication status, retaining only journal articles that were either “published” or “in press” between 2003 and 2023 and were written in English (See **Table 2-1**). Conference papers were excluded, as they are often republished as journal articles, as well as magazine articles, editorials, communications, letters to the editor, book chapters, and review articles.

Table 2-1. Keyword combinations used to search journal titles, abstracts, and keywords

Keywords	SCOPUS	WoS
Field tag* = ((“water quality” AND “Remote sensing” AND “inland water bodies”) OR (“machine learning” AND “remote sensing” AND “water quality”) OR (“water quality” AND (“vegetation indices” OR “red-edge”)) OR (“water quality” AND (“Sentinel-2” OR “Landsat” OR “MERIS” OR “MODIS” OR “optical remote sensing”)))	3 232	2 170
AND Article (Document Types) AND English (Languages) AND >2013 <2023 (Publication Years)	1 945	1 983
AND Engineering Electrical Electronic OR Oceanography or Limnology OR Meteorology Atmospheric Sciences OR Engineering	1 541	1 618

Civil OR Instruments Instrumentation OR Chemistry Analytical or Engineering Multidisciplinary OR Optics OR Chemistry Multidisciplinary (Exclude – Web of Science Categories)

* The field tag varied between SCOPUS and WoS. Field tags used for SCOPUS were TITLE-ABS-KEY, while the WoS equivalent was TS.

2.2.2 Literature inclusion and exclusion strategy

Apart from the search refinements performed (as described in Section 2.2.1), the articles were automatically screened for duplicates using titles as a criterion using the ***screen_titles*** function of the R package *revtools*. The ***screen_abstracts*** function was then used to screen abstracts for relevance. As a result, 552 articles were identified as duplicates or inappropriate document types. Next, we compared SCOPUS and WoS results and found 1 541 duplicates. The rest of the titles and abstracts of the remaining articles were screened for relevance, and 230 review articles were selected after manual verification. A total of 403 studies were read and included in this review, as shown in **Figure 2-1**.

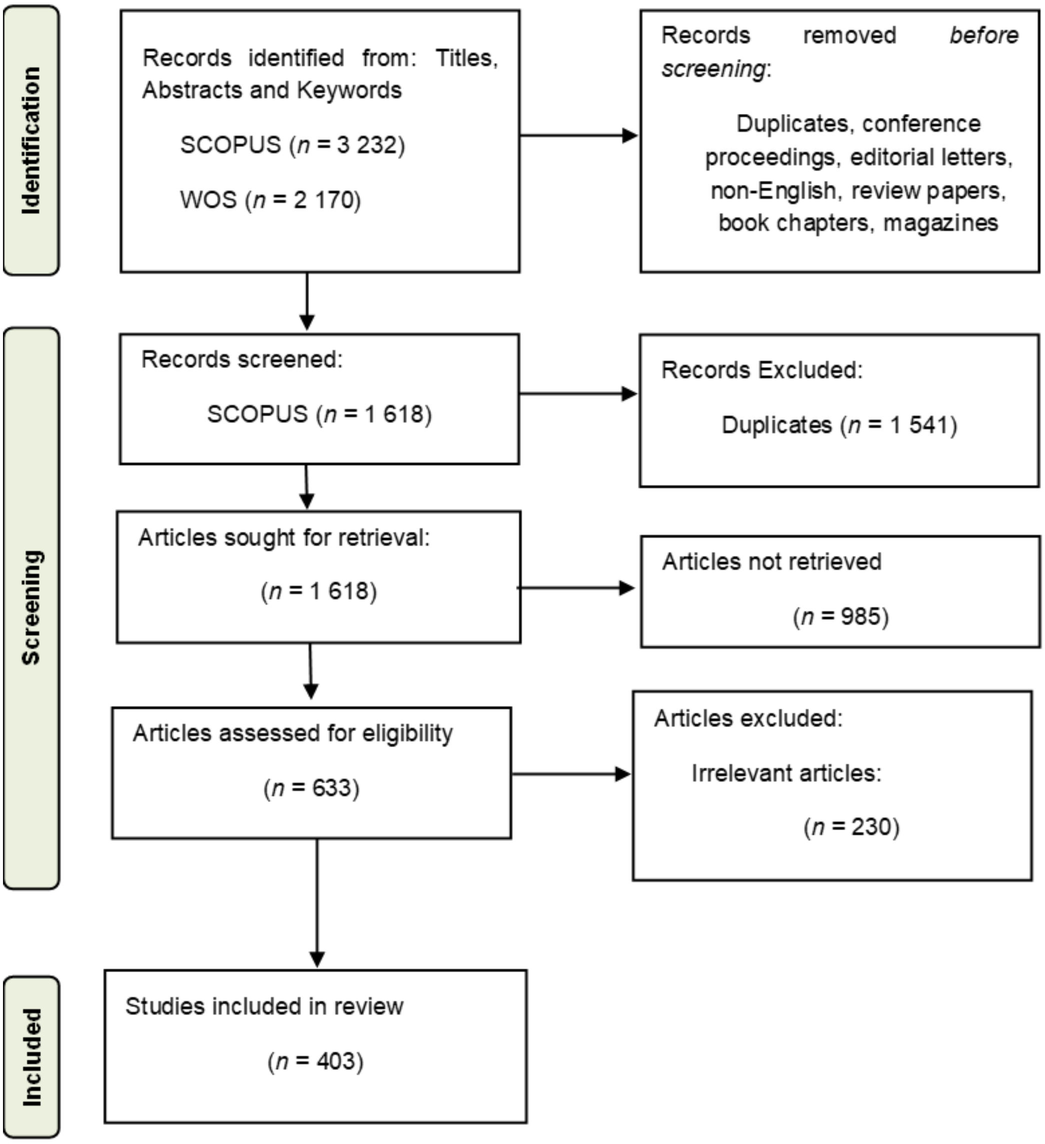


Figure 2-1. A flowchart indicating the number of relevant studies identified, screened, and included

2.2.3 Bibliometric analysis

Bibliometric analysis is a widely used meta-analytical tool that can identify interconnections of key terms related to a given topic or field from published papers. This was necessary for this study to identify and visualise the bibliometric occurrence and co-occurrence networks and clusters of keywords or terms. The final articles (i.e., titles, abstracts, and keywords) identified using the PRISMA guideline in 2.2.2 were subject to bibliographic analysis using VOSviewer (version 1.6.20) software (www.vosviewer.com). In addition, we analysed the basic frequencies and trends of publications to assess the progress of the application of remote sensing in water quality assessment using linear regression in Microsoft Excel and ArcGIS Pro 3.2.

2.3 RESULTS

2.3.1 Bibliometric analysis

The final articles show an increasing trend from 2013 to 2023; however, the number of publications declined in 2016 (See **Figure 2-2**). The most productive countries in terms of water quality research are shown in **Figure 2-3**, where China, the United States, and India are the three countries with the most research articles published during the period of this review.

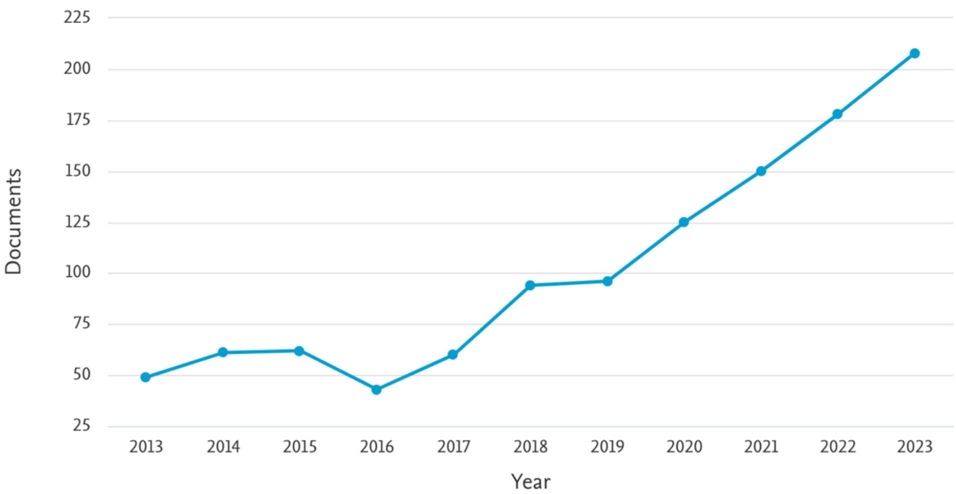


Figure 2-2. The trend in publications between 2013 and 2023

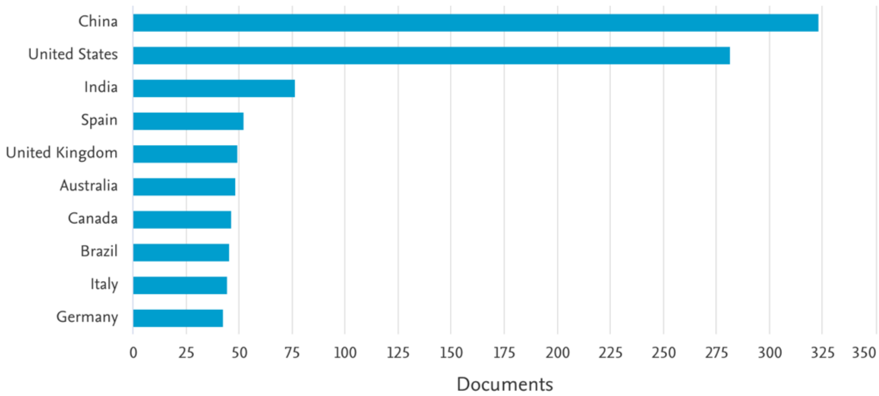


Figure 2-3. Number of publications by country

Figure 2-4 shows the major journals published on the remote sensing of water quality. The journal *Remote Sensing* started leading the others from 2018 to 2013, while previous years (that is, 2013-2015) were mainly dominated by *Remote Sensing of Environment*.

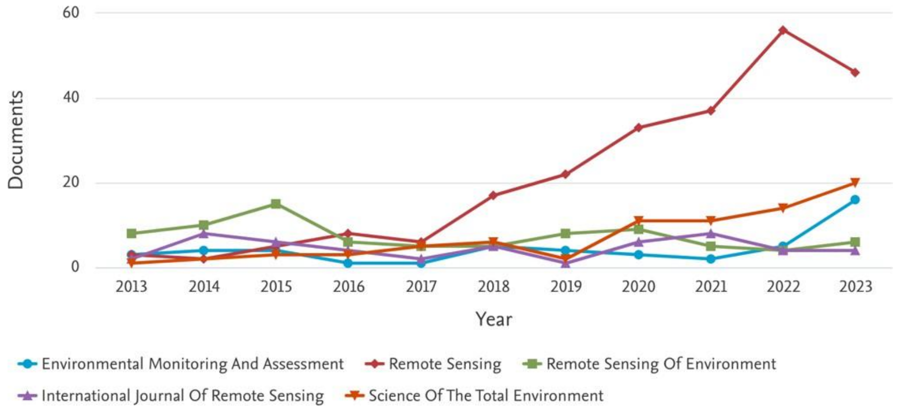


Figure 2-4. Number of relevant studies per publisher

Bibliometric analysis was performed using the meta-analytical tool, VOSviewer, to visualise the occurrence and co-occurrence networks of key terms from the retrieved literature. This allowed the identification of interconnections of key terms related to water quality and their clusters (See **Figure 2-5**). In **Figure 2-5**, the relative size of both the label and circle for an item is determined by the item's weight. The greater the weight of an item, the larger the label and the circle of the item. For some items, the label is not displayed to avoid overlapping labels. The colour of an item indicates the cluster to which the item belongs.

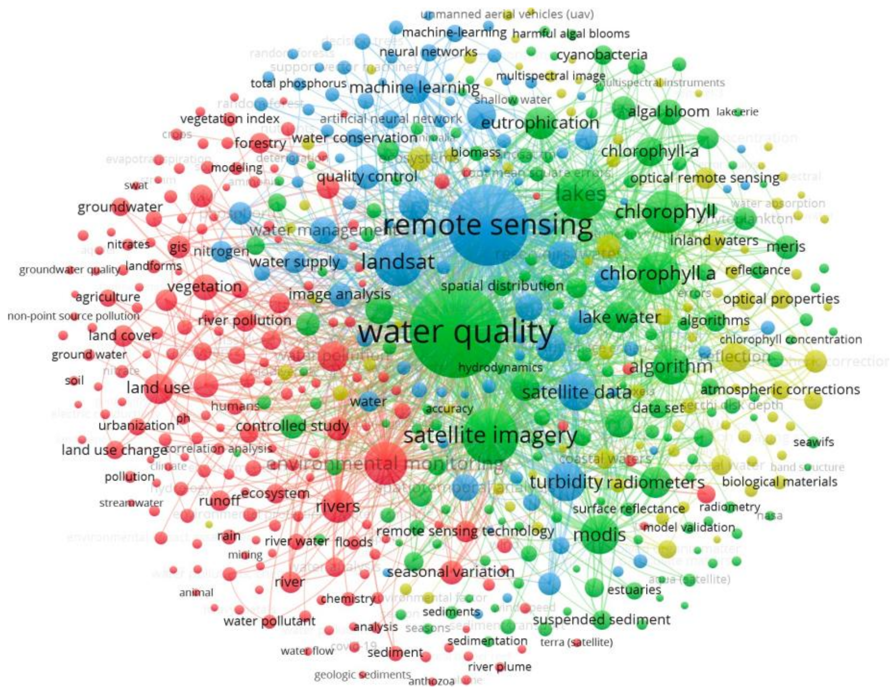


Figure 2-5. Clustering of key terms from the abstracts and titles of the included literature

The key terms from each cluster are outlined in **Table 2-2**. The main clusters can be grouped into themes or topics based on the strongly linked key terms. The first cluster focuses on the environment (land use, pollution) and climate change (drought) and its impact on the water quality. For example, it is clear that agricultural land uses are linked with the deposition of various biological and particulate matter into rivers, streams, and lakes. The second cluster focuses on the water quality key term, which is mainly linked to algal blooms, Chl-*a*, and cyanobacteria, and shows that these aspects have been studied with satellite imagery and in situ data collection approaches. The clusters are also linked with health risks and environmental regulations. The third cluster shows that remote sensing key terms are strongly linked with contemporary ML techniques such as artificial neural networks, deep learning, and random forest. It is also linked with new sensors such as Sentinel-2 and Sentinel-3 and cloud computing platforms such as Google Earth Engine. This indicates that recent studies preferred applying new technologies in a quest to improve water quality parameter estimation.

Table 2-2. Keyword clusters

Category	Topic	Keywords
Cluster 1	Environmental change and water quality monitoring	"environmental monitoring", "agricultural land", "agricultural pollution", "agriculture", "ammonia", "animals", "anthropogenic activity", "anthropogenic effect", "aquatic ecosystems", "climate", "aquatic environments", "dissolved oxygen", "drought", "effluents", "electrical conductivity", "geologic sediments", "heavy metals", "irrigation", "land use change", "ph", "particulate matter", "pollutant source", "precipitation", "runoff", "sewage", "total dissolved solids", "total suspended matter", "salinity"

Cluster 2	Water quality assessment with remote sensing	“water quality”, “Air temperature”, “algae”, “algal bloom”, “algorithms”, “annual variation”, “band ratios”, “chlorophyll-a”, “chlorophyll concentration”, “climate change”, “cyanobacteria”, “ecology”, “empirical analysis”, “enhanced thematic mapper”, “environmental parameters”, “environmental regulation”, “error analysis”, “eutrophication”, “fresh water”, “health risks”, imagery”, “in situ measurement”, “lake”, “landsat imagery”, “meris”, “model validation”, “modis”, “normalised difference water index”, “phytoplankton”, “phycocyanin”, “trophic state status”
Cluster 3	Remote sensing and ML techniques	“remote sensing”, “artificial intelligence”, “artificial neural networks”, “correlation methods”, “deep learning”, “drinking water”, “empirical algorithms”, “eutrophic environment”, forecasting”, “genetic algorithms”, “google earth engine”, “image analysis”, “image enhancement”, “landsat-8”, “learning systems”, “linear regression”, “machine learning algorithms”, “mapping method”, “multiple regression”, “neural networks”, “random forest”, “satellite data”, “secchi depth”, “sentinel-2”, “spatial variability”, “spectral indices”
Cluster 4	Reflectance characteristics	“reflectance”, “absorption”, “accuracy”, “accuracy assessment”, “aerosols”, “atmospheric correction”, “bathymetry”, “classification”, “coloured dissolved organic matter”, “earth observation”, “hyperspectral”, “ecosystems”, “inland waters”, “lake ecosystem”, “multispectral image”, “olci”, “optical properties”, “radiative transfer”, “sentinel-3”, “spectroscopy”, “turbid water”, “ unmanned vehicle”, “water absorption”, “water color”, “water depth”

2.3.2 Sources of water pollution

Water is polluted by microorganisms, and solid and liquid wastes, and this leads to the water colour being changed, as well as the taste and aroma. This then results in the water being unfit for consumption. Water pollutants can also cause various types of diseases in humans, animals, and plants. Anthropogenic activities such as urbanisation, agricultural practices, industrialisation, and population growth have led to the deterioration of water quality in many parts of the world (Singh and Gupta, 2016). Industries such as paper mills, iron and steel production, distilleries, and nuclear industries result in the release of toxic chemicals and organic and inorganic substances into the environment, increasing wastewater discharge (Singh and Gupta, 2016). In agriculture, pesticides, nitrogen fertilisers, and organic farm wastes increase water contamination through the washoff of nitrates, phosphorus, pesticides, and soil sediments (Parris, 2011; Lin et al., 2022). A study by Xu et al. (2022) clarified the main sources of pollution and their impact on river water quality. Their study in the Yangtze and Yellow River Basins in China revealed that industries were the major source of the country’s water pollution. Agricultural activities were also shown to cause water quality deterioration, but their impact was not as major as the impact caused by industries.

A further study (conducted by Lin et al., 2022) showed that globally, an estimated 80% of industrial and municipal wastewater is discharged into the environment (and hence reaches the world’s rivers) without any prior treatment. The situation is worse in developing regions such as sub-Saharan Africa. For example, Edokpayi et al. (2017) reported that the major factors responsible for the failing state of wastewater treatment facilities in developing countries may include poor operational state of wastewater infrastructure, design weaknesses, lack of expertise, and corruption (among others). South Africa’s Hartbeespoort Dam is an example of a reservoir where there is eutrophication because of industrial effluents from AMD, bacterially contaminated municipal wastewater discharge, and agricultural land use. The dam’s ecological integrity and its capacity to provide ecological services are therefore being compromised. Most countries in Africa lack basic sanitation, yet more than half a billion people still consume contaminated water, thus leading to the deaths of thousands of children. Water pollution affects not only the health of users but also the environment, where aquatic ecosystems are harmed. Water pollution leads to water scarcity and resource depletion as water utility reduces, thus impacting agricultural productivity, the economy, aquatic biodiversity, and fish yields. It also reduces tourism revenues because the polluted water is deemed unattractive (Al-Taai, 2021). Indeed, water pollution affects the quest to achieve the United Nations Sustainable Development Goals (SDGs) and associated targets such as SDG 6. Target 6.3 of the SDGs aims to "improve water quality by reducing pollution, eliminating dumping, and minimising the release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally" by 2030. Hence, to attain water resource sustainability (integrating traditional in situ and lab-based approaches and contemporary approaches such as satellites and ML algorithms) is needed. This will ensure that reliable and actionable information is available to facilitate informed water resources management.

2.3.3 Conventional methods for water quality assessments

Water quality monitoring in waterbodies such as lakes, dams, and rivers is a necessary activity to ensure that available water meets the set standards for various uses (Wanda et al., 2016; Sent et al., 2021). Assessment and monitoring of the concentration of pollutants in water have traditionally been completed through laboratory and in situ tests that focus on the biological, chemical, and physical parameters of water (Ahmed et al., 2020). Research conducted by Jaji et al. (2007) focused on the quality of the Ogun River in South-West Nigeria. Their approach involved collecting and analysing samples for physicochemical and bacteriological parameters (such as turbidity, pH, DO, TDS, nitrate, phosphate, oil, grease and faecal coliforms) as well as heavy metals using conventional methods involving water sampling and laboratory tests. The study concluded that the values of the parameters obtained were above the maximum acceptable limit (such as nitrate being no more than 50 mg/L) set by the World Health Organization (WHO) for drinking water.

In contrast, in situ testing involves preventing artefacts that can occur during sampling, transporting, and storing contaminated water and sediment for toxicity testing in a lab. The methods prevent artefacts from occurring owing to data being collected directly from the water body, leading to an understanding of the conditions and interactions within that environment in real-time. Both in situ and laboratory testing methods have been praised as being accurate for measuring water quality (Zainurin et al., 2022). However, both these conventional methods are criticised for being point-based and limited in temporal and spatial scales (Ndou, 2022). In situ and laboratory testing methods are labour-intensive, costly, and time-consuming and are unable to adequately address factors that influence the development of algal blooms and oxygen depletion (Ndou, 2022; Zainurin et al., 2022). According to Dube et al. (2015), there has been marked deterioration of water quality, particularly in sub-Saharan Africa; there is thus a need for monitoring water quality, the establishment of efficient control methods and the proposal of solutions for water protection. Owing to this motivation, other techniques have been globally developed to monitor water quality (Ahmed et al., 2020).

2.3.4 Remote sensing of water quality

The advent of satellite remote sensing in water quality monitoring has revolutionised the field by providing a systematic and synoptic perspective on water quality parameters over large water bodies, enabling efficient, cost-effective, and timely assessments of environmental changes and trends (Swain and Sahoo, 2017). The study of physical and chemical water quality parameters, such as turbidity and Chl-*a* over space and time, can, therefore, be achieved using remotely sensed data. Remote sensing techniques have proven to be better methods for monitoring water quality parameters, but these methods have not been used enough in developing countries (Lieu et al., 2022). Satellite remote sensing in water quality monitoring provides a tool to be used for monitoring inland water bodies due to its offering of an opportunity for data collection on systematic synoptic monitoring and temporal scales (Majozi et al., 2014).

Remote sensing is a combination of technology and science deployed to obtain information about an object, area, or phenomenon by analysing data collected by a technological device without being in contact with the object, area, or phenomenon under investigation. The sensing technological device collects information about an object, area, or phenomenon by measuring the object's transmission of electromagnetic energy from reflecting and radiating surfaces (Aggarwal, 2004). Different objects or surface features return varying amounts of energy across different channels of the electromagnetic spectrum, incident upon it. The amount of energy returned is defined by the structural, chemical, and/or biological properties of an object, surface roughness, angle of incidence, intensity, and wavelength of radiant energy (Aggarwal, 2004). The application of remote sensing in water quality assessment is premised on this understanding.

Remote sensing of water quality measures the physical, chemical, and biological characteristics of waterbodies (Bresciani et al., 2019; Sent et al., 2021) and offers the prospects of establishing the possible sources of water quality contamination with a catchment-level assessment of water quality (Usali & Ismail, 2010). As indicated earlier, water quality contamination emanates from various sources, including municipal waste discharges, chemical fertilisers, and pesticides used in agriculture, heavy metals, microorganisms, and sedimentation. These sources of contamination alter the chemical, physical, and biological characteristics of waterbodies, and this alteration affects the reflectance signal captured by the remote sensing satellites over water bodies. Using empirical or analytical models, the remotely sensed water leaving spectral signature can be related to the measured changes (i.e., using in situ and laboratory techniques) in a particular water quality parameter (Murray et al., 2022). Indeed, significant progress has been made in the remote sensing of water quality, especially with a focus on optically active water quality parameters such as Chl-*a*, TSS, Secchi disc depth, water temperature, and turbidity.

2.3.4.1 Chlorophyll-a

The concentration of Chl-*a* indicates the trophic status of a waterbody driven by nutrient enrichment, particularly phosphorus. Nutrient enrichment creates favourable conditions for algal production, and Chl-*a* is an essential component in photosynthesis to produce algal biomass (Brivio et al., 2001). Generally, Chl-*a* is found in plants, algae, and cyanobacteria, and it absorbs light energy in the blue and red channels of electromagnetic wavelengths and reflects in the green channel. Studies have demonstrated that Chl-*a* interaction with light energy can be used to estimate its concentration in waterbodies with varying levels of success. Pizani et al. (2020) used bands 2, 3, 4 and 5 of the Sentinel-2 sensor and bands 2, 3 and 4 of Landsat-8 in multiple regression models to estimate Chl-*a* in Tres Marias Reservoir and observed coefficients of determination (R^2) ranging from 0.71 to 0.75 for Sentinel-2 and 0.67 for Landsat-8. Brivio et al. (2001) used a band ratio from Landscape Thematic Mapper (TM) band 1, 2 and 3 to estimate chlorophyll concentration in Lake Garda, Italy in February 1992 and March 1993 with varying chlorophyll concentrations. This study reported that the regression model estimated chlorophyll concentration with a RMSE of 0.0374 in March (R^2 of 0.818) and an RMSE of 0.485 in February (R^2 of 0.67) when lower chlorophyll concentrations were measured in Lake Garda.

Moreover, Giardino et al. (2001) used TM bands 1 and 2 in a regression model to predict Chl-*a* concentrations in Lake Iseo, with an RMSE of 0.054 (R^2 of 0.999). This study had a very limited number of calibration data. Gitelson et al. (2008) used a three-band model featuring Medium Resolution Imaging Spectrometer (MERIS) spectral bands (660-670 nm, 703.75 713.75 nm, and 750-757.5 nm) and a two-band model using Moderate Resolution Imaging Spectroradiometer (MODIS) spectral bands (662-672 nm, 743-753 nm) to estimate Chl-*a* in turbid waters. The three-band model of MERIS predicted Chl-*a* with a normalised mean bias error of 18.3%, while the two-band model of MODIS predicted with a normalised mean bias error of 53.3%. Generally, the aforesaid studies indicate the applicability of remotely sensed data in Chl-*a* estimation in waterbodies. Consistent with this assertion, Matthews and Bernard (2015) reported close correspondence between Chl-*a* estimates from MERIS data and in situ measurement data of the National Eutrophication Monitoring Programme implemented in various waterbodies of South Africa.

However, it appears that the changing concentration of Chl-*a* with changes in seasons affects its estimation accuracy and that different sensors achieve varying levels of accuracy. This is supported by contrasting observations in Brivio et al. (2001) and Gitelson et al. (2008). The study by Brivio et al. (2001), using Landsat data, reported lower prediction accuracy for Chl-*a* when it is in lower concentrations. In South Africa's Vaal Dam, Malahlela et al. (2018), using Landsat-8 bands in a stepwise logistic model, reported a fluctuating trend in Chl-*a* mapping accuracy with time. In April 2014, their study mapped Chl-*a* concentration with 76.8% accuracy, while in May 2015, the accuracy dropped to 35.01%.

These variations in Chl-*a* estimation accuracies across seasons and between sensors require further scrutiny, especially in South African waterbodies. The cited literature on remote sensing of chlorophyll-a concentration in South Africa's waterbodies relied mostly on course resolution data. The availability of Sentinel-2 sensors with coastal and red-edge bands presents an opportunity to improve the estimation of Chl-*a* concentration, especially in winter seasons.

2.3.4.2 Water turbidity and total suspended sediments

The turbidity of a water body is an optical indicator of the amount of downwelling light reduced by the presence of suspended matter in water. Water turbidity is strongly related to total suspended matter in that a higher concentration of suspended particles reduces the high amount of light passing through water, leading to high water turbidity (Tyler et al., 2022; Gholizadeh et al., 2016). Therefore, total suspended matter and turbidity are important variables in the remote sensing of water quality due to their interaction with incoming sunlight. These define the scattering and absorption properties of water and are used to estimate suspended sediments based on water reflectance (Gholizadeh et al., 2016). Scattering emanates largely from suspended sediments, while absorption is dictated by Chl-*a* or particulate matter (Gholizadeh et al., 2016).

It is worth mentioning, however, that the spectral behaviour of suspended sediments hinges on the particle size distribution, with fine-grained material containing more particles and reflecting more light than coarse-grained material. Different sediment types also display unique reflectance patterns, with white clay sediment reflecting four times more light than red silt at 600 nm (Novo et al.,1989). Moreover, an increase in sediment concentration shifts peak reflectance to longer wavelengths. Meanwhile, organic matter such as Chl-*a* and CDOM lowers reflectance because of their absorptive effect. These intricacies around suspended sediments lead to varying results about estimating total suspended sediments using remotely sensed data. For instance, Myint and Walker (2002) found that the NOAA AVHRR bands, 580 to 680 nm and 725 to 1100 nm, could estimate suspended sediments of up to 200 mg/L with acceptable root mean square error (RMSE). The backscatter from inorganic suspended sediments and minimal interference from detritus, phytoplankton pigments, etc., was credited for the high performance of these bands.

The estimation of suspended sediments using the green band of SeaWiFS centred at 545-565 nm was inferior to the estimation achieved with the red band of SeaWiFS centred at 660-680 nm. The low estimation accuracy from the SeaWiFS green band was attributed to the shallow depth of water samples and increased effects from CDOM or pigments from phytoplankton species (Myint and Walker, 2010). These observations suggest that band wavelength position affects the accuracy of estimation achieved with a single-band model. Consistent with this, Nechad et al. (2010) observed a high correlation between reflectance at bands 665 nm and 708 nm of MERIS and SeaWiFS estimates of total suspended matter. The same study observed that the error of estimation increases with the use of spectral bands above 710 nm, and this contradicts Ritchie et al. (1976), who observed that the best relationship between spectral reflectance and suspended sediments can be achieved with bands centred at 700 to 800 nm.

While the use of single-band models may be attractive for their simplicity, their efficacy is limited by the concentration of total suspended sediments. Nechad et al. (2010) argued that single-band models are less appropriate for the estimation of suspended sediment in areas with low concentration (<1 g/m^3), especially in phytoplankton-dominated waters. Meanwhile, suspended sediment in the surface water exhibits seasonal variations owing to climatic factors and reservoir management (Ritchie et al.,1976). The varying concentrations of suspended sediments across seasons necessitate algorithms that are sensitive to these variations. In developing such an algorithm, an exploratory analysis of the utility of spectral reflectance from 600 to 800 nm may be useful in establishing an algorithm for estimating suspended sediments across seasons.

Consistent with the suggestion above, the use of band ratios that exploit spectral content from different wavelengths has been shown to be useful for estimating suspended sediments and turbidity. Bhatti et al. (2008) tested a two-band ratio involving green and NIR bands for estimating total suspended sediments and achieved an R^2 of 0.73 for the Altamaha and St. Marys Rivers, respectively. Meanwhile, the Sentinel-2 sensor pioneered a rich spectral configuration in the visible region of the electromagnetic spectrum. This offers an opportunity to further test the utility of the visible region for the estimation of suspended sediments across different seasons. The alternating four seasons in the COHWHS provide ideal conditions for exploring Sentinel-2 data.

2.3.4.3 Secchi disc depth

A Secchi disc is a white or black-and-white disc that is used to gauge the clarity of ocean and lake waters (Lee et al., 2015). The disc sets off slowly downward in the water column until the disc pattern becomes invisible to the observer, marking the Secchi disc depth (Lee et al., 2015; Gholizadeh et al., 2016). Secchi depth is affected by i) algae and algae-derived particles, ii) natural, coloured organic matter, and iii) soil-derived clay and silt particles (Brezonik et al., 2007; Olmanson et al., 2008). Nonetheless, the effect of these constituents on water clarity varies from one waterbody to the other. For instance, Brezonik et al. (2007) observed that soil-derived turbidity is not the most important factor in Minnesota lakes compared to algal abundance.

As a measure of water clarity, Secchi disc depth is a useful indicator of water quality that relates to i) human perception of quality, with clearer water being considered safe for swimming, and ii) algal abundance (Brezonik et al., 2007). It is strongly correlated to spectral reflectance in the blue, green, red, and NIR bands of Landsat sensor series. Lathrop (1992) observed that the ratio of TM band 2 or 3 over band 1 generates a spectral index useful for modelling Secchi disc depth and total suspended sediments in Green Bay-Lake Michigan. Fuller et al. (2004) observed that a combination of band 1, band 2 and band 3 of Landsat Enhanced Thematic Mapper plus (ETM+) relates better to Secchi disc depth across different ranges compared to the ratio of band 1 over band 3 plus band 3. Olmanson et al. (2008) implemented a least-squares multiple regression with TM band 1: TM band 3 and ETM+ band 1: TM band 3 ratio as independent variables and log-transformed Secchi depth data as dependent variables. This study observed a coefficient of determination ranging from R^2 of 0.77 to 0.80 across different periods, and the relationship between Secchi depth measurements and Landsat data was consistent across different ranges of Secchi depth.

The performance of Landsat band 1 and band 1: band 3 ratios as independent variables showed higher variability in the study of Fuller et al. (2004), where the R^2 ranged from 0.30 to 0.80, compared to the study by Olmanson et al. (2008) in Minnesota lakes. While these studies used different types of Landsat data, this should not be a problem, given that the two sensors share common spectral wavelength centres. In fact, ETM+ used in the Fuller et al. (2004) study had a superior radiometric resolution, yet its performance showed high variability compared to TM. This raises the question of whether differences in lake conditions between the two study areas could be at play, thus affecting the performance of band ratios. Lathrop (1992) observed that the band ratio of TM band 3 over band 1 was not appropriate across different ranges of Secchi depth and total suspended sediments because of a shift in spectral response at low TSS levels. Meanwhile, the band ratio of TM band 2 over band 1 was appropriate across different ranges of both variables.

Further scrutiny on the utility of remote sensing data for Secchi depth modelling may be necessary, especially with the availability of advanced satellite sensors featuring improved spectral and radiometric resolutions. For instance, the Sentinel-2 sensor series presents an opportunity to explore other wavelength centres not covered by the Landsat sensor series for Secchi depth estimation.

2.3.4.4 Water temperature

Water temperature is an indicator of water's kinetic energy. It is an important variable affecting physical and biochemical processes occurring in water. The ability of water to dissolve gases such as oxygen depends on water temperature, with lower temperatures leading to higher oxygen solubility (Christ and Wernli, 2014; Gholizadeh et al., 2016). Water temperature also influences the distribution, transportation, and interaction of some contaminants, such as nutrients (Gholizadeh et al., 2016). Importantly, water temperature varies with seasonal changes, depth, and time of day (Christ and Wernli, 2014), and this makes the monitoring of water temperature over space and time essential for understanding some physical and biochemical processes occurring in water.

The regular collection of data over Earth's surface and oceans by space-borne thermal infrared sensors, e.g., the Landsat Program, makes it attractive as a data source for studying the temperature of water bodies. Thermal infrared bands onboard Landsat sensor series are sensitive to radiant heat emitted from land surfaces and the radiant temperature of water bodies. The TM thermal bands have been used to estimate the thermal effect of effluent discharge on the west coast of Daya Bay (China) and show the spatial and temporal distribution of thermal pollution in coastal waters (Chen et al., 2003).

Meanwhile, Dyba et al. (2022) noted that the first thermal band of Landsat-8 centred at 10.9 μm is more essential than the second band at 12.0 μm in estimating water temperature. This again stressed the importance of spectral band location in executing a particular function. It is argued in the literature that the out- of-field stray light affects the calibration system of the thermal infrared sensor, and the captured thermal data can be erroneous. However, the error is minimal in the first thermal band of Landsat and should be preferred for analysis (Barsi et al., 2014). In addition, Gholizadeh et al. (2016) argue that modelling lake water temperature using satellite data benefits from understanding the seasonality of water temperature by facilitating the correction of incorrect measurements that the satellite sensor may make. For instance, winter months are historically associated with lower temperatures than summer months, and therefore, excluding weather anomalies may improve the model's performance.

However, the argument of Gholizadeh et al. (2016) is problematic, especially considering the changing climatic conditions and the low temporal resolution of the Landsat sensor. For instance, the recent Intergovernmental Panel on Climate Change report reported record high temperatures (Lee et al., 2024), and the trend in these temperatures may not be captured by Landsat-8 sensor owing to its low temporal resolution. Ultimately, following the Gholizadeh et al. (2016) logic may erroneously lead to the treatment of genuine changes in water temperatures as weather anomalies.

2.4 DISCUSSION

The review of available literature indicates that Landsat sensors, TM, MSS (Multispectral Scanner), ETM+, and OLI have received unmatched scrutiny for their utility in estimating different water quality parameters. While other satellite sensors, such as MODIS, MERIS, and SeaWiFS, have received a disproportionate amount of attention, data from Landsat sensors have been used in studying Chl-*a*, Secchi disc depth, TP, TSS, turbidity, water temperature, DO, biochemical oxygen demand, and chemical oxygen demand. This reflects the completeness of Landsat sensors with the availability of spectral bands in the visible, NIR, shortwave infrared and thermal infrared regions, making Landsat data applicable to different types of water quality parameters.

However, with its low temporal resolution and medium spatial resolution, the Landsat sensor series has important weaknesses. The 16-day revisit period of Landsat sensors makes it susceptible to data gaps, especially in areas frequented by unfavourable weather conditions. The summer period is particularly characterised by frequent cloud cover, leading to incidences of data unavailability. This makes the Landsat Program inadequate for regular monitoring of water quality. Moreover, the 30-metre spatial resolution of Landsat sensors limits its application to big water resources, leaving behind rivers. Despite these weaknesses, Landsat data remains useful in managing water quality parameters, and the availability of long historical archives makes it relevant even in the presence of new satellite sensors.

The question is whether the latest instalments of satellite sensors with higher spectral and spatial resolution improve the accuracy of estimating water quality parameters. There are indications that Sentinel-2 data delivers higher prediction accuracy than Landsat data regarding Chl-*a* and Secchi disc and performs the same for turbidity in the Tres Marias Reservoir, Brazil (Pizani et al., 2020). However, water quality issues vary with seasons, and the literature has shown that the performance of satellite data is affected by the variation in the concentration of water quality parameters across seasons (Nechad et al., 2010; Lathrop, 1992). It has been observed in South Africa’s Vaal Dam that Landsat-8 data mapped Chl-*a* concentration with contrasting accuracies between seasons (Malahlela et al., 2018).

The utility of Sentinel-2 sensor data needs to be explored for water quality monitoring under differing seasonal conditions. This is necessitated by the assertion that Sentinel-2 data may be useful for the actual management of water resources owing to its high temporal resolution. The COHWHS presents an ideal study site to explore Sentinel-2 data, because it is characterised by four seasons with varying climatic conditions and land use activities, all of which affect water quality. This study presupposes that the high spatial resolution and improved spectral configuration of Sentinel-2 make it sensitive to variations in water quality parameters.

2.5 SUMMARY AND RECOMMENDATIONS

The high-resolution, multispectral imagery of the Sentinel-2 MSI, its global coverage, and frequent revisits lead to more efficiency in spatial and temporal monitoring of water bodies. The ML algorithms can then capitalise on this data to create models that predict water quality parameters. When these techniques are combined with field data, the predictions are further refined, leading to an increase in accuracy through calibration and validation (Hassan and Woo, 2021). Therefore, all these methods (remote sensing data, field data, and ML), when combined, can assist in addressing the gap of determining the sensitivity of remotely sensed variables to optically and NOAPs (Objective 1) and determining (mapping) the spatiotemporal patterns of water quality parameters within the rivers in and around the COHWHS (Objective 2) over both space (geographical area) and time. Owing to the presence of a “red-edge” band in the NIR that is centred around the 705 nm wavelength, and high spatial, temporal, and spectral resolution (13 spectral bands), the Sentinel-2 MSI is well-equipped for addressing these literature gaps.

The availability of Sentinel-2 MSI data with high temporal resolution (for 5 to 10 days for both Sentinel – 2A and 2B) presents an opportunity to investigate the possibility of developing season-specific algorithms for estimating optically active and NOAPs within the rivers in and around the COHWHS. Studies (Arora et al., 2022; Pizani et al., 2020; Sent et al., 2021) have mainly looked at optically active water quality parameters and spatiotemporal patterns, and this means there is a need for more studies to also include NOAPs, as they can assist in showing the cause of optically active water quality parameter distribution (Gelsey et al., 2023). Some studies (Chavula et al., 2009; Lieu et al., 2022; Abbas et al., 2019; Reinhart and Piersoon, 2005) have mainly used sensors (MODIS, MERIS, SeaWiFS) designed for monitoring large water bodies, such as oceans, large lakes, and coastal marine waters. Thus, multispectral sensors with high spatial and temporal resolutions are more useful in monitoring small inland waters. With the advent and increased utility of ML models, such as random forest in remote sensing, water quality parameter retrieval would be faster and more accurate (Fu et al., 2023). A remote sensing approach is further proposed as a complementary technology for monitoring water quality in the COHWHS, as collecting water samples and laboratory testing (with ML algorithms) will be used as a method for validating accuracy. The proposed study investigates the utility of these EO technologies, such as the high-resolution Sentinel-2 MSI sensor, for detecting water quality parameters across three seasons (spring, summer, and autumn) in the rivers in and around the COHWHS.

Based on the literature review conducted here, the following recommendations can be deduced:

- The variations in Chl-*a* estimation accuracies across seasons and between sensors require further scrutiny, especially in South African waterbodies. The cited literature on remote sensing of chlorophyll-a concentration in South Africa's waterbodies relied mostly on coarse-resolution data. The availability of Sentinel-2 sensors with coastal and red-edge bands presents an opportunity to improve the estimation of Chl-*a* concentration, especially during winter seasons.
- Sentinel-2 sensors pioneered with rich spectral configuration in the visible region of the electromagnetic spectrum, offering an opportunity to further test the utility of the visible region for the estimation of suspended sediments across different seasons. This could overcome a limitation of band wavelength positional effects on the accuracy of estimating water quality parameters, especially with single-band models.
- Further scrutiny on the utility of remote sensing data for Secchi depth modelling may be necessary, especially with the availability of advanced satellite sensors featuring improved spectral and radiometric resolutions. For instance, the Sentinel-2 sensor series presents an opportunity to explore other wavelength centres not covered by the Landsat sensor series for Secchi depth estimation.
- The Landsat-8 sensor may not adequately capture the trends in these temperatures owing to its low temporal resolution.

CHAPTER 3: MATERIALS AND METHODS

3.1 INTRODUCTION

This chapter offers a detailed account of the chosen sampling locations and outlines the methodologies utilised for collecting and analysing both field and satellite data.

3.2 DESCRIPTION OF THE SAMPLE SITES

The COHWHS (See **Figure 3-1**) is situated approximately 39.2 km west of Johannesburg, within Mogale City Local Municipality (MCLM), and falling in the West Rand District of Gauteng, South Africa. In alignment with the ongoing effort to monitor water quality in the rivers encompassing the COHWHS and fulfil the objectives of the project, ten (10) surface water sampling locations suitable for remote sensing-based monitoring were identified along the Blougatspruit, Bloubankspruit, Tweelopiespruit, and the Crocodile River in tertiary catchment A21. Detailed descriptions, including geographic coordinates for each site, are provided in **Table 3-1**.

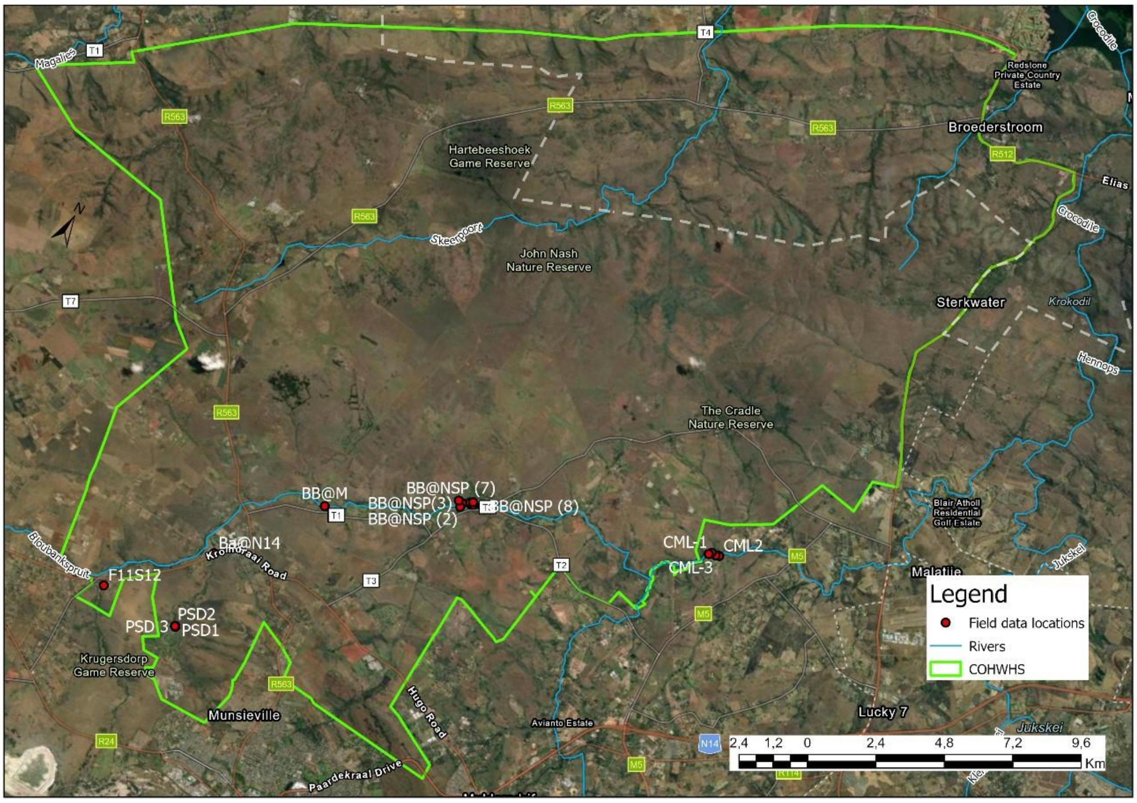


Figure 3-1. Spatial distribution of sample sites

3.2.1 Blougatspruit sampling sites

Sites PSD and BG@N14 experience perennial flow supported primarily by the discharge of treated and/or untreated municipal sewage effluent released from the Percy Stewart Wastewater Treatment Works (WWTW). Site PSD is ~1.3 km downstream of that WWTW and is situated adjacent to an instant lawn farm. Site BG@N14 is immediately downstream of the N14 bridge and is approximately 3 km downstream of site PSD (See **Figure 3-2**).


Figure 3-2. The two sites on the Blougatspruit are pictured with PSD on the left and BG@N14 on the right

3.2.2 Bloubankspruit sampling sites

Sites BB@M, BB@E and BB@NSP are located on the Bloubankspruit, after the confluence of the Rietspruit and Blougatspruit; therefore, the water quality at these sites reflects the combined influence of the mine water and the municipal sewage effluent discharge. The water quality at the BB@E and BB@NSP sites is further influenced by the Zwartkrans and Kromdraai springs, which discharge directly into the Bloubankspruit at their respective locations (**Figure 3-3**).

Figure 3-3. Two sites on the Bloubankspruit are pictured with BB@E on the left and BB@NSP on the right

3.2.3 Tweelopiespruit sampling sites

The F11S12 Dam is a private dam situated at the lower end of the Tweelopiespruit and immediately upstream of the N14 national road crossing. It experiences the perennial discharge of raw and/or treated mine water that has been blended with better-quality dolomitic groundwater. The latter is contributed by various karst springs located between the locus of mine water discharge and this site. Site TL@PF and PF Dam are located ~400 m downstream of site F11S12 Dam and within 500 m of the confluence of the Tweelopiespruit and its main stem, the Rietspruit (See **Figure 3-4**).

Figure 3-4. The two sites on the Tweelopiespruit are pictured with F11S12 Dam on the left and PF on the right

3.2.4 Crocodile River sampling sites

Sites K@GB and K@CML are located on the upper reaches of the Crocodile River, immediately downstream of its confluence with its tributary, the Bloubankspruit. The Crocodile River ranks among South Africa's most contaminated river networks. Unprocessed industrial, mining, agricultural, and domestic refuse has caused a decline in water quality along its length, resulting in notable algal blooms in various dams it passes through, such as the Cradle Moon Lake and Hartbeespoort Dam (See **Figure 3-5**).

Figure 3-5. The site K@CML on the Crocodile River

Table 3-1. Sites selected for surface water sampling

Latitude (dd.ddddd°S)	Longitude (dd.ddddd°E)	Site	Site Description
−26.06561944	27.72236944	PSD	Blougatspruit downstream of Percy Stewart WWTW
−26.06554722	27.69604444	F11S12 Dam	Private dam on Tweelopiespruit
−26.06208611	27.69538611	TL@PF	Tweelopiespruit on Pinocchio's Farm
−26.05963056	27.69914722	PF Dam	Dam on Tweelopiespruit at Pinocchio's Farm/Trading Post
−26.03969	27.72106	BG@N14	Blougatspruit near N14 bridge
−26.0087	27.7414	BB@M	Bloubankspruit near Makiti Lodge
−25.989780	27.77551	BB@E	Bloubankspruit at Ekuthuleni property
−25.98311389	27.78353056	BB@NSP	Bloubankspruit at NIROX Sculpture Park
−25.95970278	27.85448611	K@GB	Crocodile River at Glenburn Lodge & Spa Dam
−25.95821944	27.85658889	CML	Crocodile River at Cradle Moon Lake

3.3 DATA COLLECTION

3.3.1 Department of Water and Sanitation historical data

We obtained the water quality data from the DWS. The geographical distribution of the DWS stations is shown in **Figure 3-6**.

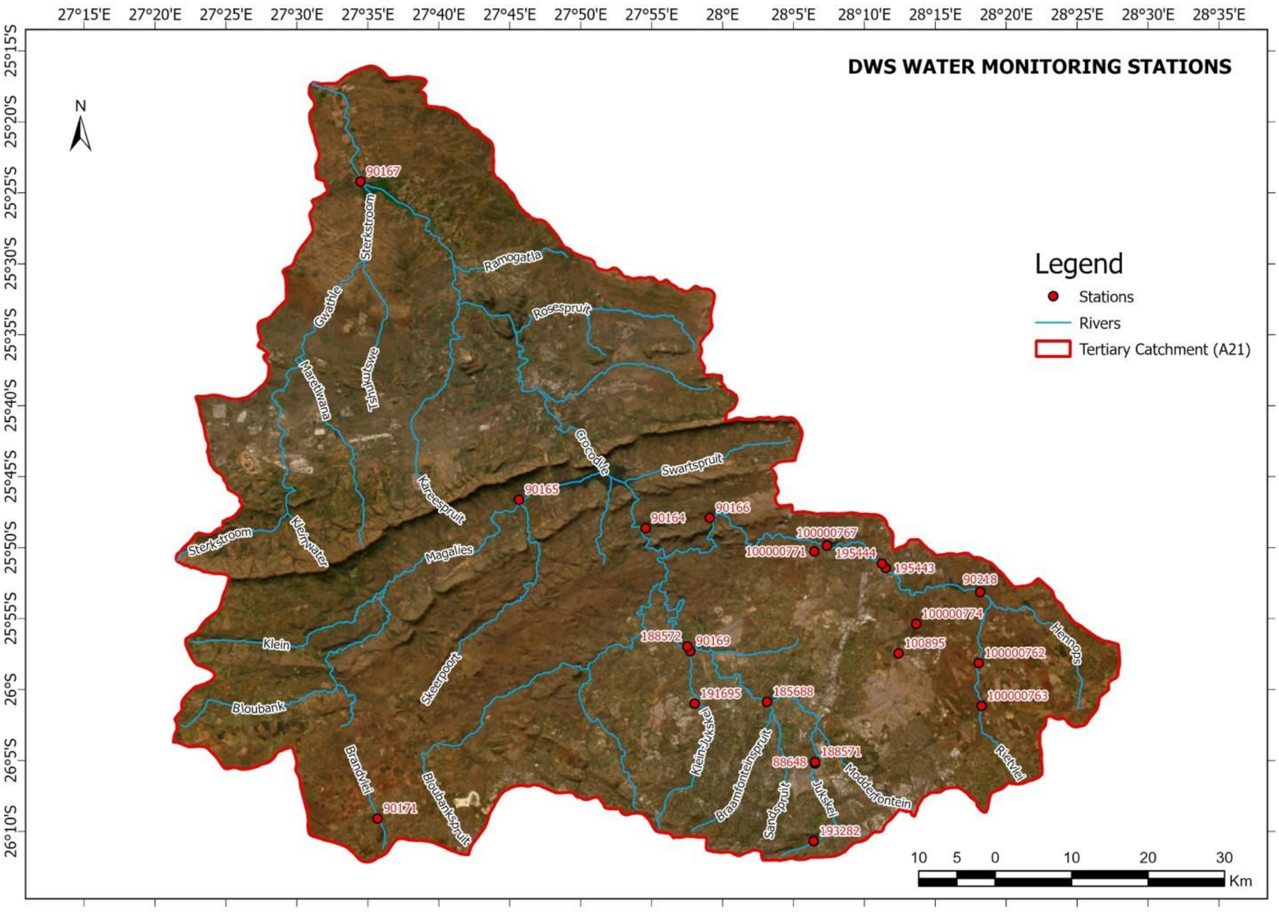


Figure 3-6. The location of the sampling sites and the studied rivers

The DWS monitoring activities in the COHWHS include the A2H049 site on the lower reaches of the Bloubankspruit (See **Table 3-2**) and Brickworks Dam (F11S12) on the Tweelopiespruit (See **Table 3-3**). The available surface water chemistry data at these stations was collated for statistical analysis and was utilised by this project for comparative purposes. **Table 3-2** shows the data associated with the period when uncontrolled raw mine water began discharging into the Tweelopiespruit (February 2010 to August 2012 and the subsequent phases where immediate AMD intervention measures were commissioned (August 2012 to February 2014). Since then, mine water treatment and management measures have been implemented to prevent the resumption of episodic uncontrolled raw mine water discharge.

It is worth noting that significant data gaps exist in the DWS dataset owing to the lack of data collection activities at the site for the period 2020 and 2024. The DWS attributed this hiatus to financial constraints and instrument failures.

Table 3-2. Synoptic overview of the Bloubankspruit water chemistry at station A2H049 in the periods February 2010 to May 2018 and July 2022 to November 2022 (Mvandaba and Shadung, 2023)

Variable	Period February 2010 to May 2018							Period July 2022 to November 2022							SANS (2015a)(1)
	n	5%ile	Mean	Median	95%ile	SD	CoV (%)	n	5%ile	Mean	Median	95%ile	SD	CoV (%)	
pH (–log10αH+)	202	7.5	8.2	8.2	8.5	0.3	3.7	5	7	7.2	7.2	7.3	0.1	1.7	5.0–9.7
SEC (mS/m)	199	61.5	93.4	90.1	126	23.2	24.8	5	81.5	81.5	124.1	137.7	26.3	32.3	<170
TDS (mg/L)	134	474.7	695.8	674.9	980.1	171.9	24.7	5	570.9	875.5	898.8	1153.8	250.6	28.6	<1200
Ca (mg/L)	191	54.3	96.3	90.7	153.7	33	34.3	5	47.9	147	152.2	281.5	105.2	71.6	n.s.
Mg (mg/L)	192	28.4	44.2	42.8	59.7	10.6	24	5	72.1	149.9	180.3	212	66.3	44.3	n.s.
Na (mg/L)	166	28.3	41.7	40.5	58.4	10.1	24.3	5	21.7	48.5	57.4	59.3	19.5	40.2	<200
K (mg/L)	169	2.9	4.1	4	5.7	1	24.4	5	5.5	5.8	5.8	6.3	0.3	5.8	n.s.
Cl (mg/L)	198	31.3	38.8	38.5	45.3	5.3	13.6	5	29.2	40.8	0.1	45.3	8.6	21	<300
SO4 (mg/L)	189	96.8	284	247.8	475	132.7	46.7	5	290.2	398.3	412.3	501.4	91.2	22.9	<500
HCO3 (mg/L)	189	242.4	422.2	404.4	621.6	121.6	28.8	5	0	134.7	142.9	159	32	23.8	n.s.
NO_3+NO_2 (mg N/L)	193	3.5	5.6	5.4	8.3	1.6	27.7	5	2.6	5.8	5.6	8.8	2.8	47.6	<11
PO_4 (mg P/L)	197	0.005	0.097	0.048	0.261	0.124	127.8	5	0.262	0.524	0.4	0.925	0.3	57.3	n.s.
Si (mg/L)	198	4.95	5.64	5.6	6.59	0.61	10.86	5	4.7	5.6	5.7	6.3	0.7	12.8	n.s.
Fe (mg/L)	61	0.004	0.019	0.012	0.072	0.024	126.3	4	0	0.019	0.021	0.025	0.007	36.1	<2
Mn (mg/L)	61	0.001	0.162	0.003	0.05	0.855	527.8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	<0.5
Al (mg/L)	60	0.003	0.019	0.009	0.057	0.026	136.8	4	0.007	0.03	0.016	0.071	0.034	115.5	<0.3

Table 3-3. Summary statistics of period-related surface water chemistry variability in the Tweelopiespruit (latest data as of September 2023) (Mvandaba and Shadung, 2023)

Variable	Statistical Parameter	F11S12 (Brickworks Dam)					
		A—B(1)	B—C(2)	C—D(3)	D—E(4)	E—F(5)	F— FIS
pH ($–log_{10}α_{H+}$)	n	173	128	83	57	276	151
	5%ile	3.9	2.7	5.3	3	5.4	4.5
	Mean	—	—	—	—	—	—
	Median	6.9	3	7	5	6.7	5.7
	95%ile	7.4	3.9	7.4	7.4	7.7	7.1
	SD	0.9	0.4	0.9	1.7	0.7	0.8
	CoV (%)	14	14	13	32	10	14
SEC (mS/m)	n	172	128	83	57	276	151
	Mean	268	332	281	329	291	298
	Median	283	330	276	323	293	299
	95%ile	329	378	350	391	323	325
	SD	48	29	34	34	41	20
	CoV (%)	18	9	12	10	14	7
SO_4 (mg/L)	n	171	128	83	56	276	151
	Mean	1636	2264	1879	2137	1832	1550
	Median	1760	2240	1870	2075	1840	1896
	95%ile	2015	2593	2148	2640	2160	1900
	SD	349	245	268	274	222	2145
	CoV (%)	21	11	14	13	12	188

(1) 09/2006-01/2010 (2) 02/2010-07/2012 (3) 08/2012-02/2014 (4) 03/2014-03/2015 (5) 04/2015-09/2020 (6) 10/2020-09/2023

3.3.2 Council for Scientific and Industrial Research historical data

For nearly a decade, the Council for Scientific and Industrial Research (CSIR) has spearheaded the collection, organisation, and analysis of both surface water and groundwater data as part of the monitoring programme for the COHWHS. This wealth of historical data serves as a cornerstone for evaluating current conditions and forecasting future trends pertaining to the property's hydrological system. In this ongoing project, data sourced from the Blougatspruit, Bloubankspruit, and Tweelopiespruit river systems is being utilised for comparative analysis. Recent microbiological assessments (See **Table 3-4**) of surface water samples obtained from BG@N14, BB@M, and PF reaffirm the presence of faecal contamination along the Blougatspruit, Bloubankspruit, and Tweelopiespruit, respectively. Furthermore, an additional sampling site, BG Spring, provides insights into the chemical composition of a spring located upstream of the Percy Stewart WWTW, directly discharging into the Blougatspruit. Notably, there is a discernible contrast between water quality upstream and downstream of the WWTW, with the most significant pollution levels observed at BG@N14 and BB@M (both located downstream of the WWTW), where the most probable number (MPN) count for *E. coli* surpasses 2 419.6 per 100 mL.

Table 3-4. Water chemistry and microbiological results of samples collected at monitoring sites along the Blougatspruit, Bloubankspruit, and Tweelopiespruit in March 2023, June 2023 and September 2023 (Mvandaba et al., 2023). Red coloured values indicate the parameters exceeded the SANS 241 standard.

Variable	BG Spring (Blougatspruit)	BG@N14 (Blougatspruit)			BB@M (Bloubankspruit)			PF (Tweelopiespruit)			SANS 241 (2015a)(1)
	Sep 23	Mar 23	Jun 23	Sep 23	Mar 23	Jun 23	Sep 23	Mar 23	Jun 23	Sep 23	
pH (–log10αH+)	7.8	7.8	7.3	7.5	7.7	7.1	7.4	7.9	6.2	5.5	5.0–9.7
SEC (mS/m)	68	72	80	110	190	180	185	100	300	900	<170
Ca (mg/L)	73	39	29	32	242	237	223	91	527	495	n.s.
Mg (mg/L)	45	9.6	11	11	56	59	64	50	83	80	n.s.
Na (mg/L)	3.7	53	67	105	80	87	97	35	126	124	<200
K (mg/L)	<0.6	11	15	14	9	8.8	8.7	2	11	11	n.s.
Cl (mg/L)	<2.0	54	65	98	49	57	64	38	55	52	<300
SO_4 (mg/L)	26	53	81	94	854	801	802	298	1735	1725	<500
HCO3 (mg/L)	336	213	188	323	115	118	169	146	5.1	2.9	n.s.
NO_3+NO_2 (mg N/L)	0.17	0.12	<0.05	<0.05	3.9	0.43	<0.05	7.4	5.3	6.4	<11
Si (mg/L)	7.5	5.7	6.5	7.6	5.6	5.7	6.2	5.8	3.5	3.1	n.s.
Fe (mg/L)	0.03	1.5	1	2	0.82	0.43	0.81	0.04	0.1	0.04	<2
Mn (mg/L)	<0.01	0.55	0.38	0.67	4.3	1.39	0.65	0.02	3.51	1.2	<0.5
Al (mg/L)	<0.02	0.12	0.21	0.47	0.26	0.16	0.21	0.03	0.08	<0.02	<0.3
Total coliform bacteria (MPN/100 ml)	1553.1	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	<10
E. coli (MPN/100 ml)	<1	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	>2419.6	108.1	106.7	201.4	n.d.

3.3.3 In situ data collection and laboratory analysis

3.3.3.1 Sampling design

The monitoring of surface water quality (encompassing chemical and microbiological aspects) within the COHWHS involved implementing a monitoring campaign designed to fulfil several key objectives. First, it entailed conducting surface water sampling and in situ data collection synchronously with the Sentinel-2 MSI satellite overpasses. Second, purposive sampling following standard sampling protocol was conducted across various upstream and downstream surface water bodies within the COHWHS, covering rivers and dams affected by pollution. Thirdly, the sampling campaign aimed to gather a minimum of 20 samples, ensuring comprehensive data collection across the designated sites.

For the inaugural wet season (i.e., high-flow) sampling campaign, 14 and 16 March 2024, were meticulously selected to align with the satellite overpass of the Sentinel-2 MSI. During this timeframe, 20 water samples were collected between 09:00 and 14:00 SAST to minimise any temporal disparity between satellite overpass and sample collection. The follow-up sampling campaign was conducted during the dry season (i.e., low-flow) in 28 and 31 August 2024 as one of the project's objectives is to define the sensitivity of the Sentinel-2 MSI to seasonal trends in surface water quality.

Gathering of surface water samples for standard laboratory chemical and microbiology analyses comprised the collection of "grab samples" at designated sample sites, following the standard protocol as described in the DWS Sampling Guide (DWS, 2000). The samples that were collected focused on optically active water quality parameters, including Chl-*a*, colour, turbidity, and temperature; and non-optical water quality parameters including, pH, DO, electrical conductivity, nitrogen (N), phosphorus (P), sulfate, and *E. coli*. All water samples are collected in pre-rinsed polyethylene (plastic) bottles that were firmly closed and labelled with the site name/site number, date, and time before being placed in an ice-filled cooler box for storage at 4°C while in transit. This use of ice and the cooler box preserves the integrity of the samples collected in the field before further processing. **Table 3-5** indicates the description of the sampling bottle type for chemical and microbiological assessment and analysis.

Table 3-5. Description of the sampling bottle type for chemical and microbiological analysis

Water quality parameter	Bottle type
Chemistry (organic and inorganic)	1 L clear polyethylene (plastic) bottle, pre-rinsed
Chlorophyll-α	1 L brown polyethylene (plastic) bottle, pre-rinsed
Microbiology	100 mL sterile polyethylene bottle (Colilert 18)

The handling and transportation of samples in water quality sampling are frequently overlooked, yet crucial elements. Once collected, water samples undergo immediate chemical alterations, necessitating preservation to maintain stability until analysis. However, it is essential to recognise that preservation techniques only delay, rather than halt, ongoing chemical and biological changes post-collection, underscoring the importance of prompt analysis. Therefore, to ensure adequate sample preservation, all the collected water samples were promptly delivered to the CSIR in Pretoria for same-day storage in a 4°C laboratory cold room. To ensure alignment with protocol requirements, the microbiological samples were stored at 4°C for up to a maximum of 72 hours before being analysed at the SANAS-accredited CSIR microbiology laboratory. The analysis method is described in Section 3.3.3.2.

To facilitate the preservation of the hydrochemical samples, the bulk transportation of the samples to the CSIR Stellenbosch Chemistry Laboratory for inorganic and organic analysis occurred within two working days of collecting the last set of samples. The water samples were packaged and sealed within polystyrene boxes for overnight delivery. The chemical analysis procedures for all the water samples followed the approved analytical methods described in “Standard Methods for the Analysis of Water and Wastewater” (APHA, 1992).

3.3.3.2 Laboratory analysis

The microbiological analysis method employed was the Colilert-18 test, which uses proprietary defined substrate technology to simultaneously detect total coliforms and *E. coli*. This method involves two nutrient-indicators, ONPG (o-nitrophenyl-β-D-galactopyranoside) and MUG (4-methylumbelliferyl-β-D-glucuronide), which serve as carbon sources and can be metabolised by the coliform enzyme β-galactosidase and the *E. coli* enzyme β-glucuronidase, respectively. As coliforms grow in the Colilert test, they metabolise ONPG, turning it yellow, while *E. coli* metabolise MUG, resulting in fluorescence. Non-coliforms, which generally lack these enzymes, are unable to grow and interfere. Any non-coliforms that do possess these enzymes are selectively suppressed by the test’s specifically formulated matrix. Results from the Colilert test are reported as MPN/100 mL, calculated based on distinctive colour changes observed in each Quanti-Ttray well after 18-24 hours of incubation, with a detection limit of 1 MPN per 100 mL. The samples were contained in 100 mL Colilert sample vessels for analysis.

The analysis of water chemistry encompasses the assessment of the concentrations of the selected variables, i.e., sulfate, phosphate, and nitrogen, using Inductively Coupled Plasma Optical Emission Spectrometry (ICP OES) detection. This detection technique is used for elemental analysis and allows for the simultaneous determination of multiple elements present in a sample. By exciting atoms in a sample to high-energy states using an inductively coupled plasma, the technique generates light emissions at characteristic wavelengths for each element. These emissions are then measured to determine the concentrations of the elements in the sample. ICP OES is widely used in various fields such as environmental monitoring, food and beverage analysis, pharmaceuticals, and metallurgy for its high sensitivity, precision, and ability to analyse a wide range of elements simultaneously. Instrumentation for analysis includes the MARS6 iWave microwave and Thermo ICap 6500. Simultaneously, ICP OES, ensures precise and simultaneous measurement of these parameters. The procedures for chemical analysis of water samples adhere to approved methods detailed in “Standard Methods for the Analysis of Water and Wastewater” (APHA, 1992).

3.3.3.3 In situ physicochemical parameters

For in situ determination of physicochemical variables (i.e., pH, electrical conductivity, DO, and temperature), a calibrated, handheld multiparameter meter (currently a Hach HQ40D) was employed. The Hach HQ40D is a multiparameter water quality meter designed for field and laboratory use. It can measure various parameters such as pH, conductivity, DO, turbidity, and specific ions in water samples. The device offers versatility, accuracy, and ease of use, making it a popular choice for environmental monitoring, water quality assessment, and research applications. The resultant field data was recorded on field data sheets (Appendix A) or stored on electronic devices (including the multiparameter meter) for efficient data retrieval when the data is analysed. The geographic coordinates of the sampling points were recorded to ensure that the collection of spectral information corresponded to the sampling points. **Table 3-6** shows the in situ and lab-based water quality parameters.

Table 3-6. Water quality parameters

		In situ	Lab-based
Optically active	Chlorophyll-*a* (Chl-*a*)		X
	Colour	X	
	pH	X	
	Electrical conductivity	X	
	Suspended Solids		X
	Temperature	X	
Non-optically active	Dissolved oxygen (DO)	X	
	Nitrogen (N)		X
	Phosphorus (P)		X
	Sulfate		X
	Escherichia coli		X

3.3.4 Satellite data

The satellite data was acquired from the Sentinel-2 Multispectral Instrument (MSI) sensor, which has a swath width of 290 km, and will be acquired from the Sentinel Scientific Data Hub (https://scihub.copernicus.eu/). The Sentinel-2 MSI sensor has a high spatial resolution and acquires images at three spatial resolutions. Its 20 m resolution will be resampled to a 10 m spatial resolution to be compatible with in situ data collected from the rivers along the COHWHS, and this will assist in simplifying the extraction of spectral data. The pixels corresponding to the sampling points will be used to obtain spectral data (reflectance and temperature). This data (when combined) was used to accurately determine the sensitivity of remotely sensed variables to optically and NOAPs using ML algorithms. The Sentinel-2 MSI sensor has 13 spectral bands (See **Table 3-7**), which have a wavelength range of 443 nm to 2 190 nm, but for this study, the bands used to determine their sensitivity to OAPs and NOAPs were those bands starting from the visible to the NIR region (490 nm to 842 nm for B2 to B8A) of the electromagnetic spectrum (Gholizadeh et al., 2016). These bands have been used in research to estimate or retrieve water quality parameters such as Chl-*a* and total suspended matter (among others), as reflected in B3 and B8A, respectively (Gholizadeh et al., 2016).

Table 3-7. Sentinel-2 Spectral bands and their spatial resolution

Spectral bands	Central wavelength (nm)	Spatial Resolution (m)
B1: Coastal aerosol	443	60
B2: Blue	490	10
B3: Green	560	10
B4: Red	665	10
B5: Red edge	705	20
B6: Red edge	740	20
B7: Red edge	783	20
B8: NIR1	842	10
B8A: NIR2	865	20

B9: Water vapour	945	60
B10: SWIR- Cirrus	1375	60
B11: SWIR1	1610	20
B12: SWIR2	2190	20

Level 1C images from the Sentinel-2 MSI sensor corresponding to the date of the fieldwork (i.e., 14 and 16 March 2024) were retrieved from the Copernicus Browser (See **Figure 3-8**). Atmospheric correction was completed on Sentinel-2 images to remove atmospheric effects, primarily caused by aerosols, water vapour, and gases. These can contribute up to 90% of the signal received by optical sensors (Ansper & Alikas, 2019), which obscures the true water-leaving reflectance needed for accurate water quality assessment. While Sentinel-2 Level-2A (L2A) imagery can potentially be corrected using the readily available Sentinel 2 Correction (Sen2Cor) algorithm, the latter was primarily designed for land applications (and relies on the "dark dense vegetation" method to estimate aerosol optical thickness). This assumption makes this algorithm 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). Consequently, Sen2Cor-corrected L2A data was not used in this study. Several other atmospheric correction algorithms were also considered but found unsuitable for the study context. For example, the Sea, Earth, and Atmosphere Data Analysis System, developed by the National Aeronautics and Space Administration, performs well in clear oceanic waters but often overcorrects in turbid inland waters and fails to handle adjacency effects adequately (Pahlevan et al., 2020a; Vanhellemont & Ruddick, 2021). Polynomial-based algorithms applied to MERIS (POLYMER) and Atmospheric Correction for OLI "lite" (ACOLITE) are widely used in optically complex waters; however, POLYMER can introduce noise in highly productive water, while ACOLITE tends to overestimate reflectance in the blue spectrum under less turbid conditions (Dörnhöfer et al., 2016; Pahlevan et al., 2020a).

Similarly, Ocean Color Simultaneous Marine and Aerosol Retrieval Tool (OC-SMART) and Case 2 Regional CoastColour (C2RCC) are effective in specific marine or coastal contexts but are limited by their assumptions and poor generalisability to rapidly changing inland water systems (Ansper & Alikas, 2019; Bui et al., 2022). The Modular Inversion and Processing System, although accurate, is computationally intensive and impractical for routine application (Kuhn et al., 2019).

Given these limitations, this project utilised the Image Correction for Atmospheric Effects (iCOR) algorithm (https://remotesensing.vito.be/services/icor, accessed on 16 May 2025), developed by the Vlaamse Instelling Voor Technologisch Onderzoek, implemented in the Sentinel Application Platform (SNAP) (De Keukelaere et al., 2018). The latter is a scene- and sensor-specific algorithm compatible with both Sentinel-2 MSI and Landsat-8 OLI data. It employs the MODTRAN5 radiative transfer model and derives atmospheric parameters directly from the imagery. Importantly, it includes the SIMEC module to correct adjacency effects, a key advantage in heterogeneous environments like inland waters (Wolters et al., 2021). Its surface-*a*daptive design, minimal user interaction, and reliable performance in both land and water targets made iCOR the most suitable option for this study. The atmospherically corrected images were resampled to 10m × 10m spatial resolution using the nearest resampling method, which retains the original values of the imagery.

Figure 3-7. Sentinel-2 Level-1C images were collected on the 14th and 16th of March 2024

3.4 DATA ANALYSIS

3.4.1 Summary statistics of non-optically active parameters

Table 3-8 presents the summary statistics for each water quality parameter collected during the high-flow (summer, 14 and 16 March 2024) and low-flow (winter, 28 and 31 August 2024) conditions. As shown in **Table 3-8**, the results showed variations between the high- and low-flow conditions considered in this study. The summary statistics indicated that the minimum and maximum values of DO concentration were relatively higher during the low-flow conditions (that is, by 0.33 mg/L and 13.94 mg/L, respectively), while the mean DO value was higher than the high-flow conditions by 4.19 mg/L. In contrast, the standard deviation indicated a wide variation of DO values per site and the mean of DO concentrations in the water during the two periods. For the high-flow conditions, the minimum and maximum values for DO concentrations were lower at 0.1 to 7.92 mg/L, respectively, while the mean concentration was 3.48 mg/L. The relatively high standard deviation indicated that the recorded concentrations of DO at each site have a wide range of high variations in DO values. The results for DO concentrations reveal that the water bodies were more contaminated during the high-flow season as compared to the low-flow season, as low DO in water means there is no oxygen for fish, invertebrates, and all organisms to survive, meaning that the water is polluted (Orr et al., 2015).

For pH, the summary statistics indicated that its concentration on the water bodies was relatively higher during the low-flow season, as the maximum value was higher by 2.56 and the mean value was higher by 0.81. The minimum value was, however, lower by 1.11 while the standard deviation was lower, indicating that most of the pH values per site were closer to the mean and that there was low variability. During the high-flow season, the pH values indicated were relatively lower, with the maximum and mean values being 8.01 and 7.31, respectively. The minimum values were, however, slightly higher at 5.55, while the standard deviation was lower, which also indicated low variability in the pH values per site. These results show that low-flowing water increases the level of pH as the presence of substances that contribute to alkaline water is enhanced because of the slow movement of the water (Saalidong et al., 2022).

For temperature, the summary statistics indicated that the minimum and maximum temperature values during the high-flow conditions were 6.7°C and 14.4°C, respectively. In contrast, the standard deviation values revealed low variability in the measurements of temperature values per site, as most temperature measurements were close to the mean value of the results. During the low-flow conditions, the temperature values were relatively low, with the minimum and maximum values being 14.40°C and 23.30°C, respectively, while the mean value was 18.03°C. The low standard deviation value during these low-flow conditions also indicates low variability in the results, indicating that there was low variability of temperature values per site when measurements were recorded. These results indicate that high-flow conditions lead to increased temperatures in water bodies, while low-flow conditions lead to decreased temperatures.

According to the mean values recorded during the low-flow conditions, EC concentration was higher during high-flow conditions, as the mean value was raised by 142.85 μs/cm. The minimum value was also higher by 42 μs/cm, while the maximum value was lower by 100 μs/cm. The standard deviation during these low-flow conditions was relatively higher than during high-flow conditions, indicating that there was high variability in the measurements of EC per site. In summary, the results reveal that EC was higher during the low-flow conditions of the water bodies and it was lower during the high-flow conditions of the water bodies, and the fluctuations in the presence of solid substances in the water may have caused this.

Table 3-8. Summary statistics of water quality parameters during high- (summer, wet) and low-flow (winter, dry) conditions

	Parameter	Min	Max.	Mean	Std. dev.
High flow	DO	0.10 mg/L	7.92 mg/L	3.48 mg/L	2.88
	pH	5.55	8.01	7.31	0.57
	Temp	20.70°C	30°C	25.71°C	2.76
	EC	226 μs/cm	2950 μs/cm	753.15 μs/cm	658.12
Low flow	DO	0.43 mg/L	21.86 mg/L	7.67 mg/L	6.2
	pH	4.44	10.57	8.12	1.26
	Temp	14.40°C	23.30°C	18.03°C	2.58
	EC	268 μs/cm	2850 μs/cm	896 μs/cm	660.32

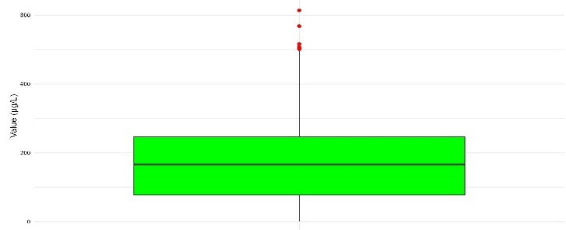
3.4.2 Lab data augmentation using multivariate empirical distribution modelling

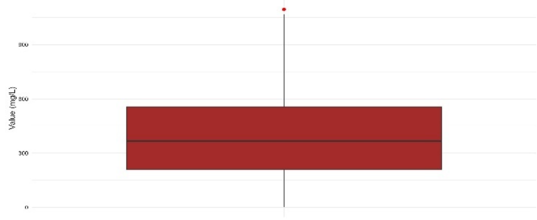
Owing to the lack of extensive data to develop robust models of optically active parameters, we used MEDM, which aims to simulate the statistical properties and inter-variable dependencies observed in laboratory and satellite measurements (Smania & Jonsson, 2021). This approach treats inputs as correlated 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 ML model training and validation (Teutonico et al., 2015). Approximately 1 000 points were simulated, resulting in some negative values for Chl-*a* and TSS that do not exist in reality, such as those <0. Therefore, these were removed from the data frame, and the remaining data points were used as inputs for ML. **Table 3-9** shows the summary statistics of data generated by MEDM in summer (i.e., wet season, high-flow conditions) and winter (i.e., dry season, low-flow conditions). Box plots were used to remove outliers from the dataset in conjunction with the interquartile range (IQR) method. The IQR was calculated as the difference between the third quartile (Q3) and the first quartile (Q1) (i.e., IQR = Q3 - Q1) (Čampulová et al., 2022). Based on the standard 1.5 × IQR rule, any data point falling below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR was classified as an outlier (Takiar, 2023). This non-parametric approach is widely used owing to its robustness and effectiveness in detecting extreme values without assuming a normal distribution. This enhances data quality and analysis accuracy (Čampulová et al., 2022).

Table 3-9: Summary statistics of data collected in summer and winter seasons generated by multivariate empirical distribution modelling (MEDM)

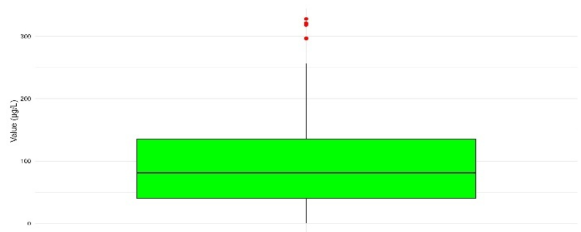
		n	Minimum	Maximum	Range	Median	Mean	Standard deviation	Coefficient of variation
Summer	Chl-*a*	542	0.672	613.751	613.079	166.921	179.797	126.217	70.20
	TSS	542	1.562	1375.985	1374.423	371.952	404.847	257.557	63.62
Winter	Chl-*a*	521	0.298	327.958	327.66	81.188	92.354	64.453	69.79
	TSS	521	0.071	605.070	604.999	158.123	177.073	119.467	67.47

Figure 3-8 (a) presents the structure of the simulated Chl-*a* values in summer after removing outliers. The simulated values towards the highest Chl-*a* values (i.e., 653 µg/L) were removed as outliers. It can be seen that the data still contains outliers after removing primary values deviating from the rest of the data, which are values between 500 µm/L and 610 µg/L. In contrast, **Figure 3-8 (b)** displays the structure of the simulated TSS values for summer. Unlike Chl-*a* values, the TSS values only included a single outlier after removing the initial high values (1 200 mg/L to 1 543 mg/L). **Figure 3-8 (c)** illustrates the structure of the simulated Chl-*a* values in winter after removing outliers. Similarly to summer values, the winter values also have outliers after removing the significant values. The structure of the simulated TSS values for winter is represented in **Figure 3-8 (d)**. Unlike summer values, the TSS values for winter are very low, ranging from 0 to 450 mg/L.a) b)





c) d)



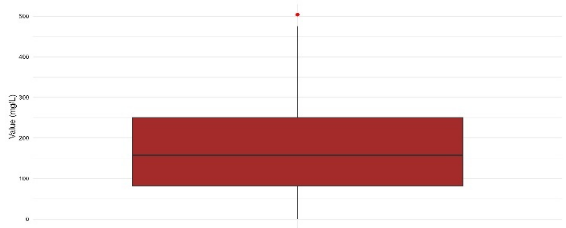


Figure 3-8: Structure of the simulated data after outliers' removal for Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)]

The final pre-processed data was divided into 80% training and 20% testing data for further analysis.

3.4.3 Pairwise bivariate correlation analysis

Bivariate correlation analysis was performed using Pearson's *r* correlation analysis (Equation 1) to determine the relationship between each water quality parameter and Sentinel-2 bands. The Pearson's *r* correlation coefficient measures the strength and direction of the linear relationship between two variables *x* and *y*.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad [1]$$

where x_i and y_i are the values of x and y for the ith observation. The Pearson correlation coefficient r was grouped into $0 \leq r \leq 0.2$ (Very weak), $0.2 < |r| \leq 0.4$ (Weak), $0.4 < |r| \leq 0.6$ (Moderate), $0.6 < |r| \leq 0.8$ (Strong), and $0.8 < |r| \leq 1.0$ (very strong) for interpretation purposes.

3.4.4 Machine learning algorithms for modelling optically active water quality parameters

To model Chl-*a* and TSS, five ML algorithms were implemented in R-statistical software, i.e., RFR, random forest survival, regression, and classification (RFSRC), GPR, and multi-output Gaussian process regression (MGPR). For consistency, each model was evaluated independently per season, using the same input variables (including the training and testing data).

3.4.4.1 Gaussian process regression and Multi-output Gaussian process regression

The GPR (Rasmussen, 2004) is a kernel-based probabilistic approach that establishes a relation between explanatory variables (e.g., spectral bands) and the output variable (e.g., Chl-*a*). To infer an unknown functional relationship from a training dataset, GPR elicits a prior GPR to constrain the possible form of the unknown function. Then, it updates the prior GPR in the light of training samples to generate the posterior GPR as the final functional model. As a non-parametric probabilistic model, GPR effectively 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 (García-Nieto et al., 2020). This approach not only enhances predictive robustness but also allows for explicit quantification of uncertainty in model predictions. GPR also supports multiple kernels, the common ones being anisotropic squared exponential and scaled Gaussian kernels. The GPR and its variants, such as variational heteroscedastic Gaussian processes, are increasingly gaining popularity owing to their competitive accuracy and capability to provide uncertainty estimates of the response variables (Camacho et al., 2021). This enables evaluations of the reliability of water quality parameter estimation for use in operational applications. It has been shown to produce high accuracy, be robust to overfitting, and have rapid training speeds.

Another variation, Multi-output GPR (MGPR) is a versatile and probabilistic regression technique that extends GPR to situations that require the simultaneous prediction of multiple correlated output variables (Zehra et al., 2018). The MGPR model has produced good results in other fields, but has not been used for water quality parameter estimation using remote sensing (Li et al., 2024; Zehra et al., 2018). Both GPR and MGPR were executed using the *gausspr* and *mgpr* R (Varvia et al., 2023) packages, respectively. Both models provided predictive outputs with associated 95% uncertainty intervals, enabling a robust quantification of model confidence and reliability in the estimation of water quality parameters. Kernel hyperparameters, signal variance, noise variance, and length scale were optimised via maximum marginal likelihood within a Bayesian framework for the MGPR model. For GPR, only the signal variance and noise variance were optimised as *rbfdot* was used.

3.4.4.2 Random forest algorithms

Random forest is a supervised ML algorithm comprising decision trees, where each tree is trained on a random subset of the input features. It is used for both classification and regression analysis. The model generates a diverse set of uncorrelated trees by introducing this randomness at each node. For regression tasks, the random forest 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). Owing to its consistent superiority against other ML algorithms, it is one of the most widely used ML algorithms (Karimi et al., 2023; Saberioon et al., 2023). Moreover, it offers the variance importance of the predictor variables, which helps understand the influence of each variable on the prediction outcomes. In this study, RFR was used since the input data is continuous.

3.4.4.3 Hyperparameter tuning of machine learning algorithms

The performance of ML algorithms is influenced by the selection of parameters in their construction. Therefore, hyperparameters play a vital role in their performance, and their tuning is crucial for optimal performance. For seasonal and back-predictions of a 5-year period, the optimal hyperparameters for the GPR and MGPR, i.e., the signal variance, length scale, and noise variance were tuned with five-fold cross-validation. In contrast, random forest and RFSRC, *ntree* and *mtry* are vital parameters. Therefore, the GridSearch was utilised to test the *ntree* values “50, 100, 150, 200, 250, 300, 350, 400, 450, and 500” and the *mtry* of “1, 2, 3, 4, 5, 6, 7, and 8” using a five-fold cross-validation. For GPR and MGPR, **Table 4-5** shows the hyperparameters used for the models.

Table 4-10: Hyperparameters of the models in summer and winter from tuning the models

	Summer		
Model	**Random forest Chl-*a***	**Random forest TSS**	**RFSRC**
ntree	500	500	300
mtry	8	8	15
	Winter		
ntree	500	500	250
mtry	1	8	5
	Summer		
Model	**GPR Chl-*a***	**GPR TSS**	**MGPR**
Noise variance	0.27	0.348	0.22
Signal variance	0.102	0.123	0.77
Length scale			9.47
	Winter		
Noise variance	0.725	0.362	0.44
Signal variance	0.108	0.105	0.34
Length scale			43.76

3.4.5 Performance measures

Evaluation of the regression model's performance is a critical aspect of quantitative data analysis in various scientific fields. Several statistical indices have commonly been employed to assess the accuracy and reliability of models. Among these, the coefficient of determination (*R^2*), RMSE, MAE and bias are commonly used in literature and were adopted for this study (Rodríguez-López et al., 2023). Serving as an indicator of model fit, R^2measures the proportion of variance in the dependent variable that is predictable from the independent variables (Niu et al., 2021). In contrast, MAE and RMSE are used to measure the dispersion of prediction errors (Tian et al., 2022). 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 measures the average difference between predicted and actual values, providing insight into the model's accuracy (Tian et al., 2022). It is particularly valuable for comparing the predictive errors of different models, as it provides a single summary measure of performance. The R^2 values range from 0 to 1 with no units, and the model performs well if it has a higher R^2 (i.e., closer to 1). With MAE and RMSE, the model with a lower value performs well, and they are expressed in the units of the estimated parameter. The model performs well in determining bias, with values close to 0 showing fewer under/overestimations. These metrics are calculated using the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad [2]$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \quad [3]$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \quad [4]$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i) \quad [5]$$

Where X_i is the actual value, Y_i is the predicted value, and n is the number of samples. Each model's predictive distribution was assessed by plotting actual versus predicted values with associated R^2, RMSE, MAE, and bias annotated on the scatter plots.

3.4.6 Transferability of multi-output machine learning models

The transferability of ML models is a critical factor in the operational scalability of remote sensing applications for inland water quality monitoring. The primary aim is to reduce the reliance on localised, resource-intensive field sampling by enabling models trained in one region to predict water quality parameters in other areas with minimal loss in accuracy. However, spatial transferability is often hindered by the variability in optical properties across water bodies, differences in atmospheric correction processes, and sensor noise (Rubin et al., 2021). Spectral heterogeneity of OAPs such as TSS and Chl-*a* can vary significantly owing to localised influences like geology, hydrodynamics, and land use, all of which can affect model generalisability (Gholizadeh et al., 2016). Therefore, evaluating the spatial transferability of ML models is essential to assess their robustness and potential for universal application in regional or national water monitoring programmes (Hoppe et al., 2024). The transferability of single-output algorithms using remote sensing has been tested in literature. However, the transferability of multi-output algorithms is limited. Therefore, only the multi-output algorithms were tested for spatial transferability in Hartbeespoort, Roodekopjes, and Buffelspoort dams. For accuracy in the transferability, the same process and variables used on the training images were used for the transfer images (i.e., AC, band combinations, resampling, and measurement units). The images were clipped to each dam's spatial extent and structured as data frames in R for model input. Predictions for TSS and Chl-*a* were generated and transformed back into raster format for summer and winter. These prediction maps were visualised using ArcGIS Pro to depict the spatial distribution of water quality parameters for both winter and summer.

To validate the maps of water quality parameters, raster values from the predicted TSS and Chl-*a* were extracted to points using the DWS samples. The metrics, R^2, RMSE, MAE and bias (See Equations 2 to 5), were used to quantify the map accuracy of the transferred models and error propagation.

CHAPTER 4: ESTIMATIONS OF WATER QUALITY PARAMETERS FROM SENTINEL-2 DATA

4.1 INTRODUCTION

This chapter presents the results of the analysis performed as part of the project and is divided into three sections. In Section 4.2, we examine the sensitivity of Sentinel-2 MSI bands to both optically and NOAPs. Section 4.3 presents a five-year spatial profile of optically active water quality parameters and Section 4.4 presents the transferability of Sentinel-2 derived models of dammed water bodies.

4.2 SENSITIVITY OF SENTINEL-2 TO SPATIAL AND TEMPORAL TRENDS OF WATER QUALITY PARAMETERS

This section aimed to determine the sensitivity of Sentinel-2 bands to water quality parameters found in the rivers and dams within the COHWHS, with specific focus on DO, pH, temperature, electrical conductivity (EC), total suspended solids (TSS) and Chl-*a*.

4.2.1 Bivariate analysis of non-optically active parameters

The correlation matrices in **Figures 4-1** and **4-2** display the relationships between various water quality parameters such as DO, pH, temperature, and EC with reflectance values from different Sentinel-2 bands during high- and low-flow conditions. During the high-flow (summer, wet) conditions, as illustrated in **Figure 4-1**, DO was found to be positively correlated with EC, but the correlation was not significant. However, DO had a significant negative correlation with EC and pH. During the low-flow (winter, dry) conditions, as illustrated in **Figure 4-2,** DO had a significant positive correlation with pH and temperature, while it was negatively correlated (not significant) with EC. During the high-flow conditions, pH had a significant negative correlation with DO and EC, while being non-significantly positively correlated with temperature. Under low-flow conditions, pH was significantly positively correlated with DO and EC, while it was non-significantly positively correlated with temperature. Under high-flow conditions, temperature, in contrast, was non-significantly positively correlated with pH, while it was non-significantly negatively correlated with EC. Lastly, temperature was found not to correlate with DO during these conditions.

Under low-flow conditions, temperature was found to be significantly positively correlated with DO, while it was non-significantly positively correlated with pH. Lastly, under low-flow conditions, temperature was found to be non-significantly negatively correlated with EC. Regarding EC, results under high-flow conditions revealed that EC was significantly negatively correlated with pH. Results also revealed that EC was non-significantly positively correlated with DO, while it was non-significantly negatively correlated with temperature. Under low-flow conditions, EC was significantly positively correlated with pH, while it was non-significantly negatively correlated with DO and temperature.

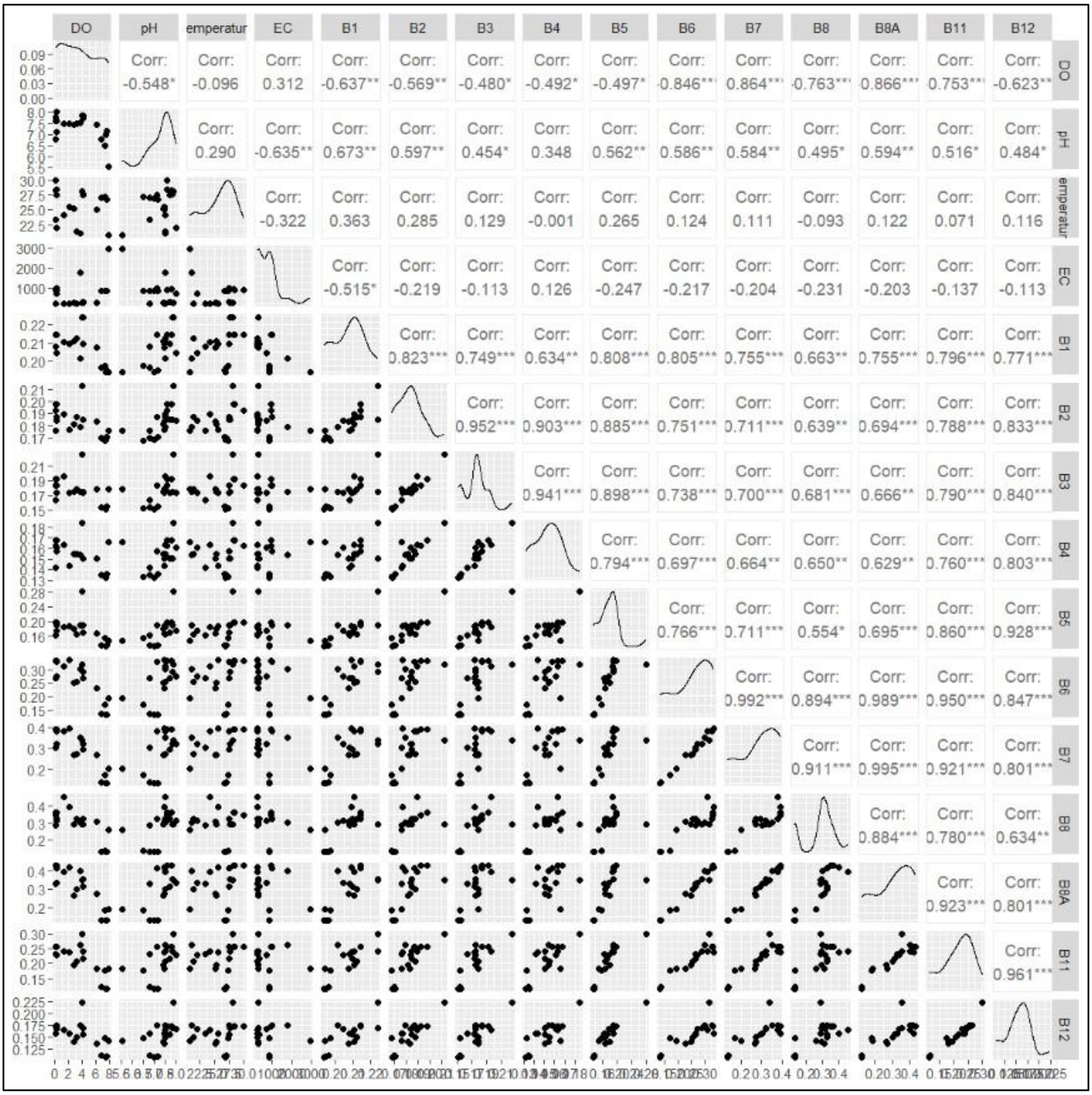


Figure 4-1. Bivariate analysis of NOAPs and Sentinel-2 bands during high-flow (summer, wet) conditions (March 2024). * , ** and * denote that the p-value is < 0.001, < 0.01, and < 0.05, respectively**

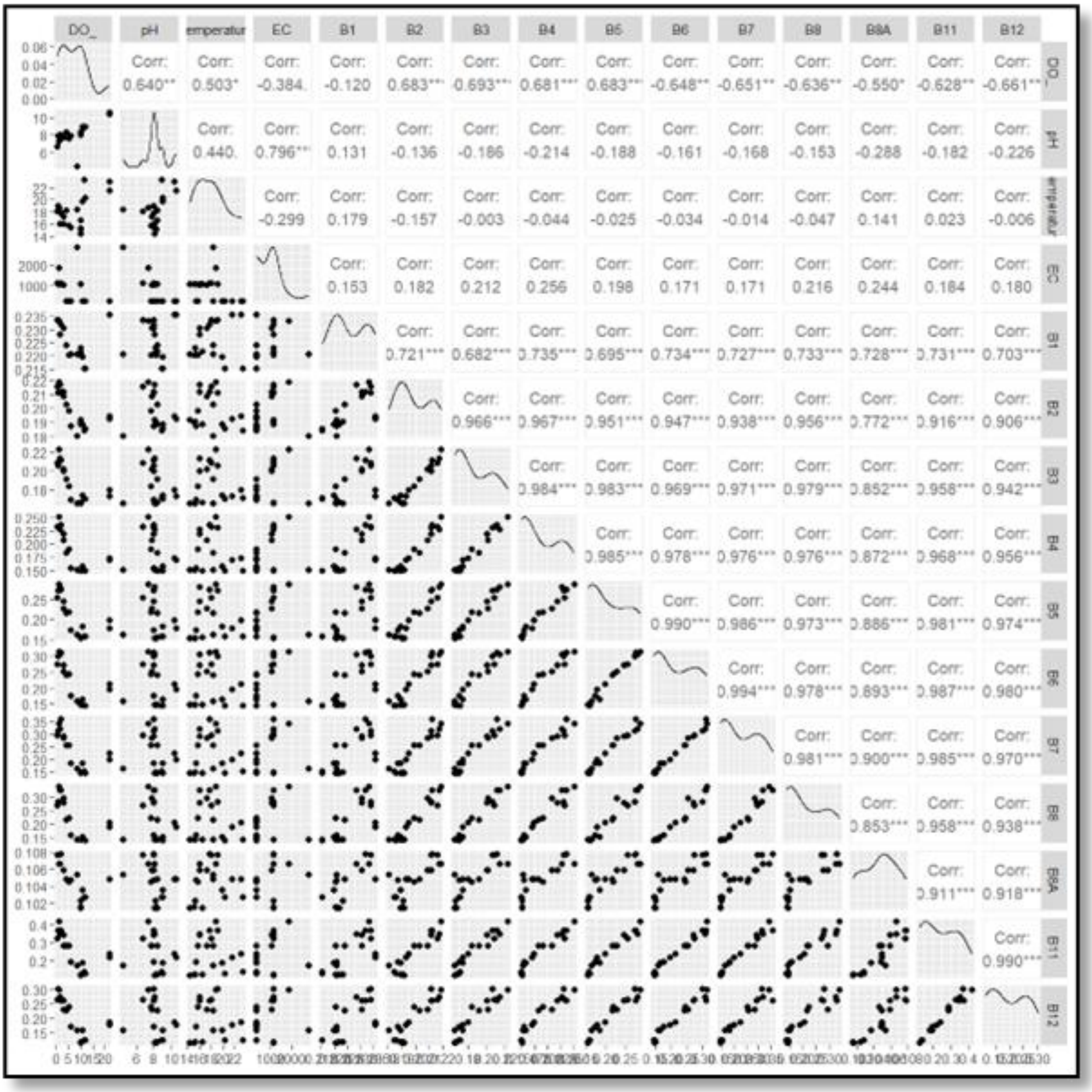


Figure 4-2. Bivariate analysis of NOAPs and Sentinel-2 bands during low-flow (winter, dry) conditions (August 2024). *, ** and * denote that the p-value is < 0.001, < 0.01, and < 0.05, respectively**

Correlation results between the Sentinel-2 MSI bands and the NOAPs (DO, pH, temperature, and EC) show variations in correlation results for high-flow and low-flow conditions, as illustrated in **Figures 4-1** and **4-2**. Under high-flow conditions, DO was found to be significantly correlated with B6, B7, B8, B8A and B11. In contrast, DO had significant negative correlations with B1, B2 and B12, while it also had non-significant negative correlations with B3, B4 and B5. Under low-flow conditions, DO had significant positive correlations with B2, B4 and B5, while it also had significant negative correlations with B3, B6, B7, B8, B8A, B11 and B12. The DO correlation results with B1 reveal non-significant negative correlations between the two variables during low-flow conditions. Correlation results for pH during high-flow conditions show that pH was significantly and positively correlated with all Sentinel-2 bands (B1, B2, B5, B6, B7, B8, B8A, B11 and B12) except for B3 and B4, which had a non-significant positive correlation. Under low-flow conditions, the correlation results for pH revealed that it had non-significant negative correlations with all the Sentinel-2 bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11 and 12) except for B1, for which there was a non-significant positive correlation between the two variables.

For temperature results during the high-flow conditions, it was found that this parameter had a non-significant positive correlation with B1, B2, B3, B5, B6, B7, B8A and B12. However, a non-significant positive correlation was found between temperature and B11, while no correlation was found between temperature, B4 and B8. Under low-flow conditions, temperature had a non-significant positive correlation with B1 and B8A, while there was a non-significant negative correlation with B2. The results for the rest of the bands revealed that temperature did not correlate with B3, B4, B5, B6, B7, B8, B11 and B12. The EC results showed a significant and negative correlation with B1, while B2, B3, B5, B6, B7, B8, B8A, B11 and B12 were non-significantly negatively correlated with EC. The results also showed that EC was non-significantly positively correlated with B4. Under low-flow conditions, EC had non-significant positive correlations with all the Sentinel-2 bands.

4.2.2 Sensitivity analysis of Sentinel-2 Bands to water quality parameters using various algorithms

The SLR, RFR, and Partial Dependence Plots (PDP) models were used to determine the sensitivity of Sentinel-2 bands to four NOAPs. The models were developed and performed in R-Studio software to identify which Sentinel-2 bands are most responsive to changes in NOAPs. Water quality parameters such as DO, EC, pH, and temperature served as the independent (predictor) variables, while the Sentinel-2 bands (B1-8, B8A, B11, and B12) served as the dependent (response) variables. The sensitivity analysis results varied according to the Sentinel-2 bands, the model and technique used to analyse the sensitivity, and the period (i.e., high-flow [summer/wet season] or low flow [winter/dry season]) being studied. Using the SLR model under the high-flow period (**Figure 4-3**), the sensitivity in B1 was mainly influenced by pH, followed by DO, and lastly EC.

During the high-flow period and using the SLR model, the results revealed that DO had the strongest influence on the sensitivity of most of the bands; however, sensitivity was not influenced by DO, as pH had a stronger influence on these bands, with the greatest R^- of 0.40 on B1. Regarding results developed by the SLR model, pH was the parameter with the second highest influence on the sensitivity of most of the 2 Sentinel-2 bands, while EC and temperature had the least influence. Both EC and temperature did, however, have a relatively strong influence on B1 and B8A, respectively, with R^- values reaching 0.27 for EC and 0.35 for temperature. For this model, the results showed that DO (as a non-optically active water quality parameter) had the strongest influence on the sensitivity of most Sentinel-2 bands during the high-flow period, followed by pH, temperature, and EC.

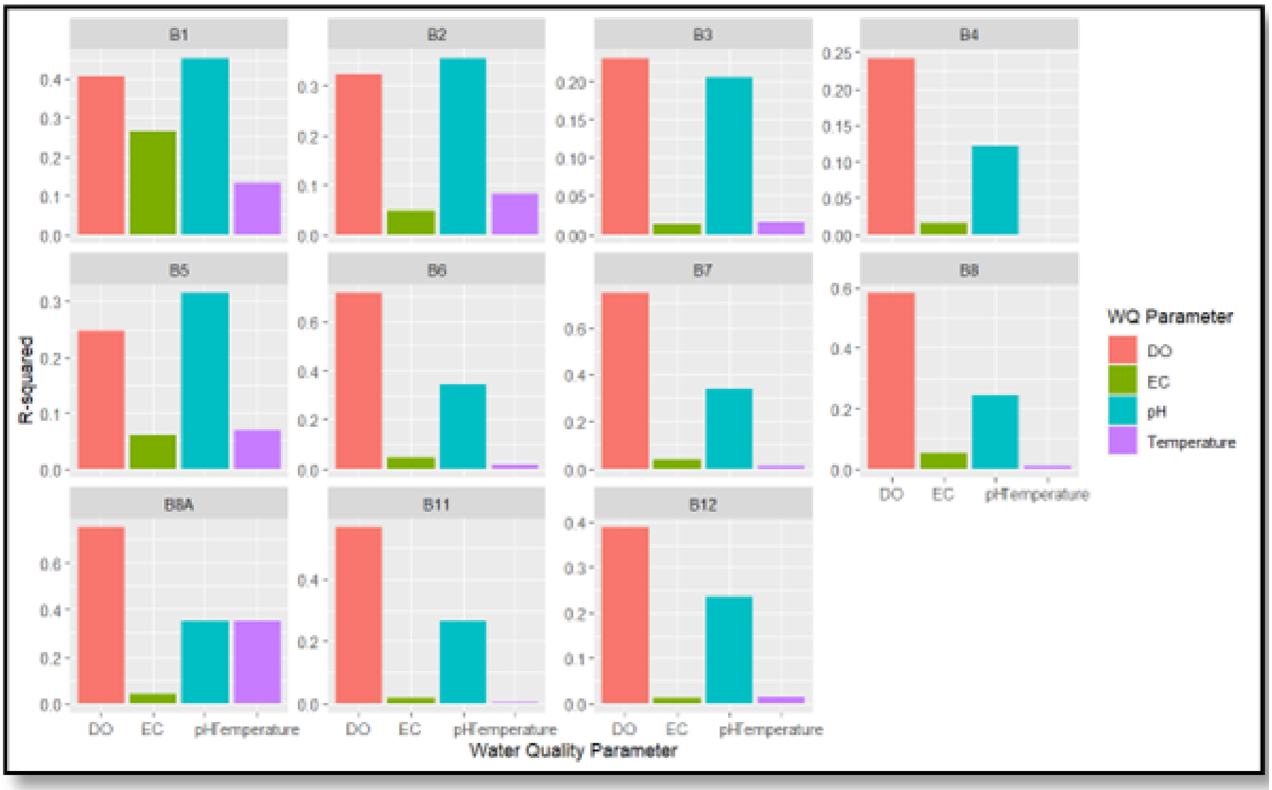


Figure 4-3. SLR model Sensitivity Analysis for the high-flow (summer, wet) period (March 2024)

For the RFR model (**Figure 4-4**), the high-flow period results were similar to the SLR model regarding the influence of DO on sensitivity. Of the parameters EC, Temperature and pH, DO had the strongest influence on the sensitivity of most of the B1, B3, B5, B6, B7, B8, B8A, B11 and B12 bands, with the highest R^2 of 28.14% being for B8A. However, for the B2 and B4 bands, temperature had the strongest influence on R^2 at 14.63% and 9.19% respectively. Temperature could also be considered the second-strongest parameter in terms of influence on sensitivity on these Sentinel-2 bands, because after DO, it was also the next strongest for B1, B3, B5 and B12. In some bands, pH showed relatively high significance on band sensitivity, as it was the second-strongest influence on B6, B7, B8, B8A and B11. However, EC had the least influence on the sensitivity of the bands, as it showed only slight significance for influence on B1, and B8 had the lowest influence (R^2 values) on most of the bands. For this model for the high-flow period, the results show that DO also has the strongest influence on the sensitivity of most of the Sentinel-2 bands during the high-flow period. When compared to the SLR model, the RFR model also concludes that temperature’s influence on the bands follows DO, pH, and EC.

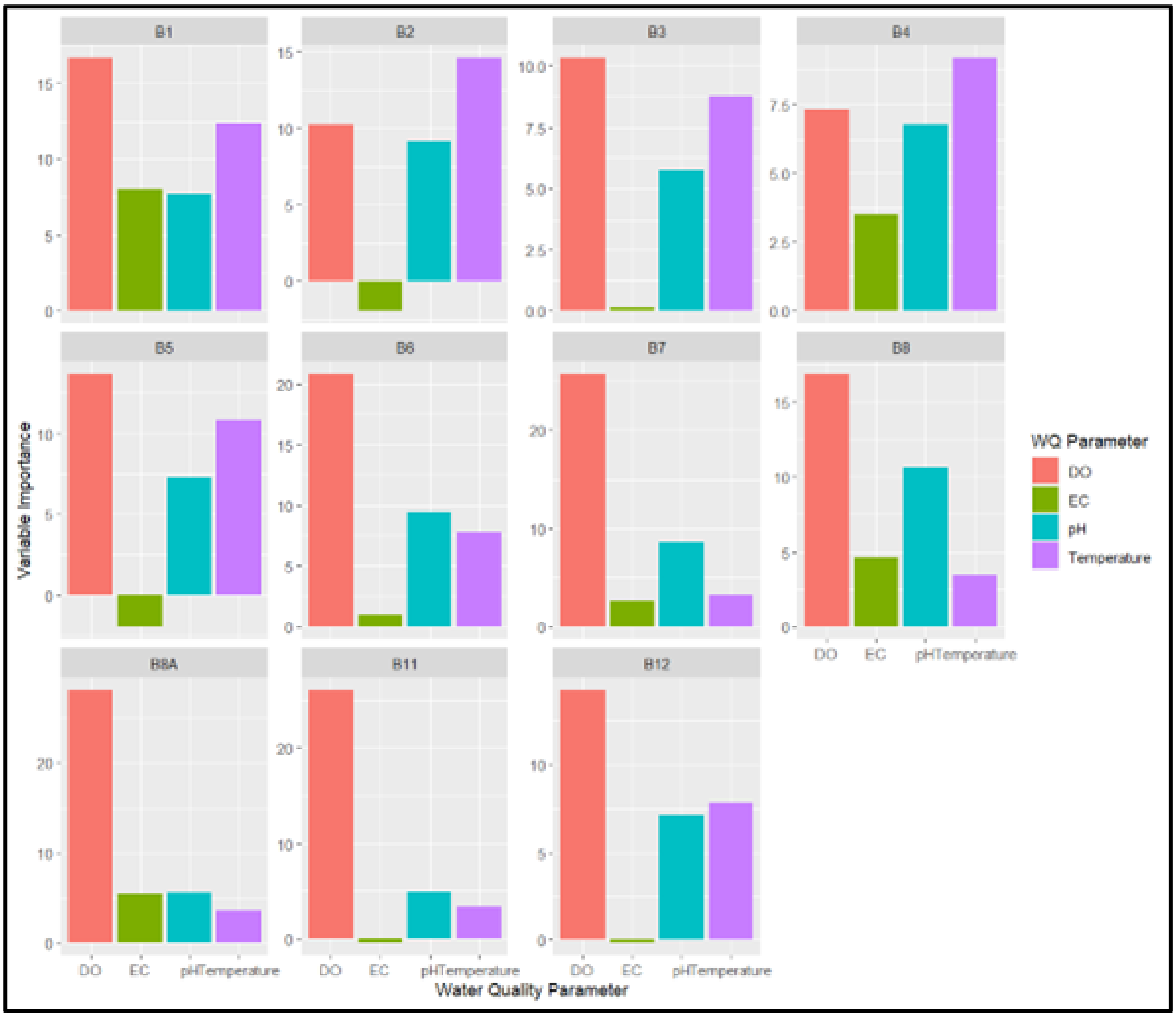


Figure 4-4. RFR model Sensitivity Analysis for the low-flow (winter, dry) period (August 2024)

4.3 FIVE-YEAR HISTORICAL PROFILE OF OPTICALLY ACTIVE WATER QUALITY PARAMETERS

In this section, we assess the performance of the Gaussian process regression (GPR) for estimating selected optically active water parameters (OAPs) and develop, through the back-prediction technique, a five-year water quality spatial profile in the COHWHS. The GPR algorithm was described in Section 3.4.5.

4.3.1 Bivariate analysis of optically active parameters

Pearson's correlation analysis *r* was performed between the water quality parameters and each of the Sentinel-2 bands (see Section 3.4.3 for description of the method). **Figure 4-5** shows results for bivariate analysis (i.e., pairwise correlations) between Chl-*a* and Sentinel-2 bands as scatterplots, kernel density estimate along the diagonal, and the *r* correlation coefficient with its significance indicated above the diagonal. As shown, B1, B5-B8, B8A, and B11 showed moderately to highly significant correlations with Chl-*a*. Among these, B1 is the only one that was positively related to Chl-*a*, while other bands showed a negative relationship, thus implying an inverse relationship. Other bands, such as B2, B3, B4 and B12, were weakly related to Chl-*a*, indicating that these bands may not directly capture its dynamics.

Figure 4-5. Bivariate analysis of Chlorophyll-*a* (Chl-*a*) and Sentinel-2 bands. *, ** and * denote that the p-value is < 0.001, < 0.01, and < 0.05, respectively**

Figure 4-6 illustrates the pairwise relationships between suspended solids and various Sentinel-2 spectral bands, with a focus on their correlations. As shown, the correlation coefficients range from strong to weak across the bands, indicating varying levels of association between suspended solids and the spectral reflectance. Notably, suspended solids exhibit a strong positive correlation with B2 (blue, $r = 0.60$), B3 (green, $r = 0.64$), B4 (red, $r = 0.63$) and B5 (red-edge, $r = 0.64$). Suspended particles in water scatter light and increase reflectance in shorter wavelengths, such as blue and green regions of the electromagnetic spectrum. The high correlations in the red and red-edge bands suggest their potential utility in suspended solids assessment, particularly given their sensitivity to chlorophyll and suspended matter. In contrast, the correlations between suspended solids and bands in the NIR and shortwave infrared (SWIR) regions, such as B8 (NIR, $r = 0.55$), B11 ($r = 0.55$) and B12 (SWIR2, $r = 0.59$), have moderate positive correlation, while other bands such as B7, B8A, and B1, also exhibit moderate correlation, but the correlation coefficients were <0.5. These lower correlations in the NIR and SWIR may be due to the reduced sensitivity of these wavelengths to suspended solids, as NIR and SWIR bands are more strongly influenced by water absorption compared to scattering by suspended particles.

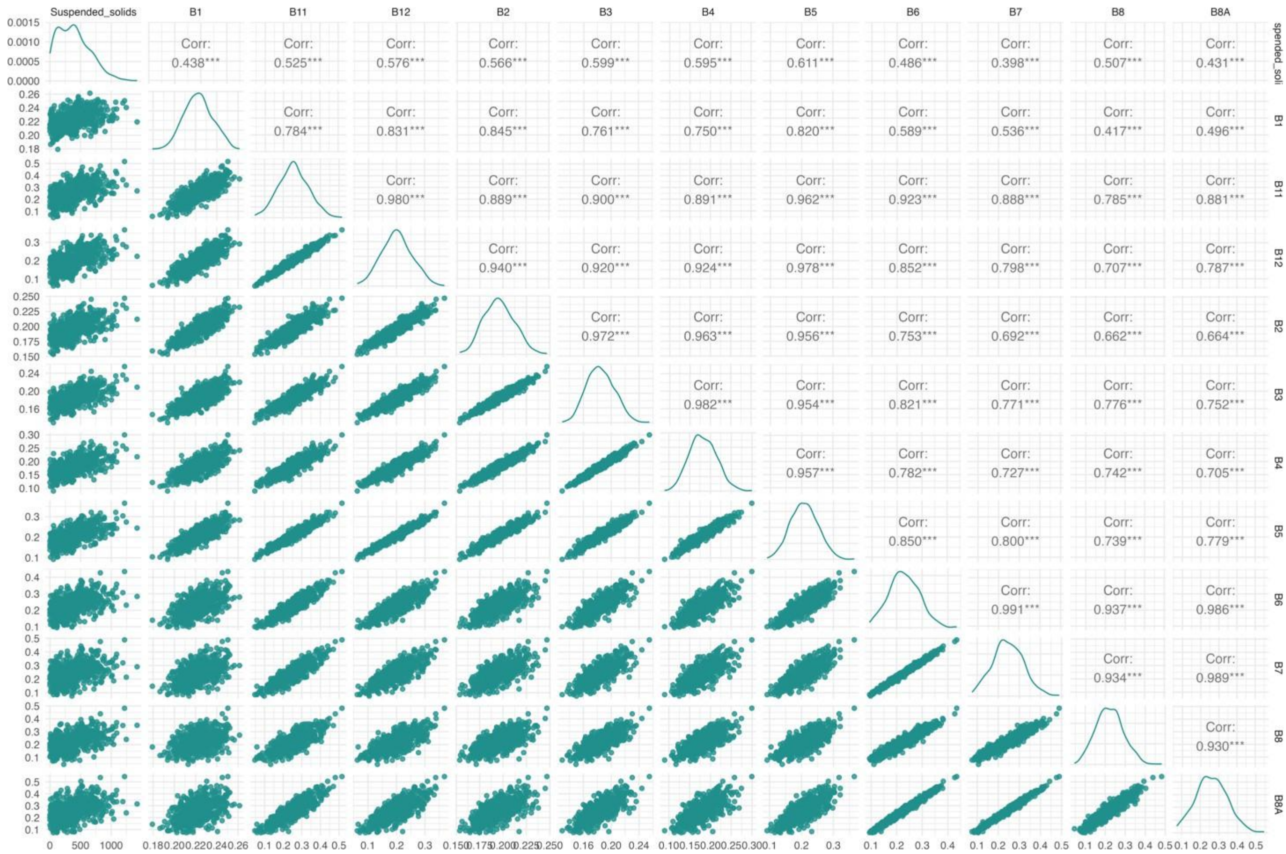


Figure 4-6. Pairwise between suspended solids and Sentinel-2 bands. *, ** and * denote that the p-value is < 0.001, < 0.01, and < 0.05, respectively**

4.3.2 Performance of optically active water quality parameters GPR models

Table 4-1 shows the performance of GPR models for estimating Chl-*a* and suspended solids. The result indicates that both water quality parameters could be estimated at a high accuracy using Sentinel-2 data. Chl-*a* had an accuracy (measured by coefficient of variation, R^2) of 75%, while suspended solids had an accuracy of 73%. The model errors as measured by RMSE and mean absolute error (MAE) between the two parameters were minimal, i.e., 57.67 µg/Land 45.65 µg/L, and 132.15 mg/Land 102.94 mg/L, for the two water quality parameters considered, respectively. A positive bias, indicative of minor overestimations was experienced by the Chl-*a* model, while a negative bias, indicative of minor underestimation, was witnessed for suspended solids.

Table 4-1. Performance measures for GPR models for chlorophyll-*a* (Chl-*a*) and suspended solids

Performance measures	Chl-*a*	Suspended solids
R^2	0.75	0.73
RMSE	57.67 µg/L	132.15 mg/L
MAE	45.65 µg/L	102.94 mg/L
Bias	5.49 µg/L	−13.96 mg/L

Figure 4-7 shows the scatterplots of the observed (truth) and estimated (response) for Chl-*a* and suspended solids. As shown, most validation samples for Chl-*a* were correctly predicted by the model, with only one sample being overestimated. Similarly, most suspended solids data samples were well estimated, with only extreme values, i.e., approximately 1 200 mg/L, being underestimated. Moreover, other points with values around 300 mg/L were overestimated by twice the magnitude, i.e., approximately 600 mg/L. Overall, the models indicate robustness to a wide range of values of the respective parameters and can be used for spatial distribution mapping of water quality parameters and back-predictions to develop a five-year spatial profile of these parameters.

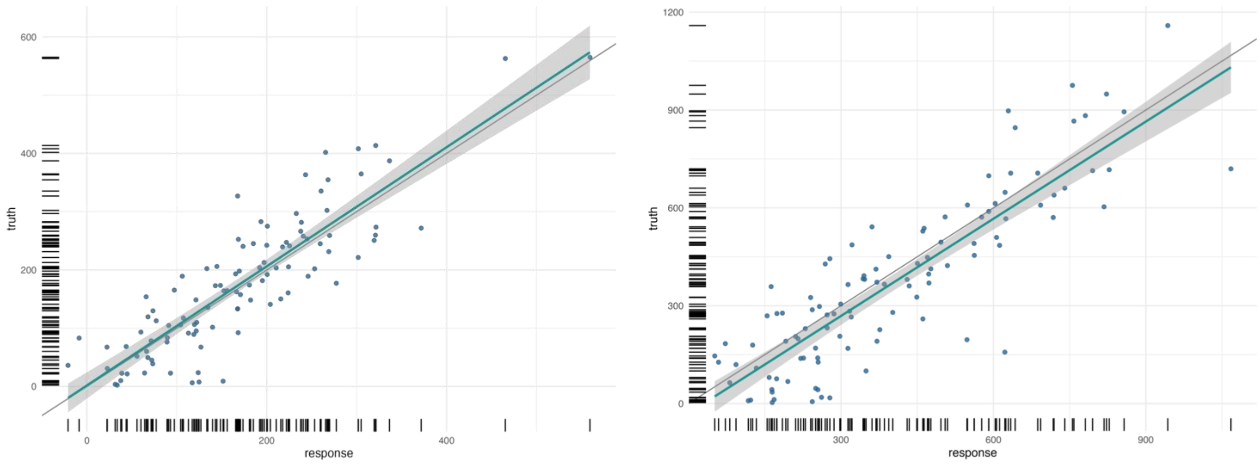


Figure 4-7. Gaussian process regression (GPR) model fit for chlorophyll-*a* (Chl-*a*) (left) and turbidity (right)

4.3.3 A five-year spatial profile of water quality

If the concentrations of the water quality parameters do not greatly exceed the measured levels, the GPR models should be able to make back-predictions (over the prior four years) based on the robust relationships they learnt between Sentinel-2 spectral reflectance and the water quality parameters. The GPR models have been widely applied in water quality assessment, primarily based on their ability to model non-linear relationships and quantify prediction uncertainty. For example, Shadrin et al. (2021) highlighted GPR's effectiveness in handling the spatial distribution of water quality parameters using variogram-based kernel functions. **Figure 4-8** shows the five-year spatial profile of Chl-*a* in Cradle Moon Lake, while **Figure 4-9** shows the five-year spatial profile of suspended solids in the same area. In both figures, the models demonstrate robustness in estimating the respective water quality parameters across time. This is depicted by, for example, when Chl-*a* levels change to very high values, i.e., >600 µg/L, in 2021 and 2023, the GPR model can respond to such changes. This is despite the model predicting a maximum of around 300 µg/L, indicating the impressive capability of the model to extrapolate (i.e., estimate values beyond those observed). In 2023, there were high levels of Chl-*a* throughout the lake, although such extreme values only occurred in the southern tip of the lake in 2021. As an OAP, Chl-*a* causes absorption in the red region, while causing a relatively higher reflectance in the NIR region of the electromagnetic spectrum. The red-edge region is sensitive to changes in Chl-*a* and chlorophyll-*b* in general, thus explaining the model's sensitivity to inter-annual changes in this parameter.

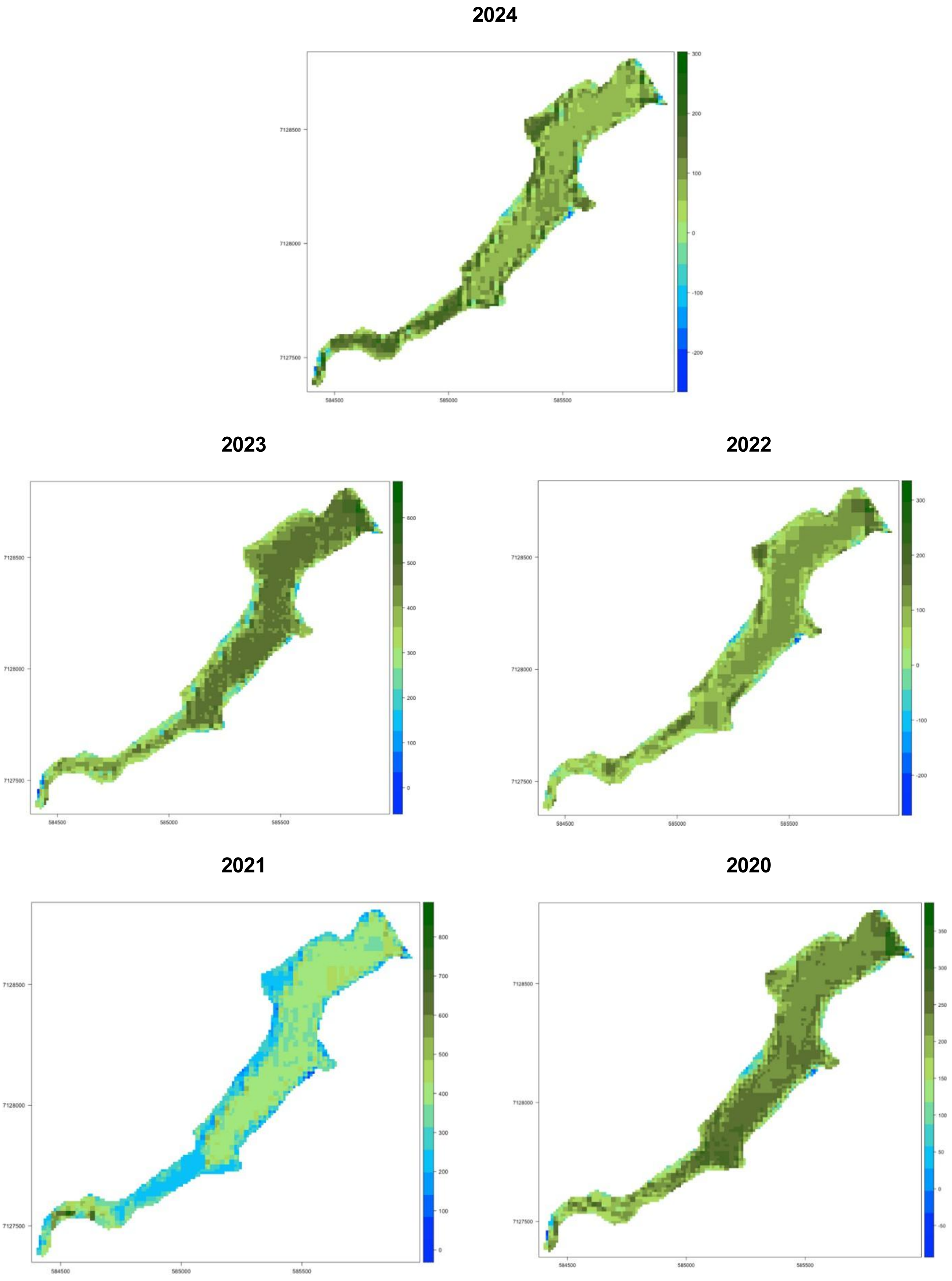


Figure 4-8. Spatial back-predictions of Chl-*a* (µg/L) in Cradle Moon Lake for a five-year period

In **Figure 4-9**, the results indicate that the suspended solids do not change as radically as the Chl-*a*. For example, the maximum values of this parameter remain more or less the same, with the areas towards the edges of the lake experiencing higher levels of suspended solids. There is a great variability in the distribution of suspended solids, with values generally ranging from 0 to 200 mg/L across most of the lake, although around the edges values are greater than 400 mg/L.

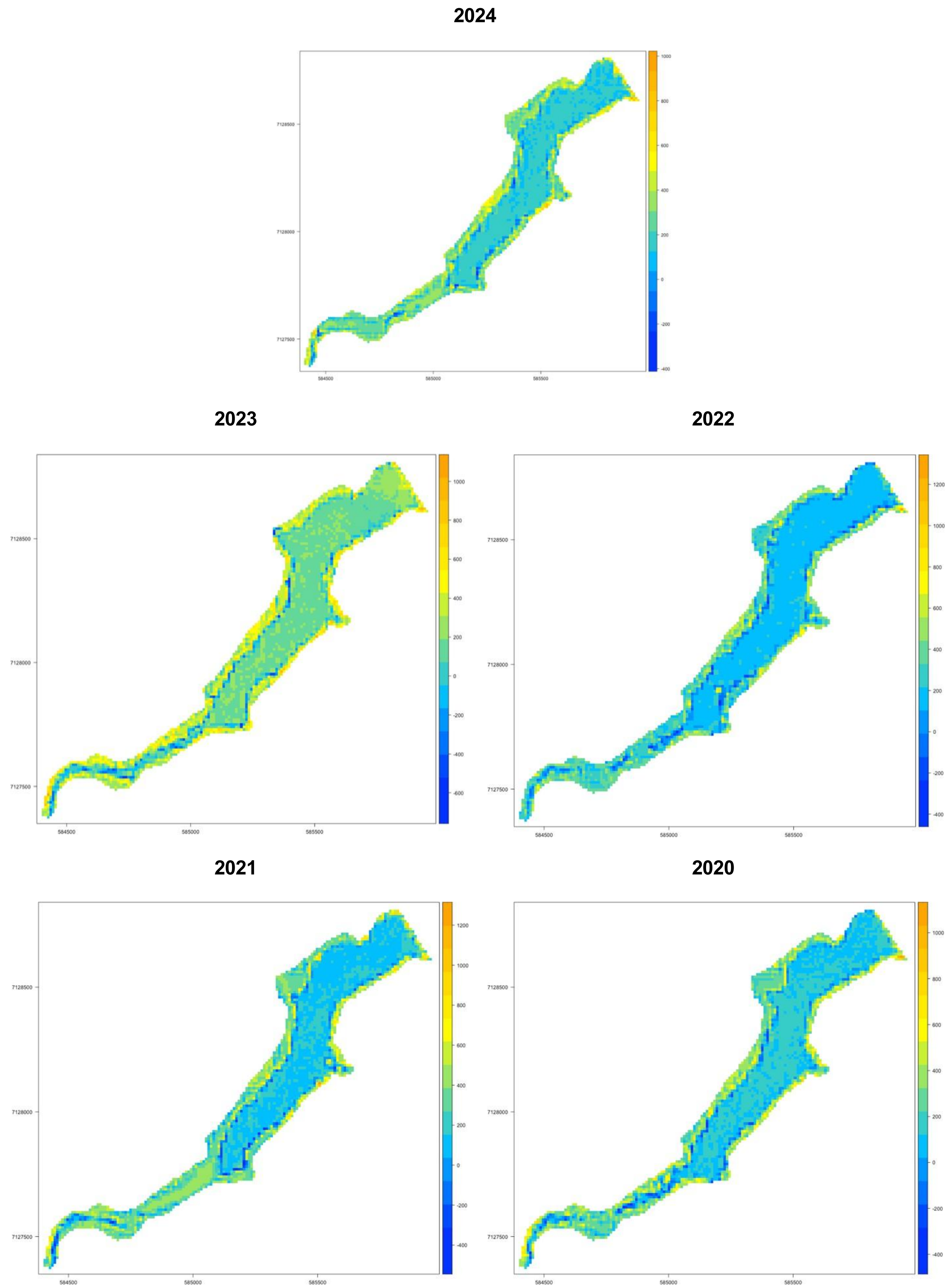


Figure 4-9. Spatial back-predictions of suspended solids (mg/L) in Cradle Moon Lake for a five-year period

4.4 TRANSFERABILITY OF SENTINEL-2 DERIVED MODELS TO DAMMED WATERBODIES ACROSS DIFFERENT SEASONS

The transferability of models across locations and times is critical to the operational application of remote sensing methods to water monitoring, as the goal is to increase spatial and temporal coverage while avoiding relying solely on field data collection campaigns for water quality data (Rubin et al., 2021). This section presents the transferability of multi-output algorithms to dammed water bodies, i.e., Hartbeespoort Dam, Buffelspoort Dam, and Roodekoppies Dam. We (i) evaluated and compared various single-output and multi-output algorithms in estimating key optically active water quality physicochemical parameters (i.e., TSS and Chl-*a*) under various seasonal conditions, and (ii) tested the transferability of multi-output algorithms in dammed water bodies, using data from DWS for validation.

4.4.1 Optically active water quality indices and algorithms testing

Over the years, researchers have explored various spectral indices, with differing levels of effectiveness depending on environmental conditions, sensor configurations, and the specific parameter being measured. Early studies (Dogliotti et al., 2015; Shi et al., 2015; Vanhellemont & Ruddick, 2014) primarily focused on single-band methods, establishing empirical relationships between spectral reflectance and water quality parameters. Although the single bands have been effective, they have been found to perform poorly in complex water bodies where several parameters exist (Tsapanou et al., 2020). Multiband approaches have been introduced to address this limitation, offering a more comprehensive analysis. Several band combinations, such as the ratio index (RI) (Neil et al., 2019), normalised difference index (NDI) (Mishra & Mishra, 2012), two-band algorithm (2-BDA) (Gitelson et al., 2008), and three-band algorithm (3-BDA) (Brivio et al., 2001), have been developed (Equations 6 to 8). These approaches select spectral bands that are physically related to the target parameter, combined with linear regression to estimate water quality parameters (Maier et al., 2021).

$$RI = \frac{,}{7} \quad [6]$$

$$NDI = \frac{,\$7}{,87} \quad [7]$$

$$3 - BDA = \frac{,\$7}{9} \quad [8]$$

Where *i*, *j*, and *k* denote the different Sentinel-2 band positions.

All Sentinel-2 band combinations were iterated at these different positions. The approach followed was proposed by Van Nguyen et al. (2019), where different band combinations were tried following the structure of existing indices. To reduce the collinearity of spectral indices created by the various band combinations, correlation analysis was performed, and we excluded highly correlated indices (i.e., >0.9). The remaining indices were used to estimate their linear, polynomial, and logarithmic relationships with Chl-*a* and TSS under various seasonal conditions, i.e., wet (summer) and dry (winter). These analyses allowed for the selection of highly correlated indices and for in situ measurements per parameter, as candidate predictors in ML models.

4.4.2 Spectral analysis

Sentinel-2 spectral bands were used in band combinations to test their correlation with Chl-*a* and TSS in summer and winter. The format for the band combinations followed the principles of RI, NDI, and 3-BDA. In order to reduce the computational time and redundancy, a heat map was created to check band combinations that were correlated to each other, and they were removed from the list. A total of 110 band combinations resulted from RI. However, after removing band combinations highly correlated to each other, there were 13 band combinations left represented by the heat map in **Figure 4-10 (a)**. The NDI and 3-BDA were left with 14 and 11 band combinations after removing correlated bands from 56 and 176 band combinations, respectively. **Figure 4-10 (b)** shows the final band combinations for NDI, while **Figure 4-10 (c)** illustrates band combinations for 3-BDA. From **Figure 4-10**, it can be concluded that all band combinations with B1 were highly correlated, while most of the band combinations with B2, B3, B4, and B5 were correlated for all spectral indices. In contrast, band combinations for B7 and B12 were not highly correlated. As with summer, RI and NDI for winter also had a total of 110 and 56 band combinations, respectively, with 13 and 14 band combinations after removing highly correlated band combinations. In contrast, the 3-BDA also had 176 band combinations, but had just 8 band combinations after removing highly correlated bands. **Figure 4-11** represents the RI, NDI, and 3-BDA final band combinations, respectively. The same pattern as in summer was realised in winter for the band combinations, including B1 being highly related to each other, and the same with band combinations, while B2, B3, B4, and B5 were less correlated.

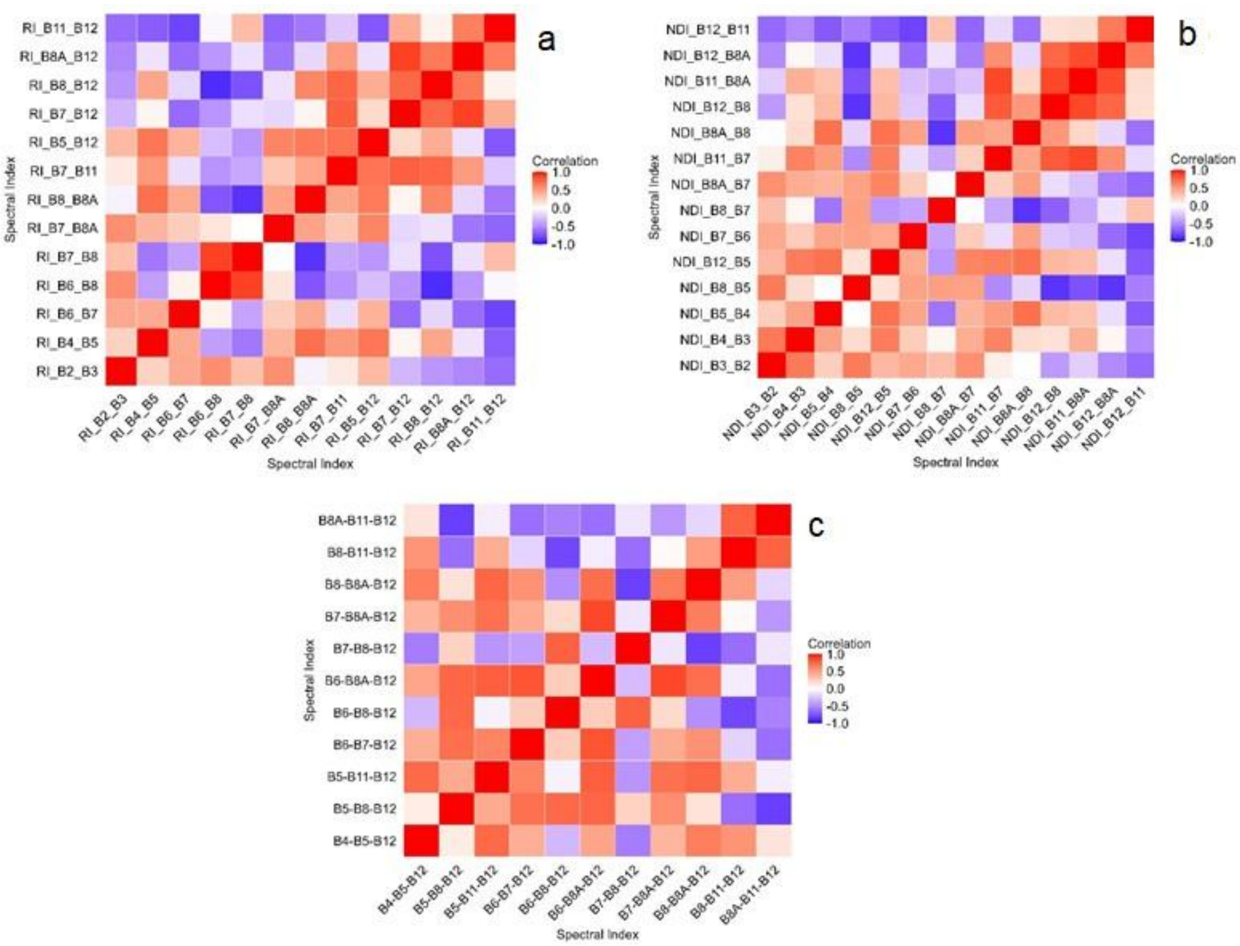


Figure 4-10. Spectral indices with low correlation in summer for (a) ratio index (RI), (b) normalised difference index (NDI), and (c) 3-band algorithm

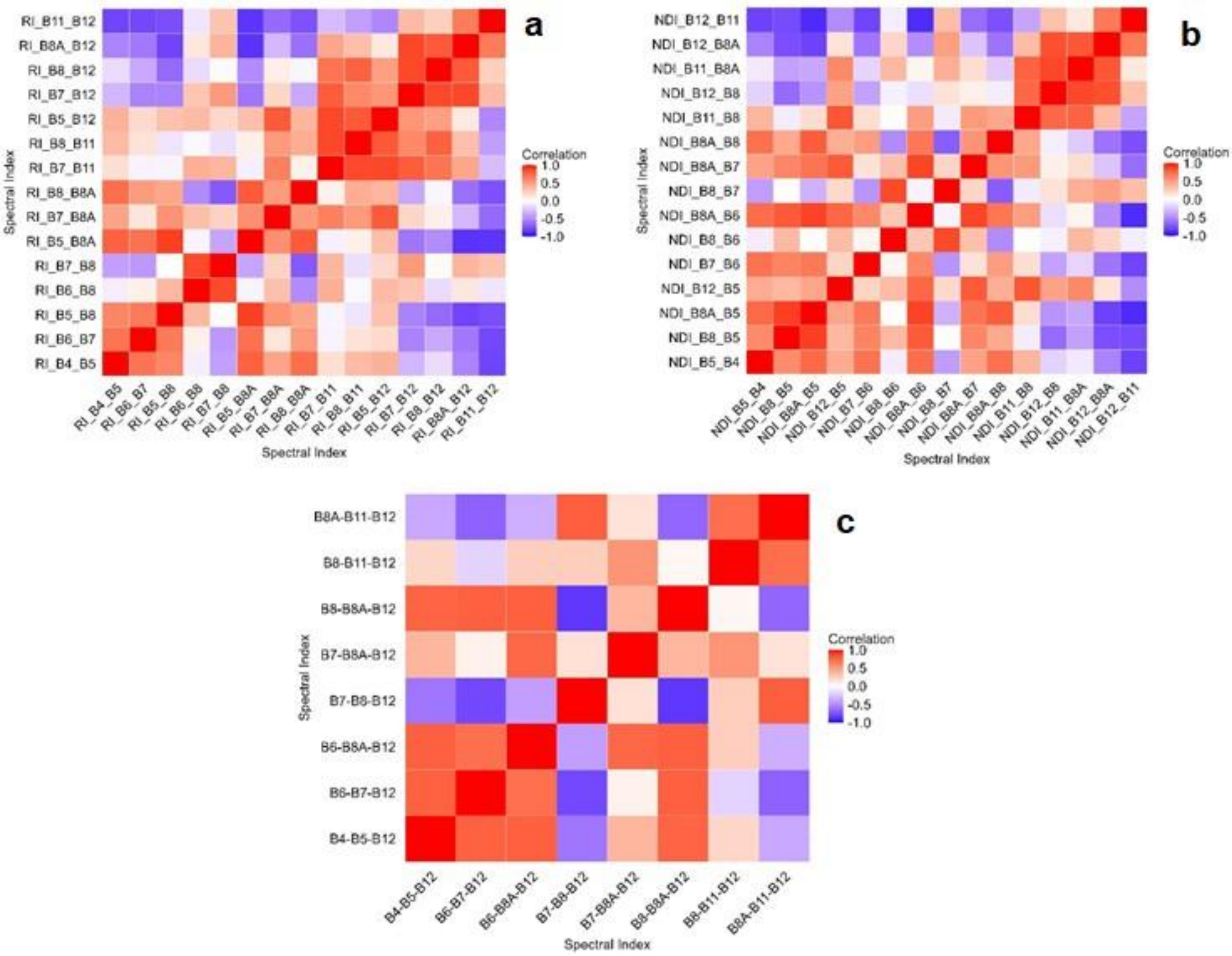


Figure 4-11. Spectral indices with low correlation in winter: (a) RI, (b) NDI, and (c) 3-band algorithm

The final band combinations from **Figures 4-10** and **4-11** were used to determine the relationship between them and Chl-*a* and TSS for summer and winter. Three types of relationships were tested, namely linear, polynomial, and logarithmic. The five highest relationships are presented in **Table 4-2** for each parameter per season. The 3-BDA showed the highest relationship to Chl-*a* and TSS for summer, and only for TSS in winter. For Chl-*a* in winter, the RI and NDI had the strongest relationship. Good correlation was found for TSS both in summer (0.53-0.74) and winter (0.51-0.64) as opposed to Chl-*a*, with poor correlation in both seasons (0.19-0.25 in summer and 0.1-0.12 in winter). Most of the correlations were described by a polynomial relationship, with only three by a linear relationship, and none by a logarithmic relationship. Interestingly, all TSS relationships were significant, while those for Chl-*a* were not.

Table 4-2. Different band combinations with the highest relationship with chlorophyll-a (Chl-*a*) and total suspended solids (TSS) in summer and winter

	Summer				Winter			
	Relationship type	Band combination	R^2	P-value	Relationship type	Band combination	R^2	P-value
Chl-*a*	Polynomial	$\frac{B8-B11}{B12}$	0.25	0.913	Polynomial	$\frac{B6-B7}{B12}$	0.11	0.179
	Polynomial	$\frac{B8A-B11}{B12}$	0.25	0.996	Polynomial	$\frac{B7-B6}{B7+B6}$	0.12	0.207
	Polynomial	$\frac{B8-B8A}{B12}$	0.19	0.649	Polynomial	$\frac{B8A-B6}{B8A-B6}$	0.11	0.264
	Polynomial	$\frac{B4-B5}{B12}$	0.19	0.638	Polynomial	$\frac{B7}{B6}$	0.12	0.215
	Polynomial	$\frac{B5-B8}{B12}$	0.22	0.427	Linear	$\frac{B7-B8A}{B12}$	0.1	0.185
TSS	Polynomial	$\frac{B6-B8}{B12}$	0.63	$9.6e^{-5}$	Polynomial	$\frac{B7-B8A}{B12}$	0.53	$3.98e^{-4}$
	Polynomial	$\frac{B7-B8}{B12}$	0.74	$4.38e^{-6}$	Polynomial	$\frac{B7-B8}{B12}$	0.53	$4.53e^{-4}$
	Polynomial	$\frac{B7-B8A}{B12}$	0.61	$2.51e^{-4}$	Linear	$\frac{B4-B5}{B12}$	0.64	$2.42e^{-5}$
	Polynomial	$\frac{B4-B5}{B12}$	0.59	$4.68e^{-4}$	Polynomial	$\frac{B6-B8A}{B12}$	0.53	$3.84e^{-4}$
	Linear	$\frac{B5-B8}{B12}$	0.51	$3.99e^{-4}$	Polynomial	$\frac{B6-B7}{B12}$	0.55	$2.95e^{-4}$

The criteria were based on the highest correlation and significance. Although not all band combinations for Chl-a were significant, they were still selected to ensure balance and equal representation in the models. (B8-B11)/B12 (R^2 = 0.25) and (B8A-B11)/B12 (R^2 = 0.25) band combinations had the highest correlation for Chl-*a* in summer, while (B6-B8)/B12 (R^2 = 0.63) and (B7-B8)/B12 (R^2 = 0.74) band combinations had the highest correlation for TSS. Their scatter plots are represented in **Figures 4-12**. For winter, B7/B6 (R^2 = 0.12) and (B7-B6)/(B7+B6) (R^2 = 0.12) band combinations had the highest correlation for Chl-*a,* while (B4-B5)/B12 (R^2 = 0.64), and (B6-B7)/B12 (R^2 = 0.55) band combinations had the highest correlation for TSS. **Figures 4-13** represent the scatter plots for the winter band combinations.

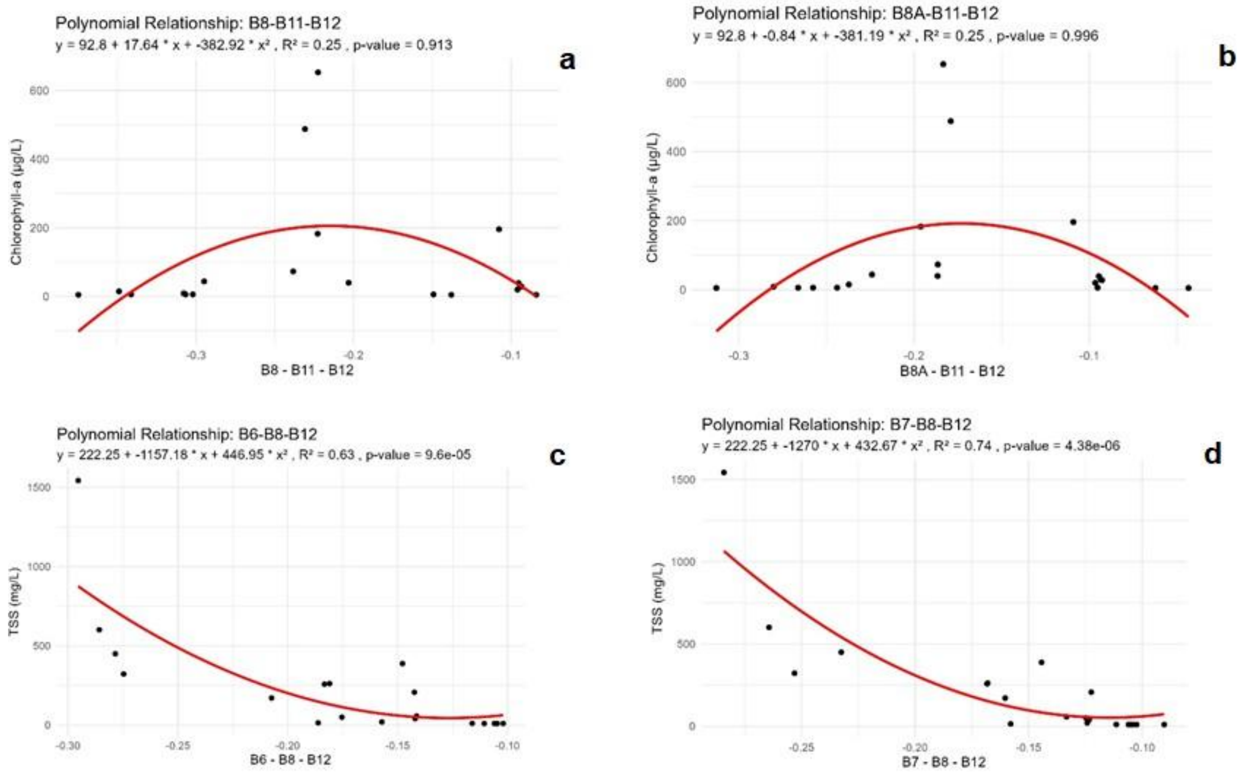


Figure 4-12: Highest correlation between band combinations and Chl-*a* [(a) and (b)] and TSS [(c) and (d)] in summer

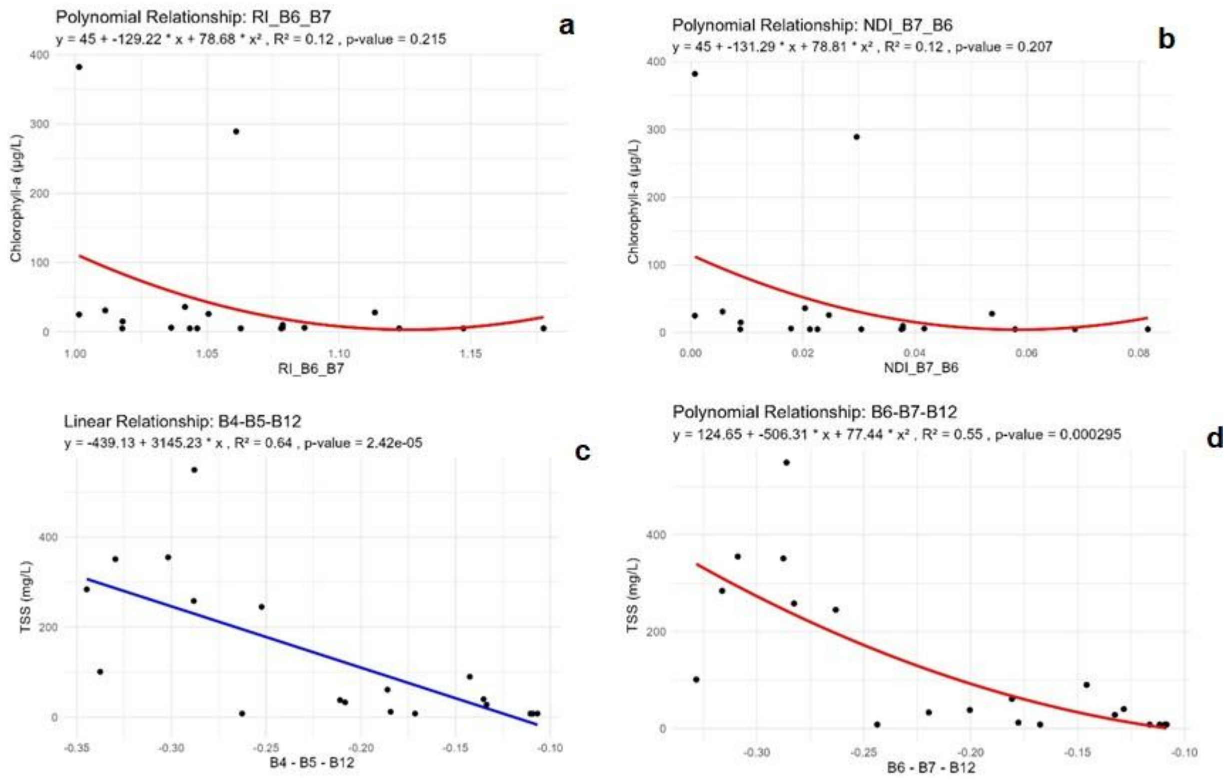


Figure 4-13: Highest correlation between band combinations and Chl-*a* [(a) and (b)] and TSS [(c) and (d)] in winter

4.4.3 Machine learning modelling

4.4.3.1 Random forest regression

Figure 4-14 shows the RFR models for Chl-*a* and TSS in summer and winter. These figures show the actual values on the X-axis and the predicted values on the Y-axis, with the regression line presented in blue. **Figure 4-14 (a)** presents the scatter plot for Chl-*a* predictions, while **Figure 4-14 (b)** presents the TSS predictions in summer. In **Figure 4-14 (a)**, most of the points were clustered around the regression line (blue) with few deviations around 300 µg/L and 450 µg/L, showing moderate performance. The model showed a moderate correlation of 0.47 with overestimations of 4.96 µg/L. The model yielded an RMSE of 96.31 µg/L and MAE of 77.89 µg/L. The TSS model also showed a moderate correlation of 0.44 with an overestimation of the predicted values at 15.73 mg/L. The model delivered a high RMSE of 191.12 mg/L and MAE of 157.29 mg/L. **Figure 4-14 (c)** showcases the scatter plot for Chl-*a* predictions in winter using the RFR model. Winter showed a poor correlation (R^2 = 0.03) between the actual and predicted Chl-*a* values. The model overestimated the Chl- *a* values (bias = 5.26 µg/L) with an RMSE of 70.18 µg/L and MAE of 54.02 µg/L. In contrast, the TSS model in **Figure 4-14 (d)** retrieved moderate predictions with a correlation of 0.49 and an RMSE of 75.38 mg/L. Moreover, the model provided an MAE of 58.08 mg/L with TSS underestimations of less than 3 mg/L. Most of the points were found close to the regression line. However, there were also scattered points after 100 mg/L.

Overall, the RFR models overestimated their predictions. However, the models captured the trends of both Chl-*a* and TSS, especially in low and extreme values, with fewer deviations around moderate values.

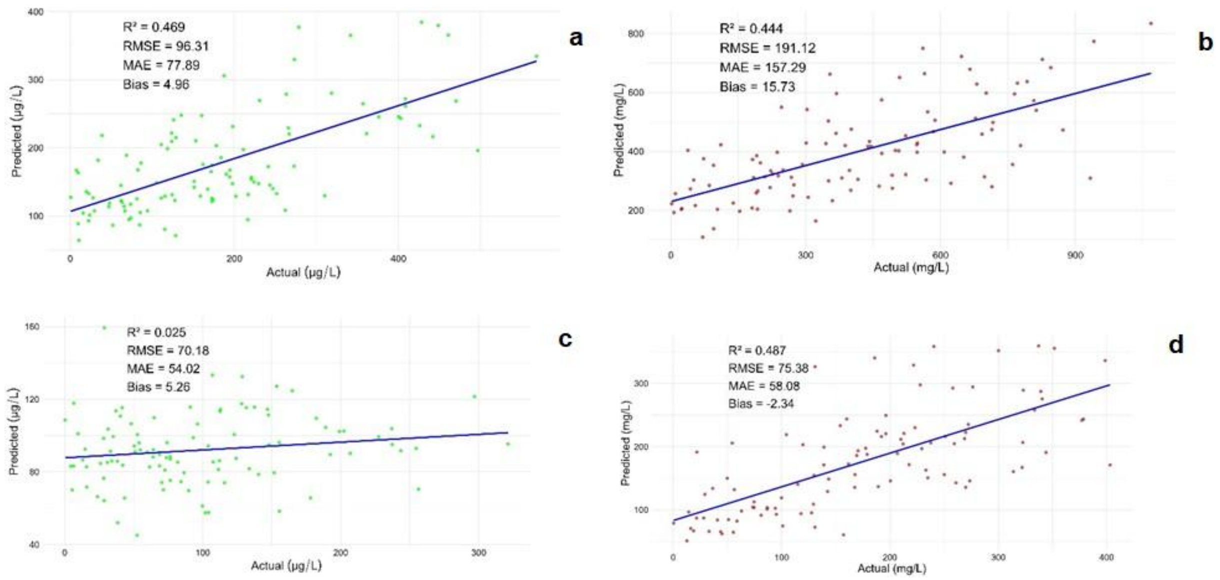


Figure 4-14: RFR scatter plot for Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)]

The advantage of random forest algorithm is that it provides variable importance measure to reveal influential variables to the model predictions. **Figure 4-15 (a)** highlights the importance of different spectral bands in Chl-*a* predictions using the random forest model in summer (high-flow, wet) conditions, while **Figure 4-15 (b**) highlights the importance of spectral bands for TSS also in summer. **Figure 4-15** (a) shows that band 1, band 5, band 8, and band 12 contribute significantly to summer predictions for Chl-*a*. In contrast, band 3, band 6, and band 8A make the lowest contributions. For TSS predictions, bands 6, 7, and 8A contribute very little, while the rest make significant contributions, with bands 3, 5, and index (B7-B8)/B12 contributing more than the others. **Figure 4-15 (c)** presents spectral bands' importance for Chl-*a* predictions in winter (low-flow, dry) conditions, while **Figure 4-15 (d)** shows spectral bands' importance for TSS predictions also in winter using the random forest model. For Chl-*a* predictions, all bands make significant contributions except for bands 8 and 8A, which provide the lowest contributions. Similar to the summer model, band 1 makes the highest contribution compared to the rest, followed by index (B6-B7)/B12. Bands 3, 5, and 12 make similar contributions to the Chl-*a* predictions. For TSS in winter, bands 2 and 4 make more contributions, while band 7 does not contribute at all. Although the arrangement of bands is not the same as the summer model, bands 6, 7, and 8A are still at the bottom, showing the most limited contribution

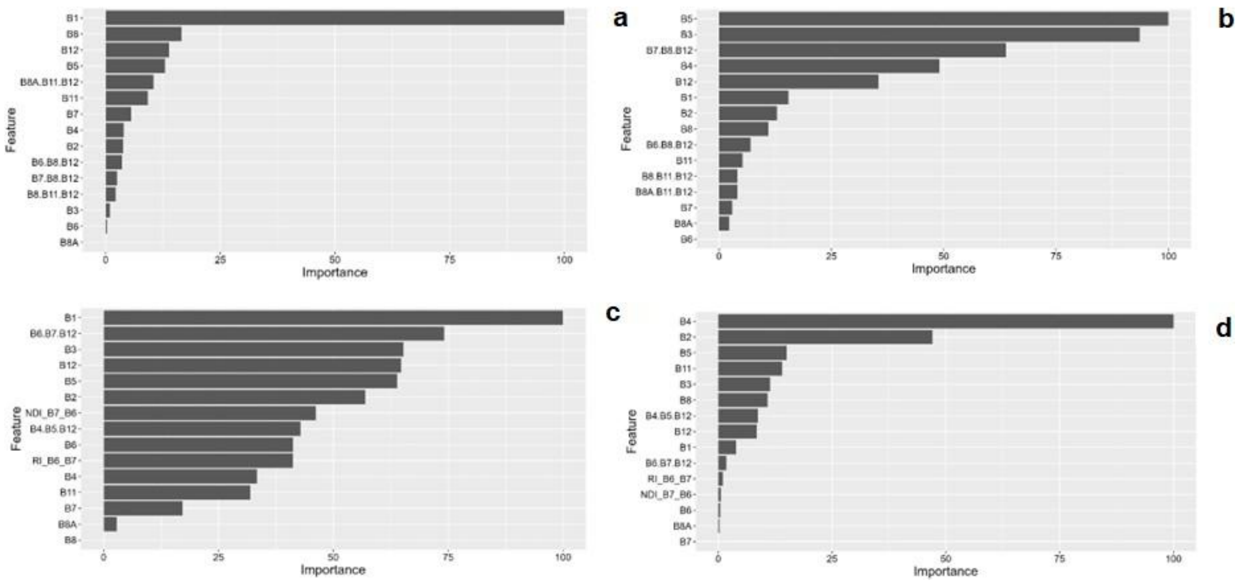


Figure 4-15: Random forest variable importance for Chl-*a* (left) and TSS (right) during summer [(a) and (b)] and winter [(c) and (d)]

4.4.3.2 Gaussian process regression

The predictions of the GPR (single output) models for summer (high-flow, wet) and winter (low-flow, dry) seasons are represented in **Figure 4-16**. **Figure 4-16 (a)** illustrates the results of the GPR model on Chl-*a* predictions in summer. The model produced a good R^2 of 0.62 with an overall overestimation of 8.4 µg/L. Moreover, the model showed a good fit, with all points other than 450 µg/L being close to the regression line. The model had an RMSE and MAE of 84.69 µg/L and 66.76 µg/L, respectively. **Figure 4-16 (b)** shows the GPR predictions of TSS in summer. Like the Chl-*a* model, the TSS model showed good results, achieving 0.52, 182.73 mg/L, and 66.76 mg/L for R^2, RMSE, and MAE, respectively. While this is better than the RFR model for summer, it did overestimate TSS by 18.56 mg/L. Similar to **Figure 14 (a)**, the scatter plot in **Figure 4-16 (b)** has values around the regression line; however, there are outliers throughout the mid and high TSS values. For winter models, **Figures 4-16 (c)** and **(d)** illustrate the Chl-*a* and TSS predictions, respectively. **Figure 4-16 (c)** shows a poor correlation between the predicted and the actual Chl-*a* values, with points scattered over the plot except at high values. However, the model achieved an RMSE of 71.23 µg/L and MAE of 53.56 µg/L with a bias of 7.59 µg/L. The model failed to explain the Chl-*a* trend, similar to the RFR winter model, producing an R^2 of 0.02. The GPR model for TSS performed better than the Chl-*a* model. However, it achieved an RMSE of 71.43 mg/L with an MAE of 53.56 mg/L, similar to the winter Chl-*a* model. Moreover, it achieved an R^2 of 0.54 and a bias of −3.79 mg/L, showing the model's good correlation and low underestimations.

The GPR model over-predicted the lower ranges of Chl-*a*, while predicting the upper ranges well. The model achieved an R^2 of 0.62 in summer, while in winter, the model was poor at predicting Chl-*a*, achieving an R^2 of just 0.01. Meanwhile, the GPR model performed better when predicting TSS both in summer and in winter, with R^2 values of 0.51 and 0.54, respectively. However, the model still over-predicted the lower ranges of TSS in both seasons.

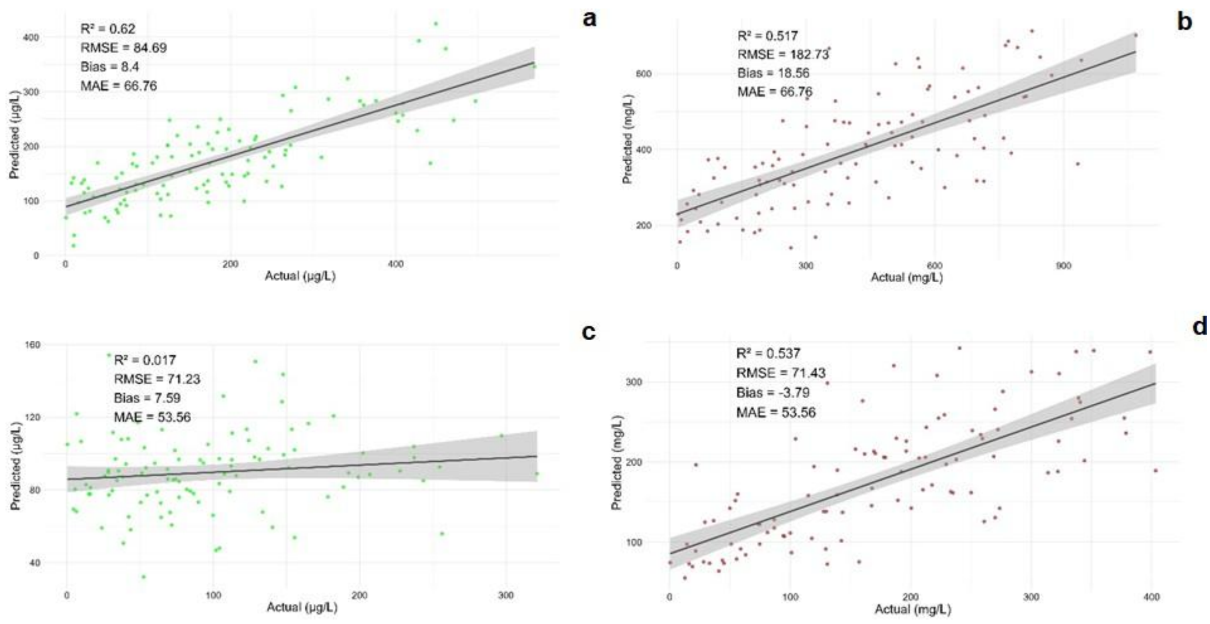


Figure 4-16: Gaussian process regression (GPR) scatter plot for Chl-*a* (left) and TSS (right) in summer [a) and b)] and winter [c) and d)]

4.4.3.3 Random forest survival, regression, and classification

The Chl-*a* and TSS predictions from the RFSRC model in summer are illustrated in **Figures 4-17 (a)** and **(b)**, respectively. For Chl-*a*, the model achieved an R^2 of 0.49 while achieving 0.43 for TSS. The model produced an RMSE of 94.68 µg/L for Chl-*a* and 193.65 mg/L for TSS. The model overestimated both parameters with 4.47 µg/L for Chl-*a* and 8.68 mg/L for TSS. The predictions of the winter RFRSC model are illustrated in **Figures 4-17 (c)** and (**d)** for Chl-*a* and TSS, respectively. The model yielded a poor correlation (R^2 = 0.03) for Chl-*a* predictions but managed to achieve a good correlation (R^2 = 0.47) for TSS. This pattern was also noticed in the RFR and GPR models. Using the correlation between Chl-*a* and TSS, the model produced a balanced prediction with a bias of 4.9 and -4.4 for Chl-*a* and TSS, respectively.

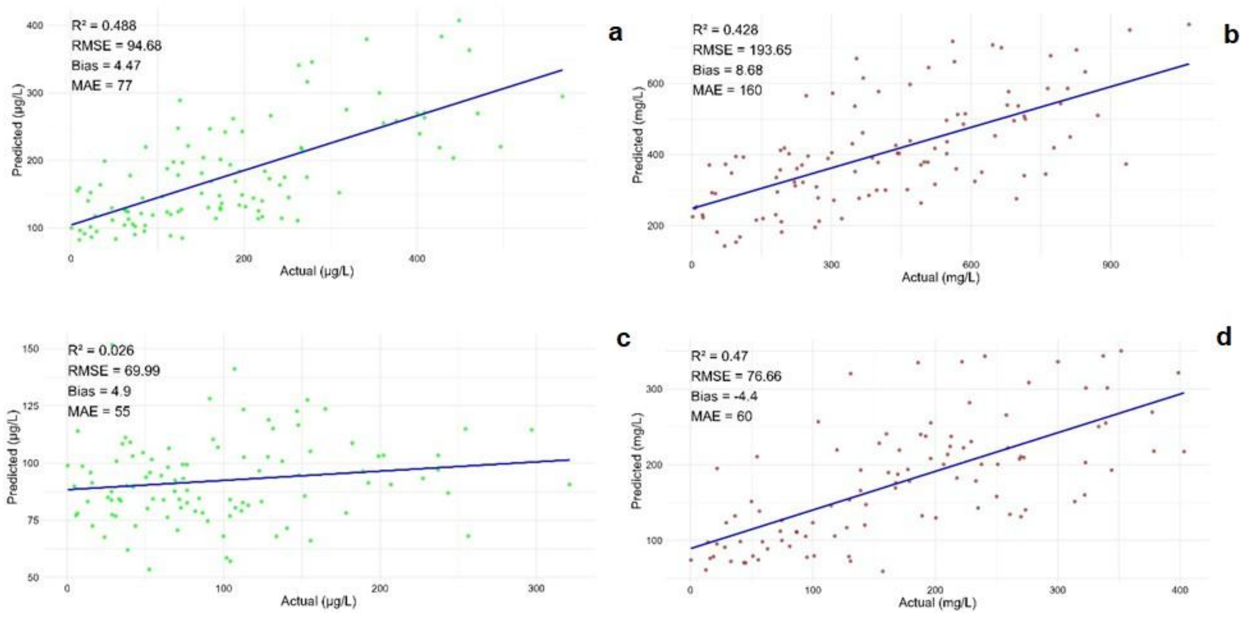


Figure 4-17. RFSRC scatter plot for Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)]

Similar to the regular random forest, the multi-output random forest, i.e., random forest survival, regression, and classification (RFSRC), also provides variable importance. **Figure 4-18 (a)** illustrates the spectral band contributions for Chl-*a* predictions in summer. In contrast, **Figure 4-18 (b)** illustrates the spectral band contributions for TSS predictions. **Figure 4-18 (a)** shows that band 1 makes a significant contribution compared to other bands. Band 5 has the second greatest contribution, while bands 3 and 12 contribute slightly less than band 5. The results indicate the same contributions as the summer RFR model, where bands 1, 5, and 12 contribute significantly towards Chl-*a*. Indices (B6-B8)/B12 and (B8-B11)/B12 have the lowest contribution in Chl-*a* predictions. For TSS, bands 1, 5, and index (B7-B8)/B12 have the highest contribution, while index (B8-B11)/B12 has the least contribution. **Figures 4-18 (c)** and **d)** show the spectral band importance of Chl-*a* and TSS predictions in winter, respectively. As shown in **Figure 4-18 (c)**, all bands contribute to the model; however, bands 7 and 11 make the least contribution to Chl-*a* predictions, while bands 1, 2, and 12 make the highest contribution. In **Figure 4-18 (d)**, bands 2 and 4 contribute the most to TSS predictions. The same observations are observed from the RFR model in winter. Following band 1, band 7 and index (B7-B6)/(B7+B6) make the lowest contribution to TSS prediction.

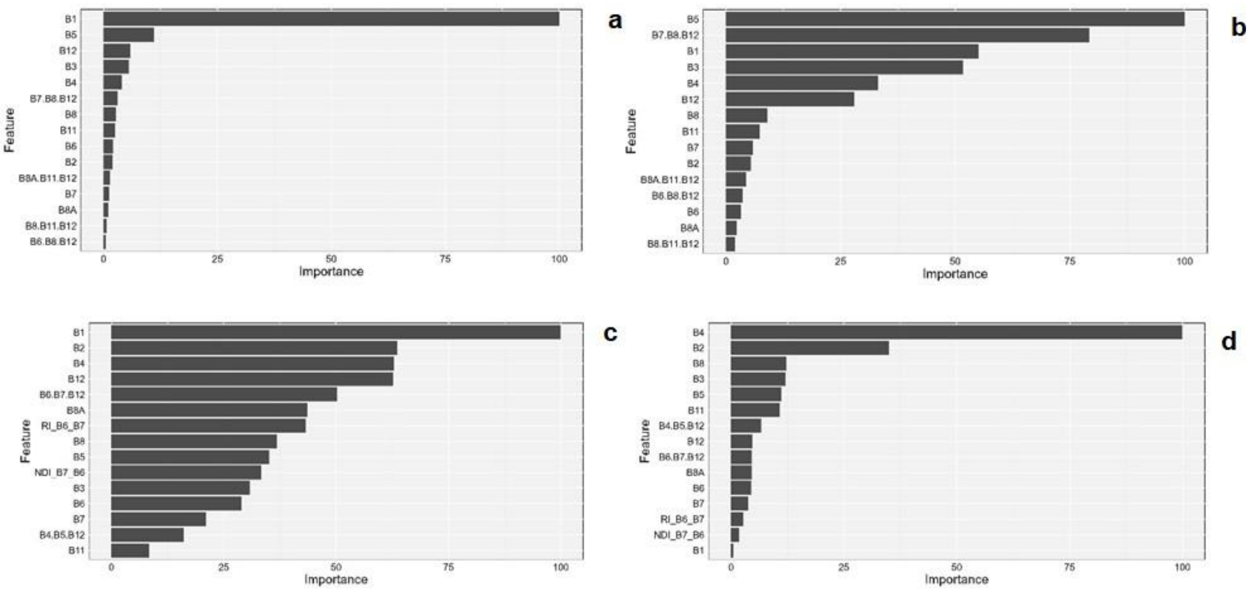


Figure 4-18: RFSRC variable importance for Chl-*a* (left) and TSS (right) during summer [(a) and (b)] and winter [(c) and (d)]

4.4.3.4 Multi-output Gaussian process regression

The predictions of the MGPR model for summer and winter are illustrated in **Figure 4-19** for Chl-*a* and TSS. In summer (See **Figure 4-19 (a)** and **(b)**), the model achieved excellent correlations for both parameters, indicating a good fit between actual and predicted values. For Chl-*a* in **Figure 4-19 (a)**, it achieved an R^2 of 0.95 and 0.89 for TSS in **Figure 4-19 (b)**. The model had a low RMSE of 28 µg/L and MAE of 23.04 µg/L, with a slight overestimation of 1.24 µg/L for Chl-*a*. For TSS, the model also performed well, with an RMSE of 84.58 mg/L, MAE of 67.77 mg/L, and overestimated the predictions with a bias of 7.13 mg/L. The model best fitted the values for Chl-*a* and TSS, with all the points aligned with the regression line and falling within the 95% confidence estimation (grey area). For winter, the MGPR model maintained its robustness for TSS. **Figure 4-19 (d)** shows the TSS predictions, where the model achieved an R^2 of 0.94 with an overestimation of 3.12 mg/L. The RMSE and MAE were 26 mg/L and 20.78 mg/L, respectively. The model had poor explanatory performance for Chl-*a* values (See **Figure 4-19 (c)**) with an R^2 of 0.23 and a bias of 5.35 µg/L. Additionally, it achieved an RMSE of 62.16 µg/L and MAE of 49.26 µg/L. However, similar to the summer model, the winter model best fitted the values for Chl-*a* and TSS, with all the points aligned with the regression line and falling within the 95% confidence estimation.

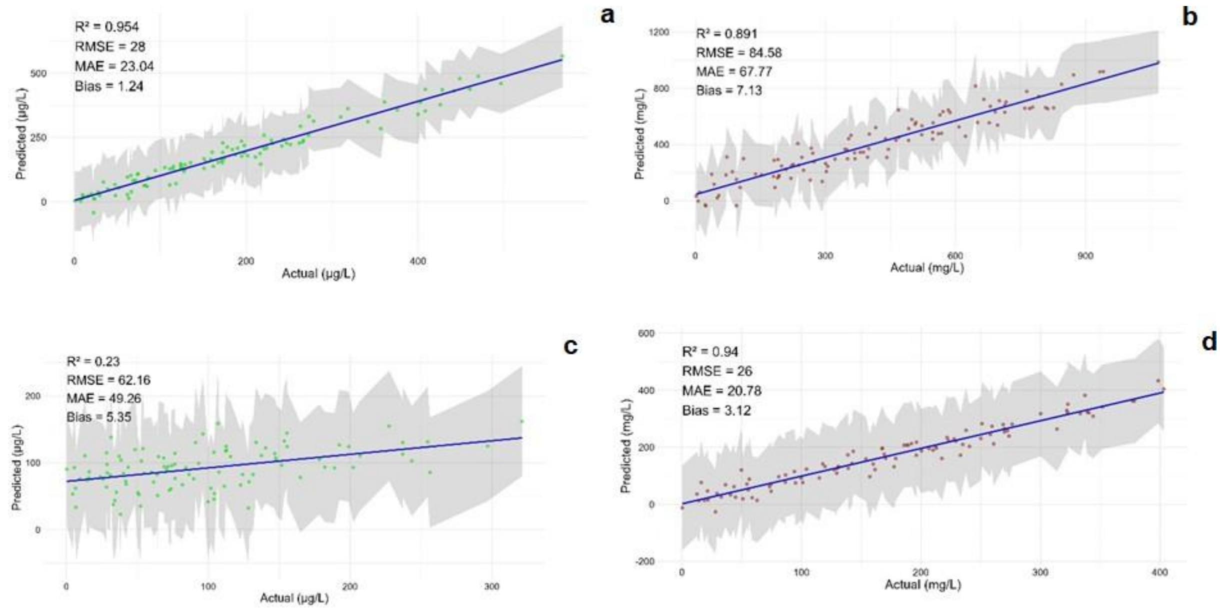


Figure 4-19. Multi-output Gaussian process scatter plot for Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)]

4.4.4 Density distribution of model predictions

Figure 4-20 shows the density distributions of GPR and MGPR (in green and yellow, respectively), with smooth density curves for each model. In summer, the actual concentrations are more skewed towards low to moderate concentrations of Chl-*a*, with the MGPR and GPR predictions following the same trend. However, the GPR density curve is not as widespread as the actual data and the MGPR predictions. Only a few predictions fell in the low concentration range (<50 µg/L), with peak predictions for 100-200 µg/L. Both models failed to predict Chl-*a* during winter, especially at low (<20 µg/L) and high (>170 µg/L) concentrations. The models showed peaks in the moderate (70-110 µg/L) concentration range, with GPR higher and narrower than MGPR. Both in summer and winter, the concentrations for TSS were across a wide range, with the density curve being flatter. Similar to Chl-*a* predictions, the MGPR depicts comparable trends to the actual concentrations, with overlaps in the low, moderate, and high concentrations. In contrast, the GPR predictions were more concentrated on the moderate concentrations, failing with low (<150 mg/L and <40 mg/L) and high (>750 mg/L and >350 mg/L) concentrations in summer and winter, respectively. **Figure 4-21** shows the density distributions of random forest and RFSRC in red and purple, with smooth density curves for each, respectively. The models show similar distributions that differ from actual concentrations, with poor predictions for low and high concentrations for both parameters in summer and winter. The predictions of the models are skewed towards moderate concentrations, with Chl-*a* having a narrower density curve. Overall, the MGPR shows realistic predictions for both seasons, with only small deviations from the actual concentrations.

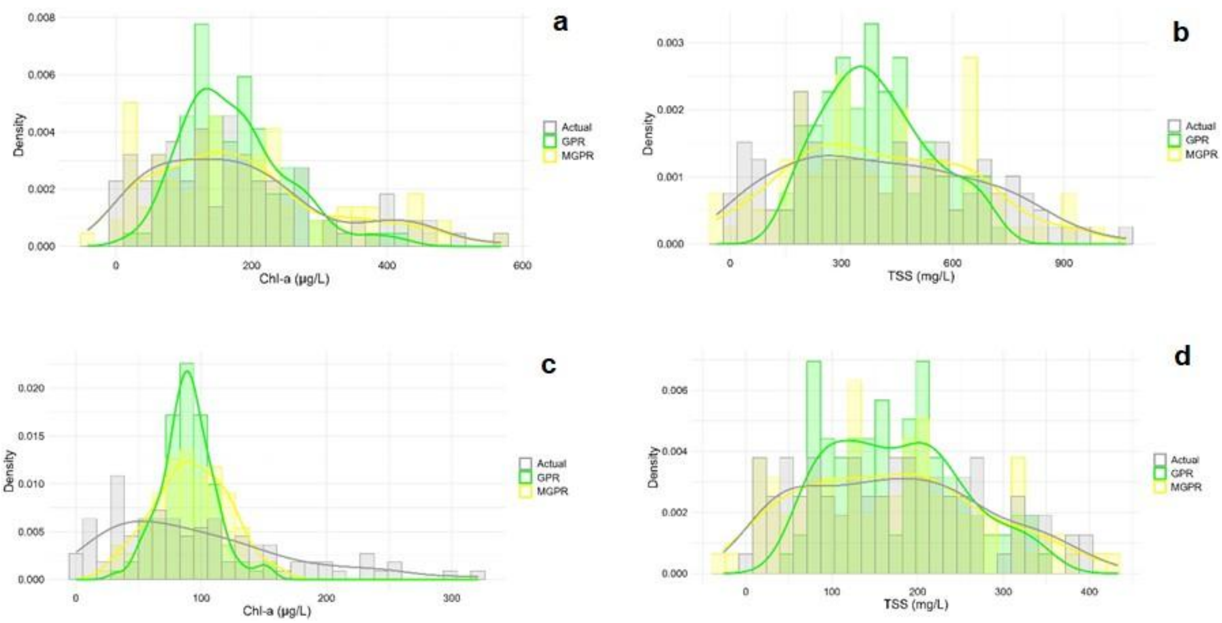


Figure 4-20. Density distribution of Gaussian process regression (GPR) and multi-output Gaussian process regression (MGPR) predictions of Chl-*a* (right) and TSS (left) in summer [(a) and (b)] and winter [(c) and (d)]

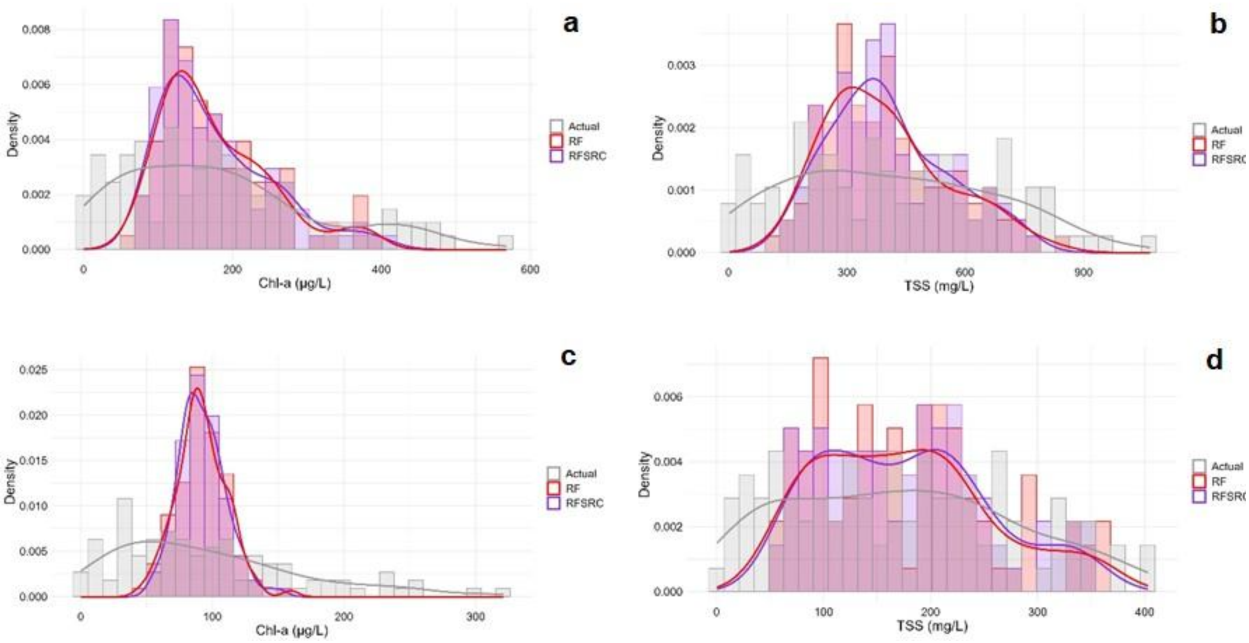


Figure 4-21. Density distribution of random forest and RFSRC predictions of Chl-*a* (right) and TSS (left) in summer [(a) and (b)] and winter [(c) and (d)]

4.4.5 Spatial distribution maps

Figure 4-22 shows the spatial distribution using the RFSRC models. The low concentrations of Chl-*a* and TSS are represented in blue, while the highest concentrations are shown in yellow. In contrast, the moderate concentrations for Chl-*a* and TSS are represented in green and red, respectively. The concentrations of both parameters in summer were double those in winter. The highest concentrations for both Chl-*a* and TSS were generally found at the edges of the dam, while the lowest concentrations were found in the middle of the dam, both in summer and winter. Chl-*a* concentrations in winter were an exception, being vice versa. Similar to **Figure 4-22**, **Figure 4-23** shows the spatial distribution of concentration parameters using the MGPR models. The concentrations of parameters in **Figure 4-23** were dominated by moderate to high concentrations.

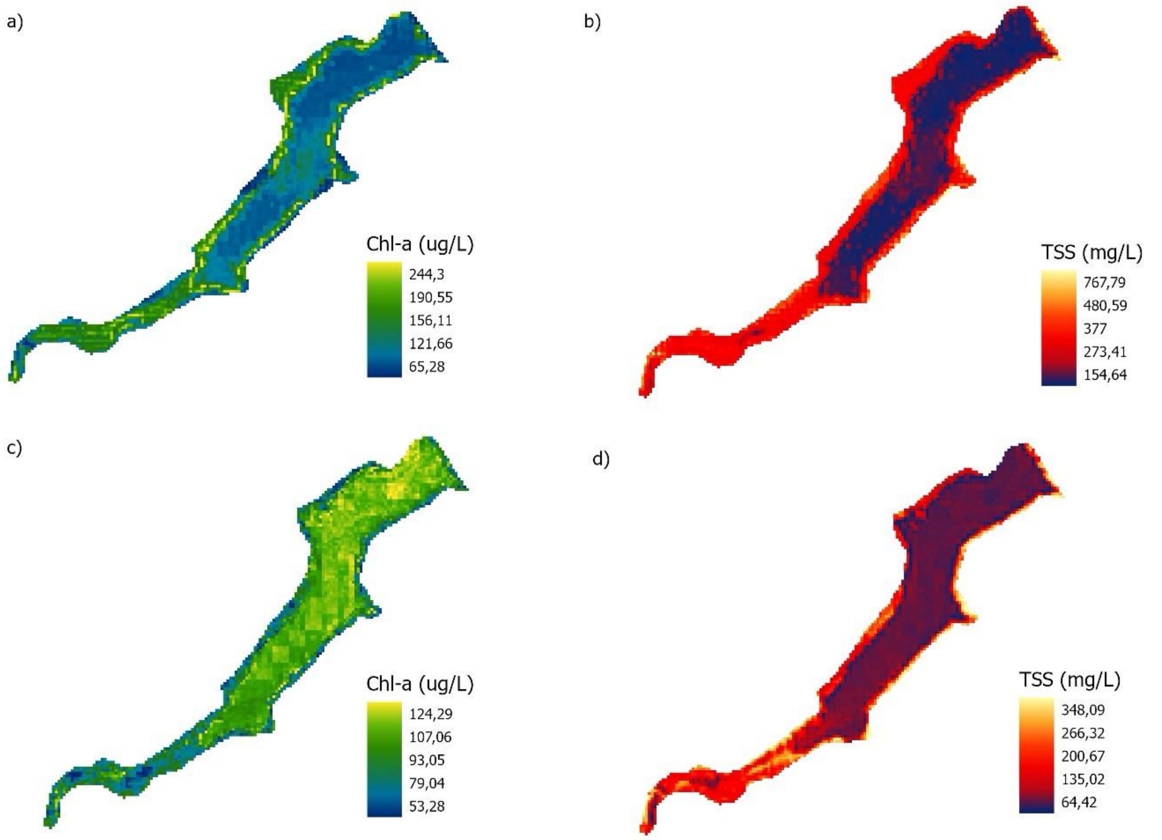


Figure 4-22. Spatial distribution of Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)] using the RFSRC model

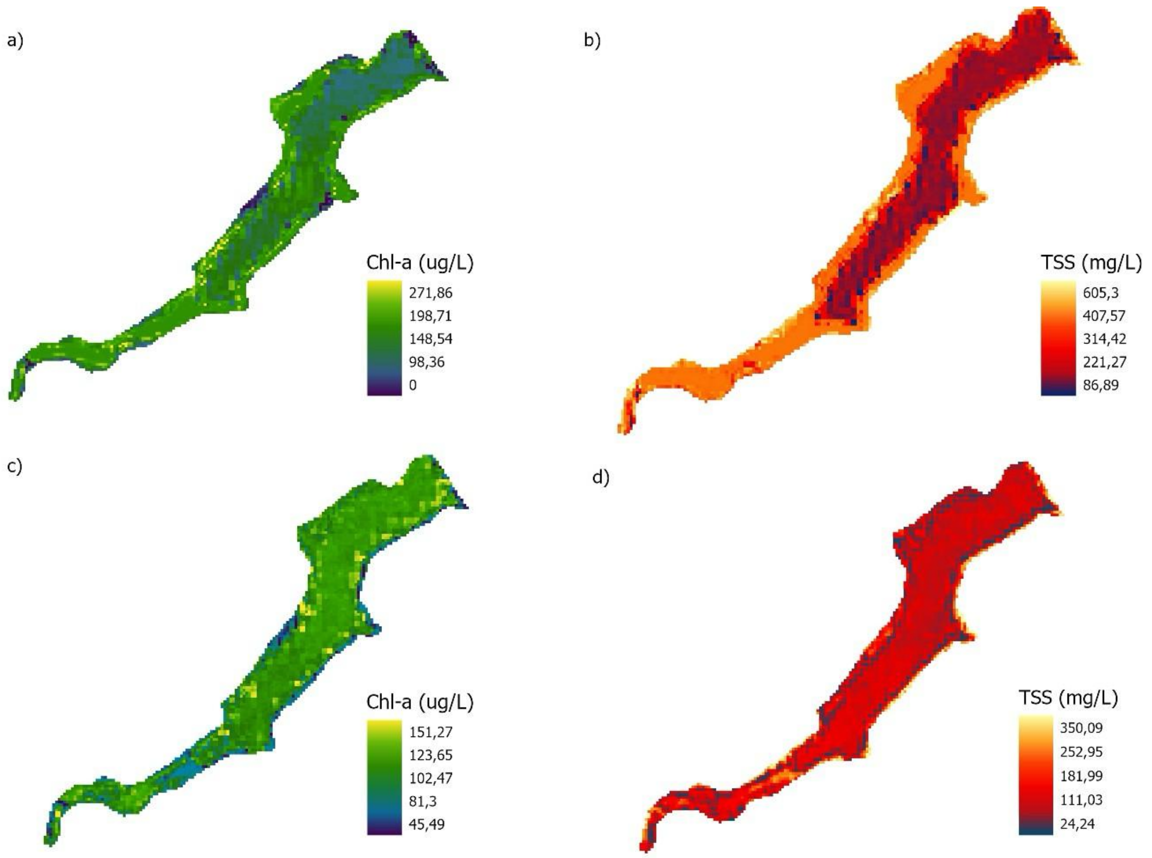


Figure 4-23. Spatial distribution of Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)] using the multi-output Gaussian process regression (MGPR) model

4.5 MODEL TRANSFERABILITY

4.5.1 Spatial distribution

Figure 4-24 illustrates the spatial distribution of Chl-*a* and TSS, respectively, in summer ([a] and [b] and winter [c] and [d]). **Figure 4-24 (a)** shows the spatial distribution of Chl-a in summer, while **Figure 4-24 (c)** shows the spatial distribution of Chl-*a* in winter. Low values of Chl-*a* are shown in blue, while the highest are shown in dark green to yellow. The concentrations for summer ranged from 55 µg/L to 422 µg/L, while in winter they ranged from 46 µg/L to 141 µg/L. High concentrations were found at Roodekoppies Dam in summer and Hartbeespoort Dam in winter. Moreover, these concentrations were found in the middle of the dams, while low concentrations were found near the edges. In contrast to Chl-*a* distributions, high TSS concentrations (dark brown) were found along the edges, and low concentrations (blue) were found in the middle, except in Roodekoppies Dam. The TSS distribution is represented in **Figure 4-24 (b)** for summer and **Figure 4-24 (d)** for winter. The concentration ranges for summer were 156 mg/L to 773 mg/L and 60 mg/L to 364 mg/L for winter.

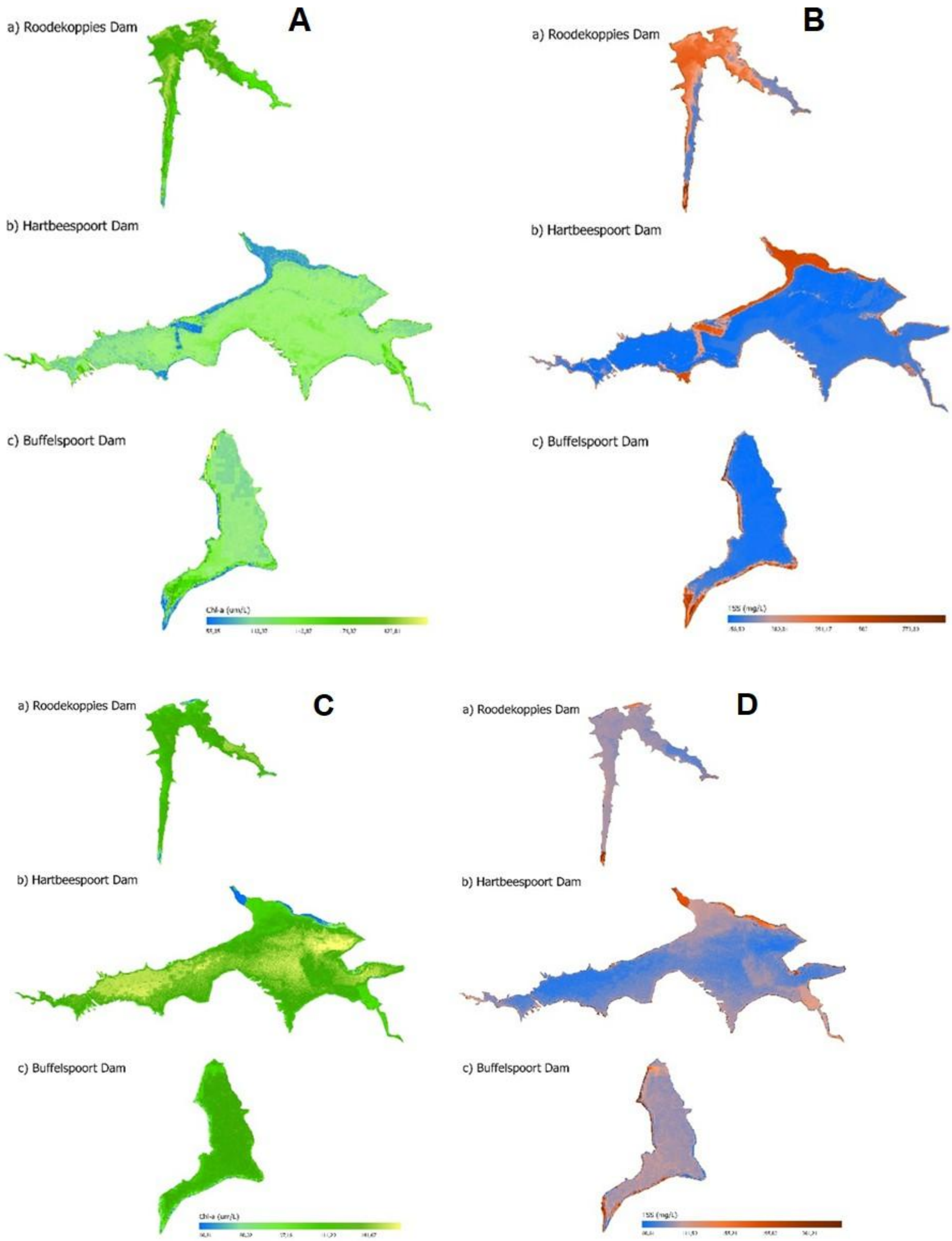
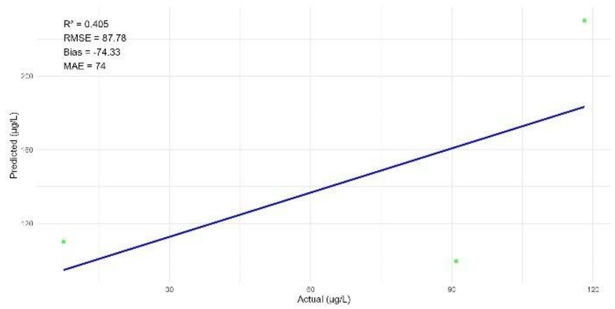


Figure 4-24. Spatial distribution of Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)] using RFSRC

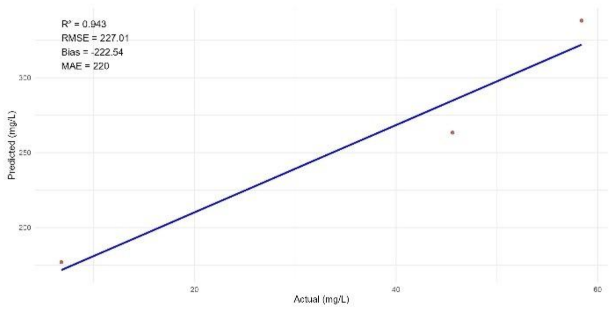
4.5.2 Validation of transferred OAP maps

Figure 4-25 shows the scatter plots for Chl-*a* and TSS in summer ([a] and [b]) and winter ([c] and [d]). All the models performed well with an R^2 of over 90%, except for Chl-*a* in summer.

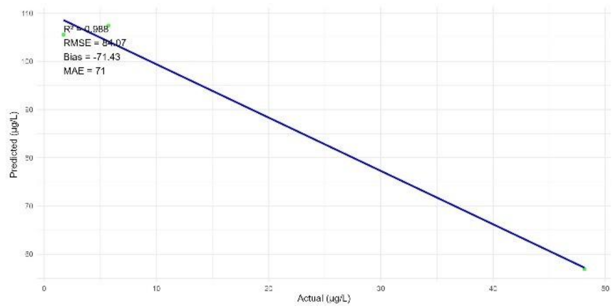
a)



b)



c)



d)

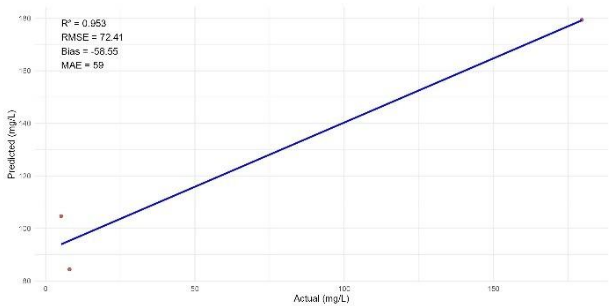


Figure 4-25. Scatter plots of Chl-*a* (left) and TSS (right) in summer [(a) and (b)] and winter [(c) and (d)] using RFSRC

4.6 DISCUSSION

The study aimed to develop a robust model for estimating water quality parameters (i.e. Chl-*a* and TSS) in the Cradle of Humankind in South Africa. First, it evaluated different spectral indices, including RI, NDI, and 3-band algorithms, in estimating Chl-*a* and TSS. Two indices with the highest R^2 were selected for each parameter per season, regardless of the type, and were combined with individual bands to build ML models. Second, four ML algorithms were tested, which included two single-output RFR and GPR, their multi-output models, RFSRC and MGPR. Lastly, the multi-output algorithms were tested for their transferability in three nearby dams (i.e., Hartbeespoort, Roodekopjes, and Buffelspoort Dams).

4.6.1 Performance of spectral indices in estimating Chl-*a* and TSS

Previous research has extensively tested various spectral indices to estimate Chl-*a* and TSS, particularly in inland and coastal water bodies. Among these, RI (Pahlevan et al., 2020b), NDI (Shahzad et al., 2018), 2-BDA (Dall'Olmo & Gitelson, 2006), and 3-BDA (Gitelson et al., 2008) have received significant attention. These spectral indices primarily utilise bands within the visible and near-infrared regions, which are sensitive to pigment absorption and scattering processes in water. However, the performance of these indices has varied across geographical regions and water types owing to the complex and dynamic nature of optical water quality parameters, especially in turbid and deep waters. Moreover, many of these studies focused on single or homogeneous water bodies, thereby limiting the generalisability of their findings.

This study addresses the challenge of estimating Chl-*a* and TSS in a heterogeneous aquatic system that exhibits both clear and turbid water conditions, typical of the Cradle of Humankind region. All Sentinel-2 bands, excluding B9 and B10 (commonly reserved for atmospheric correction), were evaluated using established spectral index structures such as RI, NDI, and 3-BDA to account for this variability. This comprehensive approach is aligned with the methodology proposed by Van Nguyen et al. (2019), who examined red and red-edge band combinations for Chl-*a* retrieval. However, the current study extends this work by incorporating eleven Sentinel-2 bands to form a broader range of band combinations across multiple indices. Given the large number of potential combinations, computational efficiency was managed by using heat maps to identify and eliminate highly correlated indices, thus reducing redundancy and retaining indices with unique spectral contributions.

Interestingly, most of the optimal spectral indices identified for both Chl-*a* and TSS included band 12 (SWIR 2), in both summer and winter seasons. This finding supports the recommendation by Knaeps et al. (2015), who emphasised the utility of SWIR bands for TSS retrieval in highly turbid waters, where NIR (Band 8) sensitivity declines. This study evaluated various band combinations seasonally, and the five best-performing indices for each parameter were identified (see **Table 4-2**). For Chl-*a* during the summer, the 3-BDA yielded the highest correlation (R^2 = 0.25), albeit still modest, and outperformed both RI and NDI indices. In contrast, during the winter, RI and NDI achieved slightly better correlations (R^2 = 0.12) than 3-BDA, though these relationships remained weak and statistically insignificant. Similar results were observed by Mpakairi et al. (2024) in Nandoni Dam, South Africa, where spectral indices failed to produce statistically significant models for Chl-*a* estimation, underscoring the limitations of index-based approaches in certain water bodies. In contrast, TSS estimation showed markedly stronger results. The 3-BDA consistently outperformed RI and NDI in both seasons, with R^2 values of 0.64 in summer and 0.74 in winter, suggesting a more robust and stable relationship. This superior performance of the 3-BDA index aligns with findings from other studies conducted in turbid waters (Buma and Lee, 2020; Toming et al., 2016; Watanabe et al., 2018), further validating its effectiveness. The polynomial nature of the relationships observed for these indices reflects the non-linear interactions between reflectance values and in-water constituents, a phenomenon commonly reported in optical water quality studies (Gholizadeh et al., 2016). These findings emphasise the importance of selecting appropriate spectral bands and indices tailored to the optical properties of diverse aquatic systems. While traditional indices such as RI and NDI may offer value in clearer waters or specific contexts, their limitations in turbid environments suggest the need for advanced approaches, such as ML models or multiband combinations, to capture the complex dynamics of Chl-*a* and TSS more effectively.

4.6.2 Performance of machine learning algorithms

It is ideal to have extensive ground points; however, the number of points in this study is small due to limited funds and resources. The ML algorithms generally benefit from large and diverse datasets to capture system variability and improve generalisation capabilities. To achieve the desired sample size, synthetic data was created by simulating the satellite and laboratory data using MEDM. This non-parametric method offers several advantages in water quality monitoring. It preserves the complex interdependencies among water quality variables like Chl-*a* and TSS, ensuring that the simulated data accurately reflects real-world conditions (Smania and Jonsson, 2021). The MEDM is particularly useful in water quality owing to its flexibility, as it does not assume any predefined statistical distribution, making it robust against skewed data, outliers, or multimodal distributions that are common in water quality datasets. Furthermore, the method allows for realistic simulation of both high and low-value extremes, helping to ensure that the model is well-calibrated across the full range of possible values. This was evident in our results, where there was substantial variation in both TSS and Chl-*a* values, yet the synthetic dataset maintained a good representation of this variability. Studies such as Teutonico et al. (2015) have demonstrated the usefulness of MEDM, supporting its effectiveness and reliability, which aligns with the outcomes observed in our research. Nevertheless, the unconstrained nature of the method sometimes results in the generation of unrealistic negative values; in this study, these were excluded to maintain physical plausibility. The simulated data was used to train and test the performance of ML algorithms, divided into an 80% training dataset and a 20% testing dataset. These included RFR, RFSRC, GPR, and MGPR. Each model was implemented with tailored hyperparameter tuning strategies that closely align with methodologies in previous literature.

The RFR model, implemented separately for Chl-*a* and TSS, utilised a five-fold cross-validation approach with an expanded grid of *mtry* values ranging from 1 to 8 and 500 trees per model. Such hyperparameter tuning is essential to avoid overfitting and ensure generalisability. Strong performance has been consistently shown by RFR in environmental applications, particularly owing to its ensemble structure and resistance to multicollinearity. For example, Leggesse et al. (2023) and Kupssinskü et al. (2020) achieved R^2 values above 0.7 when using RFR to predict Chl-*a* and TSS in various aquatic systems. In this study, RFR performed moderately in summer with R^2 values of 0.49 for Chl-*a* and 0.44 for TSS, in line with results from N. Joshi et al. (2024) and Arias- Rodriguez et al. (2023), who reported comparable R^2 values of 0.55 and 0.41, respectively. However, RFR performance deteriorated during winter, dropping to an R^2 of 0.03 for Chl-*a* and 0.49 for TSS, likely reflecting seasonal shifts in water optical properties and phytoplankton dynamics. Such patterns were also observed by Arias-Rodriguez et al. (2020), who noted lower Chl-*a* predictability in colder months. Sigma optimisation through cross-validation was employed by GPR, implemented using the *gausspr* function in R with a radial basis function (RBF) kernel. This model allows for flexible non-linear regression while incorporating uncertainty quantification, making it particularly suitable for environmental systems with high variability. In this study, GPR outperformed RFR in both summer and winter. It achieved an R^2 of 0.62 for Chl-*a* and 0.52 for TSS in summer, and an R^2 of 0.02 for Chl-*a* and 0.54 for TSS in winter. These findings corroborate previous comparisons by Arias-Rodriguez et al. (2020) and Nguyen et al. (2021), who demonstrated the superior generalisation ability of GPR over RFR, especially when dealing with small or noisy datasets. Although both RFR and GPR produced comparable RMSE and MAE values, GPR exhibited lower bias and better trend capture, likely because of its probabilistic framework and kernel-based flexibility.

The RFSRC model extended the RFR approach to multivariate outputs using a unified tree structure (Shen et al., 2020). In summer, the model produced average correlations for both Chl-*a* and TSS that were 0.49 and 0.43, respectively. The same trend observed in random forest and GPR in winter, of low correlation in Chl-*a* and good correlation for TSS, is followed by the RFSRC model. This suggests that the multivariate structure of RFSRC may offer limited benefits under constrained training data conditions. In contrast, the MGPR model, implemented via the *mgpr* package using an RBF kernel and linear mean function, demonstrated superior predictive power. Hyperparameter tuning was carried out using a five-fold cross-validation approach embedded in a simulated annealing optimiser with constraints on kernel width, correlation length, and noise variance. This flexible and rigorous optimisation framework allowed MGPR to simultaneously model both input-output relationships and inter-output dependencies, making it ideal for the joint prediction of Chl-*a* and TSS. In summer, MGPR achieved exceptional performance with R^2 values of 0.95 for Chl-*a* and 0.89 for TSS. In winter, performance remained strong for TSS (R^2 = 0.94) and moderately improved for Chl-*a* (R^2 = 0.23), still outperforming all other models. These findings are consistent with those of Varvia et al. (2023), who highlighted the strength of multi-output Gaussian processes in capturing cross-target correlations and handling complex, high-dimensional environmental datasets.

Beyond prediction, RFR and RFSRC models enabled analysis of variable importance, shedding light on the spectral drivers of Chl-*a* and TSS. Band 1 (ultra-blue) was the most influential predictor for Chl-*a* in both summer and winter, aligning with Joshi et al. (2024), who identified high sensitivity of this band to phytoplankton absorption. Band 8 (NIR) was also significant during summer, suggesting enhanced reflectance in vegetated or organic-rich waters. For winter Chl-*a* prediction, additional contributors included bands 2 (blue), 3 (green), 12 (SWIR 2), and the index B6/(B7-B12), indicating a broader spectral influence in turbid conditions. For TSS, bands 5 (red edge) and 3 (green) were key in summer, aligning with Nechad et al. (2009) and Saberioon et al. (2023), who linked these bands to high backscattering properties of suspended sediments. In winter, bands 2 and 4 dominated TSS predictions, likely owing to their sensitivity to changes in particle size and concentration under low-light conditions. These spectral trends were also reflected in RFSRC outputs, which highlighted bands 1, 2, 4, and 12 as essential for winter predictions and bands 3 and 5 for summer, reinforcing the utility of both visible and shortwave infrared wavelengths in inland water monitoring.

This study contributes substantially to the evolving field of remote sensing-based water quality monitoring, particularly in regions where data scarcity and resource constraints hinder traditional in situ assessments. Focusing on the Cradle of Humankind, a UNESCO (United Nations Educational, Scientific and Cultural Organization) World Heritage Site located in South Africa, this research explores inland waters that range from clear, oligotrophic systems to highly turbid, eutrophic environments. Such spectral and optical diversity presents a significant challenge for remote estimation of parameters like Chl-*a* and TSS. Despite this heterogeneity, the MGPR model consistently outperformed other models across seasons, underscoring its robustness and adaptability. The ability of MGPR to simultaneously model multiple interdependent water quality variables and capture complex non-linear relationships makes it especially suitable for environments characterised by variable turbidity, seasonal shifts, and multiple pollution sources. These results offer valuable insights for the development of transferable models that can be adapted to other regions facing similar environmental variability. In the broader context of South Africa, where water scarcity is exacerbated by pollution, population growth, and under-resourced water governance structures, such tools are urgently needed. The DWS often struggles to meet monthly water quality sampling targets, particularly in rural or remote catchments. In this light, the integration of freely available satellite imagery with ML methods offers a cost-effective, spatially continuous alternative for supplementing or replacing conventional monitoring strategies. This not only addresses national water management needs but also serves the broader public by enabling evidence-based decision-making, safeguarding ecosystem health, and promoting community water security.

4.6.3 Transferability of multi-output algorithms

The RFSRC models were able to estimate Chl-*a* and TSS in the transfer sites using the learnt relationships between the laboratory measurements and Sentinel-2 reflectance from the Cradle of Humankind. The spatial distribution maps were created using the predicted Chl-*a* and TSS images on the transfer sites. These maps show that both Chl-*a* and TSS values are high in summer compared to winter, which is the same trend observed on the training site. Higher Chl-*a* and TSS values are observed on the Roodekoppies Dam in summer. In winter, the highest Chl-*a* values are observed on the Hartbeespoort Dam, while all the dams show the same levels and patterns for TSS. High Chl-*a* values are found in the middle of the dams both in summer and winter. In contrast, high TSS values are found by the dam edges except for Roodekoppies Dam where high values are at the centre. The summer model maintained its moderate performance in estimating Chl-*a* by achieving an R^2 of 0.46 on the transfer sites, similar to the training site (R^2 of 0.49), as evidenced in **Figure 4-20** and **Figure 4-21**. On the contrary, the model highly improved its accuracy in estimating Chl-*a* in winter by achieving an R^2 of 0.99. In addition, the model also improved its estimations of TSS both in summer and winter by explaining over 94% of the data in the transfer sites. The results show that the RFSRC model is transferable by achieving better accuracy on the transfer sites compared to the training site. However, the small validation data may have had an impact on the improvements of the model, as only a few datasets had to be explained.

4.7 SUMMARY

South Africa faces water quality challenges linked to agricultural development, informal settlements, AMD, and infrastructure issues. To address escalating water quality concerns, consistent monitoring is crucial for policymakers and water managers to implement effective mitigation strategies. Although studies (Obaid et al., 2024; Oliphant et al., 2018; Molekoa et al., 2022) have been conducted in South Africa using remote sensing, none have been undertaken using Sentinel-2 and ML, specifically advanced machine learning regression algorithms such as GPR. The study aimed to build robust ML models to estimate OAPs and NOAPs in the COHWHS.

This chapter performed several analyses, showing the following key results:

- Sentinel-2 spectral bands are sensitive to water quality parameters. However, this sensitivity declines with a change in season. During the period of high flow, Sentinel-2 bands had a higher relationship with DO and pH, and this relationship deteriorated during the period of low flow. The results presented in this report emanate from the analysis of optical non-active water quality parameters. It is recommended that results from optically active water quality parameters be included, enabling the study to explain the relationship between Sentinel-2 data and DO and pH.
- In producing a five-year profile of water quality parameters, GPR was used. The GPR model for Chl-*a* achieved an accuracy of 75% and RMSE of 57.67 µg/L, while the model for suspended solids achieved 73% accuracy and RMSE of 132.15 µg/L. Using the validated GPR models for each respective OAP, spatial predictions for the years 2020 to 2023 were made using back-predictions. Overall, the GPR models developed in this study were able to adjust to changing conditions observed in the past five years (2020-2024). See Chapter for further details.
- By evaluating and comparing RFR, RFSRC, GPR, and MGPR, this project identified MGPR as the superior model, particularly in the summer season. An R^2 of 0.95 and 0.89 was achieved by MGPR for Chl-*a* and TSS, respectively, in summer. Further, high accuracy for TSS was maintained in winter (R^2 = 0.94), even in the face of spectral variability across water types. This model's ability to leverage interdependencies between output variables and account for the non-Gaussian nature of environmental data makes it especially effective in heterogeneous aquatic systems. The high performance of MGPR across both clear and turbid water bodies in the Cradle of Humankind highlights its robustness and indicates its suitability for deployment in other diverse environments. The RFSRC was valuable and transferable with improved accuracy for transfer sites, achieving over 94% accuracy in summer and winter, except for Chl-*a* in summer. Remote sensing, when combined with ML, offers a scalable and timely solution for addressing the country's pressing water quality challenges, especially considering increased anthropogenic pressures and climate variability. Ultimately, this study underscores the power of combining spectral data with multivariate ML for sustainable water resource management in resource-limited settings. It paves the way for innovative, technology-driven approaches that enhance monitoring capabilities, inform policy, and protect vulnerable freshwater ecosystems.

CHAPTER 5: COMMUNITY ENGAGEMENT

5.1 INTRODUCTION

This chapter provides details on the community engagement strategy and activities undertaken to ensure that the findings of the project are effectively communicated and its methodologies are adopted by a wide range of stakeholders for the enduring protection of the COHWHS.

5.2 APPROACH TO COMMUNITY ENGAGEMENT

This section presents the approaches to community development followed in this project, including stakeholder identification and engagement activities. A systematic approach was taken to plan, implement, and manage the community and stakeholder engagement activities for the project. As much as possible, our goal was to design activities that are participatory, inclusive, and strategic, to ensure that the scientific outputs of the project are translated into tangible benefits for stakeholders; in turn, we sought to ensure stakeholder input enriched our research process. Our community engagement strategy was built upon the following core principles:

1. **Inclusivity and equity**: Deliberate effort was made to identify and involve a diverse range of stakeholders, from government regulators to local civil society groups, ensuring multiple perspectives were heard.
2. **Transparency and trust**: All communications were open and honest about the project’s capabilities, limitations, and objectives. We aimed to build trust by being reliable partners.
3. **Mutual benefit**: Engagement was designed not as a one-way transfer of information, but as a dialogue where both the project team and stakeholders could learn and benefit.
4. **Capacity development**: A primary goal was to move beyond awareness-raising to actively build skills and knowledge that would remain within the stakeholder organisations after the project is concluded.

5.2.1 Stakeholder identification and analysis

We identified several stakeholders, categorised them, and analysed the best engagement approach. The analysis was based on two primary criteria, i.e., their influence over water management and policy, and their interest or impact regarding water quality in the COHWHS. This resulted in four key stakeholder categories (influence-interest matrix) as listed in **Table 5-1**.

Table 5-1. Identification of stakeholders and analysis

Stakeholder segments	Key organisations	Rationale for engagement
Category 1 stakeholders (high influence, high interest)	• COHWHS Management Authority • Department of Economic Development • MCLM • DWS	These are regulatory authorities or have a direct management responsibility for COHWHS. Engaging these stakeholders is critical for the adoption of Earth Observation-based water quality research and products and their integration into the current suite of monitoring efforts and formulation policy activities.
Category 2 stakeholders (high influence, lower interest)	• Percy Stewart WWTW • Mining companies • Dept. of Mineral Resources & Energy	These stakeholders have a significant potential impact on water quality through their operations, and they have regulatory power. It is important to keep them informed of the project findings to promote accountability and influence corrective actions.
Category 3 stakeholders (lower influence, high interest)	• Civil society (e.g., NGOs) • Local communities • Landowners	These stakeholders are directly affected by water quality in the COHWHS, and they are potential advocates and valuable sources of ground-level information.
Category 4 stakeholders (lower influence, low interest)	• Other academic and research community members • General public	Other academic and research institutions beyond the project partners should be engaged through broad and low-intensity communication. These are potential future collaborators. The general public beyond the COHWHS area is also engaged for awareness purposes and for broader public support.

5.2.2 Community engagement activities

Activities for engaging with the community were planned, some of which have been completed, while others are ongoing until the project is concluded (See **Table 5-2**).

Table 5-2. Planned community engagement activities and the respective target stakeholders

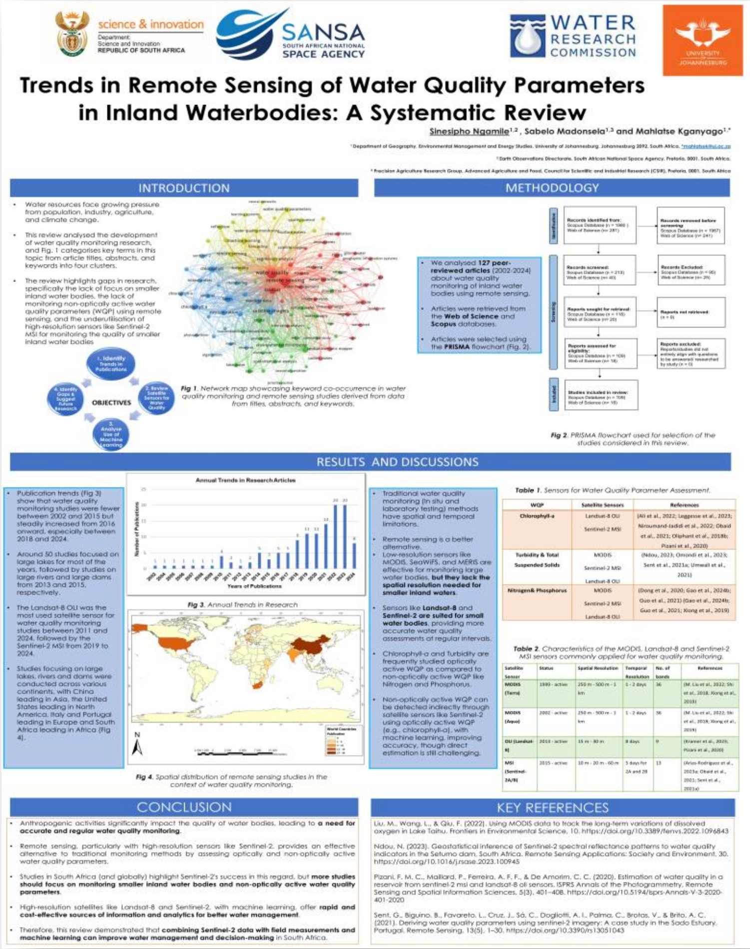
Activity	Target stakeholder	Status
Stakeholder engagement workshop	COHWHS Management Authority, Department of Economic Development, MCLM, landowners, DWS	Aug 2025 (Completed)
Simplified videos	General public	Aug 2025 (Ongoing)
Career days/science communication	School learners, general public	Aug 2025 (Completed)
Op-eds (The Conversation, WRC and internal newsletters)	General public	Oct 2025 (Completed)
Public lectures	University students, academics, general public	Sept 2025 (Completed)
Conferences	University students, academics, governments, general public	Aug/Sept 2025 (Completed)
Training on remote sensing of water quality	DWS	April/may 2026 (To be completed)

5.3 ENGAGEMENT ACTIVITIES AND OUTCOMES

This section presents the different initiatives undertaken as part of the community engagement of the project. The students were supported to attend and present at conferences, a stakeholder workshop was held with various stakeholders of the COHWHS, and various public awareness activities were undertaken.

5.3.1 Conferences

Ms Sinesipho Ngamile presented a systematic literature review in the form of a poster titled “Trends in Remote Sensing of Water Quality Parameters in Inland Waterbodies. A Systematic Literature Review” at the 38th Annual Conference of the South African Society for Atmospheric Sciences (SASAS) at the Potchefstroom Campus of the North-West University (NWU) from 29 to 30 October 2024 (See **Figure 5-1**).



science & innovation
SANSA
SOUTH AFRICAN NATIONAL SPACE AGENCY
WATER RESEARCH COMMISSION
Trends in Remote Sensing of Water Quality Parameters in Inland Waterbodies: A Systematic Review
Sinesipho Ngamile, Sabelo Madonsela and Mahlatse Kganyago
INTRODUCTION
METHODOLOGY
OBJECTIVES
RESULTS AND DISCUSSIONS
Annual Trends in Research Articles
Table 1. Sensors for Water Quality Parameter Assessment.
Table 2. Characteristics of the MODIS, Landsat-8 and Sentinel-2 MSI sensors commonly applied for water quality monitoring.
CONCLUSION
KEY REFERENCES

Figure 5-1. Ms Sinesipho Ngamile presented a poster at the 38th Annual SASAS conference at North-West University (NWU)

At the 3rd International Conference on Earth and Environmental Sciences (ICEES), held at the Durban International Convention Centre from 17 to 20 August 2025, two project-affiliated MSc students, Ms Elizabeth Rathupetsane and Ms Sinesipho Ngamile, presented their research findings. Ms Ngamile delivered a presentation titled “Determining the Sensitivity of Sentinel-2 Bands to Water Quality Parameters Under High- and Low-Flow Conditions in the COHWHS, South Africa” (Paper No. 41) (See **Figure 5-2**). Her work garnered significant attention, stimulating a substantive technical question-and-answer session. Attendees notably recognised the potential application of her methodology for systematic water quality monitoring in other African regions that lack robust in situ monitoring infrastructure. Ms Rathupetsane subsequently presented her research, “Comparing Single- and Multi-Output Machine Learning Approaches for Assessment of Optically Active Water Quality Parameters in the Cradle of Humankind” (Paper No. 156). This presentation also generated notable interest, leading to further detailed discussions with participants following the formal session.

Figure 5-2. Ms Elizabeth Rathupetsane (Left) and Ms Sinesipho Ngamile (Right) presenting their master's work at the ICEES conference

Ms. Rathupetsane further disseminated the project's findings at the Society of South African Geographers (SSAG) conference, hosted by the University of Johannesburg (UJ) from 3 to 5 September 2025 (See **Figure 5-3**). The conference served as a key forum for interdisciplinary dialogue, bringing together students and scholars across diverse fields in Geography and Environmental Science.

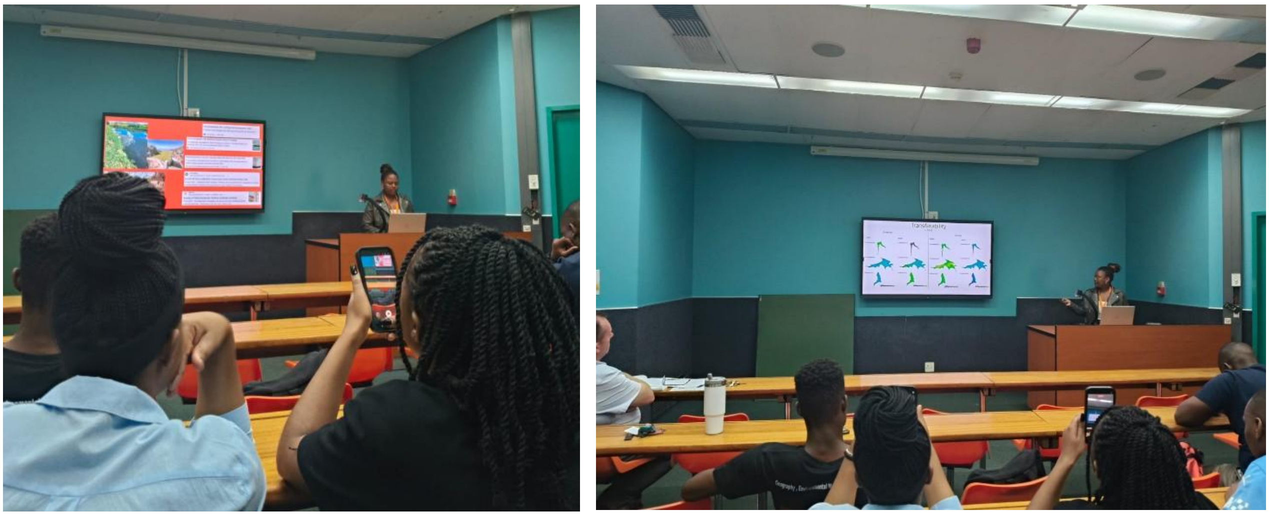


Figure 5-3. Ms Elizabeth Rathupetsane presenting her master's work at the SSAG student conference

Ms. Rathupetsane also presented her MSc research at the Gauteng Environmental Research Symposium (GERS), held from 16 to 17 September 2025 (See **Figure 5-4**). The symposium, hosted by the Gauteng Department of Environment in partnership with the UJ Process, Energy and Environmental Technology Station, was convened under the theme *Innovating Environmental Governance and Research for a Resilient, Sustainable and Inclusive Future.* This theme directly aligns with the project's focus on addressing critical environmental challenges in Gauteng, such as those arising from rapid urbanisation, industrialisation, and climate change. The forum successfully brought together government, academia, and researchers to advance sustainable environmental practices. In a significant achievement, Ms. Rathupetsane's research was awarded "Best Research" in the student track, recognising the exceptional quality and relevance of her contribution to the field.

Figure 5-4. Ms Elizabeth Rathupetsane (Master's student) presenting her work at the Gauteng Environmental Research Symposium (GERS) (left) and receiving a certificate for best student research (right)

5.3.2 Stakeholder workshop

Dr. Mahlatse Kganyago, Ms. Vuyelwa Mvandaba and Ms. Elizabeth Rathupetsane represented the project at a dedicated stakeholder workshop focused on the COHWHS. The hybrid event, co-convened by the CSIR and the Gauteng Provincial Government, was attended by approximately 27 in-person and 15 online participants from key stakeholder organisations.

5.3.2.1 Summary of presentations delivered

- The workshop featured a series of presentations that established a comprehensive context of the challenges and ongoing efforts in the region:
- **COHWHS Management Authority** opened the proceedings, outlining the mandate to protect the Outstanding Universal Value of the World Heritage Site. She emphasised that water degradation

poses a direct risk to the World Heritage Site status of the COHWHS, tourism, and public health, and detailed coordinated monitoring and remediation actions with partners, including the DWS and MCLM.

- **CSIR** presented progress on the long-term water resources monitoring programme. Key concerns identified were AMD from the Western Basin and municipal effluent from the Percy Stewart WWTW. While water clarity has improved in some areas, microbiological quality remains poor, with persistent *E. coli* contamination. She also highlighted challenges, including data gaps from reduced DWS monitoring and budget constraints.
- **COHWHS Association** provided a community perspective, cataloguing pollution from mine water, sewage infrastructure failures, and illegal dumping. He presented a timeline of escalating concerns since 1999 and stressed the urgent need for coordinated intervention, monitoring, and enforcement to address the risks of sinkhole formation and contaminated water supplies.
- **Federation for a Sustainable Environment (FSE)** focused on the impacts of mine water pollution, characterised by low pH, high salinity, and heavy metals. She noted the issue has gained international attention and highlighted the federation's work in community advocacy and workshops on environmental rights.
- **Eco Tabs Gauteng Pty Ltd** introduced his company's bioremediation methods as a cost-effective and ecological alternative to chemical treatments. He reported observable water quality improvements downstream following a pilot intervention in the Bloubank River.
- **Office of the Member of the Executive Council (MCLM)** detailed current and planned interventions for the Percy Stewart WWTW. She noted the R170 million total requirement for full refurbishment and reported that R52.5 million was secured for the 2025/26 financial year. Key steps include emergency repairs, staff appointments, and comprehensive training programmes to strengthen operational capacity.

5.3.2.2 Project presentation and stakeholder dialogue

Dr. Kganyago then introduced the project's research objectives and key achievements, focusing on the contribution of EO technologies. This presentation stimulated a robust and productive dialogue. A strong consensus emerged on the need for a synergistic monitoring ecosystem that integrates Internet of Things (IoT) devices, satellite remote sensing, and in situ measurements. Participants emphasised the critical value of obtaining rapid water quality insights for proactive management (See **Figure 5-5**).

Key discussion points elaborated these themes:

- FSE stressed that monitoring must lead to concrete action and rapid response, a role where technology can be pivotal.
- One participant raised concerns about inconsistent enforcement, particularly against public entities, questioning who bears the cost when enforcement fails.
- Another participant inquired about educational programmes for landowners and schools to clarify responsibilities and raise environmental awareness.
- A serious concern reiterated throughout was the persistent threat of the COHWHS losing its World Heritage status if tangible improvements are not demonstrated.

In his concluding remarks, Dr Kganyago synthesised a key recommendation from the discussions: the need for the standardisation of in situ data collection protocols. He stressed that such standardisation is paramount to ensuring that ground-based measurements are directly comparable and usable for validating and calibrating satellite remote sensing models.

Figure 5-5. Dr Mahlatse Kganyago (Project Lead) presenting at the stakeholder workshop

Furthermore, a dedicated video summarising the application of remote sensing as a complementary tool to in situ monitoring was screened during the workshop and subsequently shared with all participants. The resource garnered immediate practical interest, with one participant from The Cradle Nature Reserve (and involved in science engagement) formally requesting the presentation and video materials for use in their educational outreach with school learners. This request demonstrates a direct pathway for expanding the project's reach, effectively building awareness and promoting an understanding of geospatial technology among younger audiences beyond the project's immediate stakeholder circle.

5.3.3 Public engagements and awareness

The project team maintained an active presence on relevant digital platforms to communicate project milestones, including fieldwork progress, publication alerts, and student achievements (See **Figure 5-6**). These strategic updates successfully generated significant online engagement, reaching thousands of viewers and serving to broaden public awareness of the project, stimulate interest in its findings, and inspire a wider audience in the fields of environmental science and remote sensing.

The project's outreach effort was further strengthened when Ms Elizabeth Rathupetsane (research student) was invited to present at a webinar organised by AfriGEO on 16 September 2025. This event, which featured young professionals from across the continent, provided a platform to share the project's methodologies and preliminary findings with a diverse and engaged audience of EO enthusiasts from Africa. In a related achievement, the project's leadership was recognised through the nomination of Dr Mahlatse Kganyago as a finalist for an emerging researcher award, highlighting the WRC's contribution to developing scientific excellence.

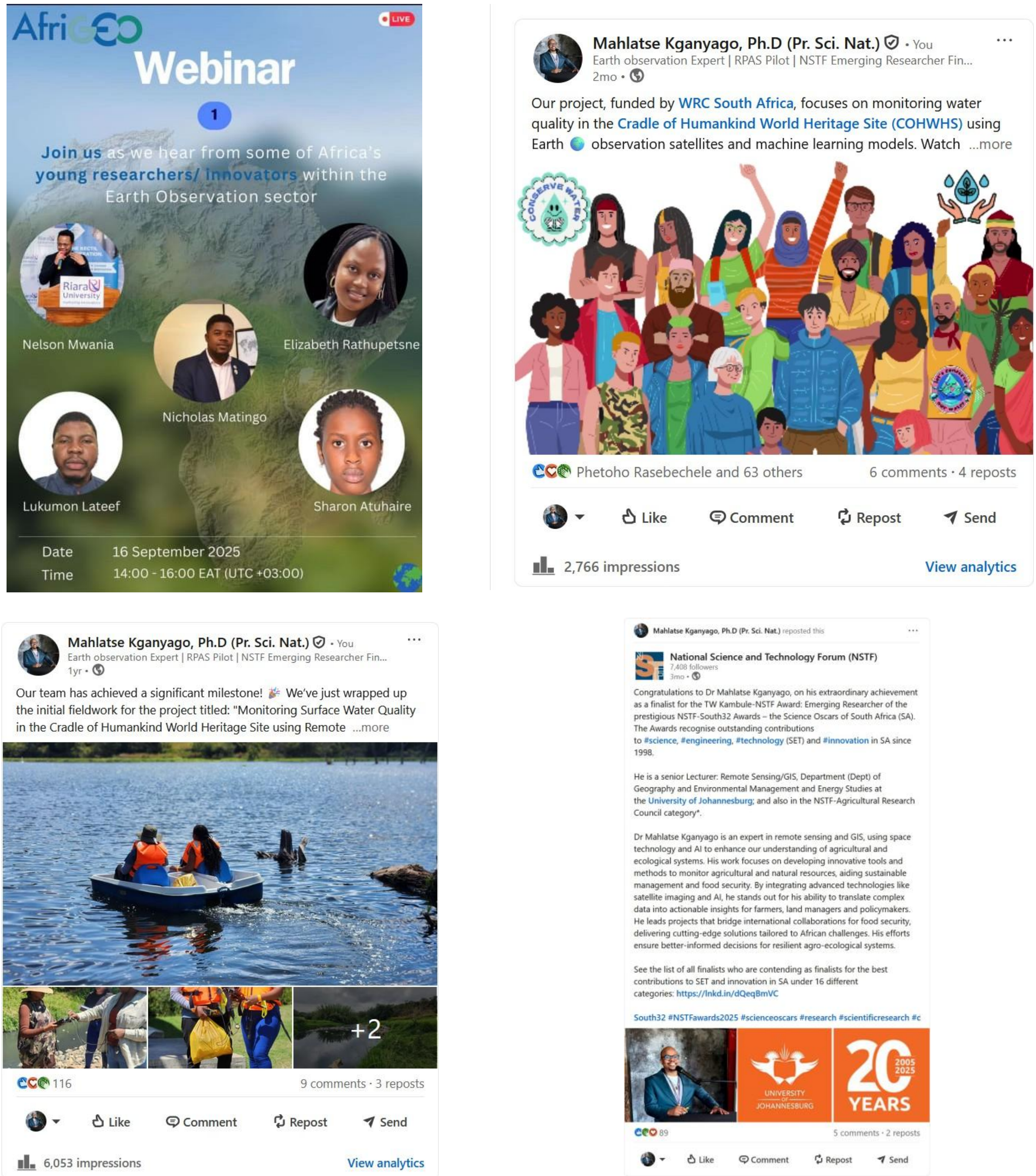


Figure 5-6. A poster of the AfriGEO Webinar where Ms Elizabeth Rathupetsane presented her master's work (top-left), a post of an awareness video (top-right), a post of the fieldwork (bottom-left) and a post of the nomination of Dr Kganyago as an emerging researcher (bottom-right)

5.3.4 Knowledge dissemination

The project has successfully disseminated its scientific findings through international peer-reviewed journals (See **Figure 5-7**). To date, two articles have been published in reputable international journals, with several additional manuscripts currently under review. Although only recently published, these works have demonstrated notable early traction and global reach. Collectively, they have been downloaded over 1 500 times and viewed approximately 600 times (as of the date of this report). This significant level of early engagement indicates a strong interest in the research and that it is contributing valuable knowledge to the international scientific community on the application of remote sensing for water quality monitoring.

frontiers | Frontiers in Remote Sensing

TYPE Original Research
PUBLISHED 21 July 2025
DOI 10.3389/frsen.2025.1631403

OPEN ACCESS

EDITED BY
Xiaojun Li,
INRAE Nouvelle-Aquitaine Bordeaux, France

REVIEWED BY
Pedzisai Kowe,
Midlands State University, Zimbabwe
Liqiao Tian,
Wuhan University, China
Tarek Seleem,
Suez Canal University, Egypt

*CORRESPONDENCE
Mahlatse Kganyago,
mahlatsek@uj.ac.za

RECEIVED 19 May 2025
ACCEPTED 03 July 2025
PUBLISHED 21 July 2025

CITATION
Ngamile S, Kganyago M, Madonsela S and 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.* 6:1631403. doi: 10.3389/frsen.2025.1631403

COPYRIGHT
© 2025 Ngamile, Kganyago, Madonsela and Mvandaba. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Characterising the spatio-temporal patterns of water quality parameters in the cradle of humankind world heritage site using Sentinel-2 and random forest regressor

Sinesipho Ngamile[1,2], Mahlatse Kganyago[1]*, Sabelo Madonsela[1,3] and Vuyelwa Mvandaba[4]

[1]Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg, South Africa, [2]Earth Observations Directorate, South African National Space Agency, Pretoria, South Africa, [3]Precision Agriculture Research Group, Advanced Agriculture and Food, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa, [4]Hydrology Research Group, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa

Introduction: Water quality assessment is essential for monitoring and managing freshwater resources, particularly in ecologically and culturally significant areas like the Cradle of Humankind World Heritage Site (COHWHS). This study aimed to predict and map the spatio-temporal patterns of both optically and non-optically active water quality parameters within small inland water bodies located in the COHWHS.

Methods: High-resolution Sentinel-2 Multispectral Instrument (MSI) satellite data and two random forest models (Model 1 [consisting of sensitive spectral bands] and Model 2 [consisting of spectral bands + indices]) were used alongside *In-situ* measurements of chlorophyll-a, suspended solids, dissolved oxygen (DO), pH, Temperature, and electrical conductivity (EC) were integrated to establish empirical relationships and assess spatial variability across high-flow and low-flow conditions.

Results: The results indicated that DO could be predicted with the highest accuracy under low-flow conditions, followed by EC. Specifically, Model 2 achieved an R^2 of 0.88 and an RMSE of 1.37 for DO, while Model 1 achieved an R^2 of 0.63 and an RMSE of 291.48 for EC. For optically active parameters, suspended solids showed the highest prediction accuracy under high-flow conditions using Model 2 (R^2p = 0.55; RMSE = 118.19). Due to the over-pixelation of other smaller water bodies within the COHWHS in Sentinel-2 imagery, Cradlemoon Lake was selected to show distinct seasonal (high- and low-flow) and spatial variations in optically and non-optically active water quality parameters.

Discussion: Variations in the results were influenced by runoff dynamics and upstream pollution: lower Temperatures and suspended solids under low-flow conditions increased DO concentrations, whereas higher suspended solid concentrations under high-flow conditions likely reduced light penetration,

frontiers | Frontiers in Environmental Science

TYPE Review
PUBLISHED 03 March 2025
DOI 10.3389/fenvs.2025.1549301

OPEN ACCESS

EDITED BY
Yong Q. Tian,
Central Michigan University, United States

REVIEWED BY
Pedzisai Kowe,
Midlands State University, Zimbabwe
Biao Xiong,
China Three Gorges University, China

*CORRESPONDENCE
Mahlatse Kganyago,
mahlatsek@uj.ac.za

RECEIVED 20 December 2024
ACCEPTED 10 February 2025
PUBLISHED 03 March 2025

CITATION
Ngamile S, Madonsela S and Kganyago M (2025) Trends in remote sensing of water quality parameters in inland water bodies: a systematic review. *Front. Environ. Sci.* 13:1549301. doi: 10.3389/fenvs.2025.1549301

COPYRIGHT
© 2025 Ngamile, Madonsela and Kganyago. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Trends in remote sensing of water quality parameters in inland water bodies: a systematic review

Sinesipho Ngamile[1,2], Sabelo Madonsela[1,3] and Mahlatse Kganyago[1]*

[1]Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg, South Africa, [2]Earth Observations Directorate, South African National Space Agency, Pretoria, South Africa, [3]Precision Agriculture Research Group, Advanced Agriculture and Food, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa

Monitoring water quality is crucial for sustainable water management and meeting the United Nations Sustainable Development Goals. Urbanisation, agricultural practices, industrial activities, and population growth increase the presence of biological, chemical and physical properties in water bodies. Traditional water quality monitoring methods (laboratory and *in situ* measurements) are limited spatially, temporarily and are costly. Satellite remote sensing has been shown to provide a systematic, cost-effective, and near-real-time alternative. This paper analysed 142 peer-reviewed articles published between 2002 and 2024 from Web of Science and Scopus databases. The final included articles in the review were achieved through the PRISMA flowchart. The review revealed that low-resolution sensors with long-term records, such as MODIS, were commonly applied to study large lakes. In contrast, sensors such as Landsat-8 and Sentinel-2 were applied for both lakes and dams. These sensors contain necessary spectral regions for monitoring water quality, where it was shown that the 500–600 nm region is critical for chlorophyll assessment, while the 640–670 nm region is used for turbidity. The Secchi disk depth and the total suspended solids were assessed using regions 860–1040 nm and 1570–1650 nm. Water quality research also focused on countries such as China, India, Brazil, and South Africa, with an emphasis on optically active parameters. There is, however, limited research on non-optically active parameters, such as nitrogen, phosphorus, and temperature, especially in small inland water bodies. Therefore, there is a need for more research in these areas, using direct and indirect methods of water quality parameter estimation with the integration of machine learning algorithms.

KEYWORDS
sustainable development goals, chlorophyll-a, total suspended solids, nitrogen, temperature, water contamination, biochemical and biophysical properties, machine learning algorithms

Figure 5-7. Publications in open journals

5.3.5 Capacity building

The project incorporated a vital capacity-development component through the recruitment of two Master of Science by Research candidates**.** Ms Sinesipho Ngamile was appointed in 2023 and has subsequently completed her studies (**Figure 5-8**), graduating with distinction (*cum laude*). Her research yielded two peer-reviewed academic publications in international outlets. Ms Elizabeth Rathupetsane, appointed in 2024, is also passed her dissertation with distinction (*cum laude*). Both students have been integral to the project's success, making significant contributions to its core scientific deliverables and actively participating in all community engagement activities.

Figure 5-8. Ms Sinesipho Ngamile graduating at the University of Johannesburg (UJ) on 3 October 2025

5.4 SUMMARY

The planned community engagement activities for the project *Monitoring Surface Water Quality in the Cradle of Humankind World Heritage Site using Remote Sensing Technology* have largely been successfully achieved, with the main objective of fostering collaboration, building capacity, and ensuring the project's scientific outputs are accessible and relevant to a broad range of stakeholders. The multi-faceted approach, encompassing academic training, multi-stakeholder dialogue, targeted skills transfer, and public communication, has clearly extended the project's impact beyond academic circles.

The engagement activities have proven highly effective. The commendable performance of our MSc students, evidenced by their successful presentations at national and international conferences, the award of "Best Research" to Ms Rathupetsane, and both students passing of their MSc degrees with Distinction, demonstrates the project's success in developing high-calibre local expertise. The stakeholder workshop has, and planned DWS training will, effectively bridge the gap between cutting-edge science and practical application, promoting a shared understanding of the value of integrated monitoring systems and empowering key institutions with new tools and perspectives. Furthermore, the significant early engagement with peer- reviewed publications and the proactive use of digital and media platforms have amplified the project's reach, building awareness and inspiring a new generation of scientists and environmental stewards.

CHAPTER 6: CONCLUSIONS & RECOMMENDATIONS

6.1 CONCLUSIONS

The project Monitoring Surface Water Quality in the Cradle of Humankind World Heritage Site Using Remote Sensing Technology investigated the utility of high-resolution Sentinel-2 data for monitoring water quality parameters across winter and summer seasons in the rivers and dams in and around COHWHS. The primary objective of the investigation was to conduct empirical analysis of ground truth data together with Sentinel-2 data to develop remote sensing models that can be used for estimating water quality parameters across different seasons. The study demonstrated the sensitivity of Sentinel-2 data to different water quality parameters including Chl-*a*, TSS, DO, pH, EC and temperature. However, this sensitivity declines with a change in season. For instance, during the summer period, Sentinel-2 bands had a higher relationship with DO and pH and this relationship deteriorated during the winter period.

Moreover, the project demonstrated the transferability of some advanced machine learning models to other waterbodies. For instance, the multi-output GPR performed exceptionally well in predicting Chl-a and TSS in the neighbouring waterbodies. Multi-output GPR can account for the non-Gaussian nature of environmental data and that makes it especially effective in heterogeneous aquatic systems. The high performance of MGPR across both clear and turbid water bodies in the Cradle of Humankind highlights its robustness and indicates its suitability for deployment in other diverse environments. Overall, the results of the project demonstrate that Sentinel-2 data, when combined with machine learning, offers a scalable and timely solution for addressing the country's pressing water quality challenges, especially considering increased anthropogenic pressures and climate variability. Ultimately, the project presented scientific tests that should pave the way for innovative, technology-driven approaches towards monitoring water quality.

More importantly, the project has contributed to human capacity development in the water science space through formal training of two master's students i.e., Ms Sinesipho Ngamile and Ms Elizabeth Modjadji. In addition, the results of the project's scientific tests have been published in scientific journals and academic conferences as part of knowledge dissemination and public engagement activities. The project team has engaged with COHWHS Association, Mogale City leadership, COHWHS Management Authority, EcoTabs, African Potential, the Gauteng Department of Environment, and the University of the Witwatersrand and other stakeholders to contribute to the development of tools for water management in the province.

6.2 RECOMMENDATIONS

- Although promising results were obtained with both Chl-*a* and TSS, and spatial and temporal predictions were achieved, it remains unknown whether the models will perform as well in other water bodies. Therefore, it may be critical to apply the model in other water bodies discernible from freely available Sentinel-2 data.
- The multi-output algorithms used in this project have shown high accuracy and good transferability. Given the limited exploration of the algorithms, these should be further tested in different water bodies across South Africa and other regions to evaluate their transferability and generalisation capability.
- During the transfer learning process, some dams (such as Hartbeespoort Dam and Roodekopies Dam) were heavily infested by water hyacinth, leaving small portions of discernible water for analysis. Therefore, such effects were not removed, and Chl-*a* and total suspended solids models were applied in the whole dam. Future research should consider masking vegetation as a precursor step to modelling, especially if a considerable portion of water is not covered, allowing robust analysis.

- Machine learning algorithms tested here, showed varying performance by different water quality parameters. This indicates that there is no single algorithm that is universally applicable. Future research and operational efforts should explore ensemble models which takes the best predictions of each model into consideration. Based on the results here, the different machine learning algorithms can be applied for different water quality parameters.
- This study only assessed the performance of multi-output algorithms on Chl-*a* and TSS; however, there are more water quality parameters. Therefore, their performance should also be evaluated on other water quality parameters and extended to multi-year and seasonal scales to account for temporal variability. The transferability of the models should be evaluated with many field measurements.
- It is worth noting that the models developed here were based on limited data (i.e., 20 samples per season) and only two seasons (summer and winter) were analysed. Hence, due to the unstable nature of water and related chemical mixing ratios, influenced by various temporally varying factors, our results are only applicable to seasons when data was collected. To fully capture the trends in water quality parameters, various seasons should be considered in the future and the intensity of sampling could increase to a weekly/monthly basis. One way to achieve this is to explore IoT technology, where per minute water quality trends can be captured.
- Overall, this report demonstrates that the project has advanced the technical capabilities for monitoring the COHWHS and also cultivated a foundational network of informed and engaged stakeholders. The enthusiastic participation and continued interest from government, civil society, and the public confirm the critical need for this work and provide a strong foundation for its future application and development.

REFERENCES

ABBAS MM, MELESSE AM, SCINTO LJ and REHAGE JS (2019) Satellite estimation of chlorophyll-a using moderate resolution imaging spectroradiometer (MODIS) sensor in shallow coastal water bodies: Validation and improvement. *Water (Switzerland)* **11 (8)**. https://doiorg/103390/w11081621

ADJOVU GE, HAROON S, DAVID J and AHMAD S (2023) Overview of the application of remote sensing in effective monitoring of water quality parameters. Remote Sensing **15 (7)** 1-35.

AGGARWAL S (2004) Principles of remote sensing. In: *Satellite remote sensing and GIS applications in agricultural meteorology*, Vol. 23.

AGNOLI L, URQUHART E, GEORGANTZIS N, SCHAEFFER B, SIMMONS R, HOQUE B, NEELY MB, NEIL C, OLIVER J and TYLER A (2023) Perspectives on user engagement of satellite Earth observation for water quality management. Technological Forecasting and Social Change **189** 122357.

AHMED T, ZOUNEMAT-KERMANI M and SCHOLZ M (2020) Climate change, water quality and water-related challenges: A review with focus on Pakistan. International Journal of Environmental Research and Public Health **17 (22)** 1-22.

American Public Health Association, American Water Works Association, and Water Environment Federation (1992) Standard Methods for the Examination of Water and Wastewater (18th ed) Washington, DC: American Public Health Association.

ANSPER A and ALIKAS K (2019) Retrieval of chlorophyll a from Sentinel-2 MSI data for the European Union water framework directive reporting purposes. Remote Sensing **11(1)**. https://doiorg/103390/rs11010064

ARIAS-RODRIGUEZ LF, DUAN Z, SEPÚLVEDA R, MARTINEZ-MARTINEZ SI and DISSE M (2020) Monitoring Water Quality of Valle de Bravo Reservoir, Mexico, Using Entire Lifespan of MERIS Data and Machine Learning Approaches. Remote Sensing **12(10)** 1586. https://doiorg/103390/rs12101586

ARIAS-RODRIGUEZ LF, TÜZÜN UF, DUAN Z, HUANG J, TUO Y and DISSE M (2023) Global water quality of inland waters with Harmonized Landsat-8 and Sentinel-2 using cloud-computed machine learning. Remote Sensing **15 (5)**. https://doiorg/103390/rs15051390

ARORA M, MUDALIAR A and PATERIYA B (2022) Assessment and Monitoring of Optically Active Water Quality Parameters on Wetland Ecosystems Based on Remote Sensing Approach: A Case Study on Harike and Keshopur Wetland over Punjab Region, India. In: *The 9th International Electronic Conference on Sensors and Applications*, Vol 27, 1-9.

BANGIRA T, MATONGERA TN, MABHAUDHI T and MUTANGA O (2024) Remote Sensing-Based Water Quality Monitoring in African Reservoirs, Potential and Limitations of Sensors and Algorithms: A Systematic Review. Physics and Chemistry of the Earth **134** 103536. https://doi:101016/jpce2023103536

BARSI JA, SCHOTT JR, HOOK SJ, RAQUENO NG, MARKHAM BL and RADOCINSKI RG (2014) Landsat-8 thermal infrared sensor (TIRS) vicarious radiometric calibration. Remote Sensing **6 (11)** 11607-11626.

BHATTI AM, RUNDQUIST DC, NASU S and TAKAGI M (2008) Assessing the potential of remotely sensed data for water quality monitoring of coastal and inland waters.

BRESCIANI M, GIARDINO C, STROPPIANA D, DESSENA MA, BUSCARINU P, CABRAS L, SCHENK K, HEEGE T, BERNET H, BAZDANIS G and TZIMAS A (2019) Monitoring water quality in two dammed reservoirs from multispectral satellite data. European Journal of Remote Sensing **52 (4)** 113-122.

BREZONIK PL, OLMANSON LG, BAUER ME and KLOIBER SM (2007) Measuring water clarity and quality in Minnesota lakes and rivers: A census-based approach using remote-sensing techniques. Cura Reporter **37 (3)** 3-313.

BRIVIO PA, GIARDINO C and ZILIOLI E (2001) Determination of chlorophyll concentration changes in Lake Garda using an image-based radiative transfer code for Landsat TM images. International Journal of Remote Sensing **22 (2-3)** 487-502. https://doiorg/101080/014311601450059

BUCHANAN M (2010) The Karst System of the Cradle of Humankind World Heritage Site. The South African Karst Working Group.

BUI Q, JAMET C, VANTREPOTTE V, MÉRIAUX X, CAUVIN A and MOGRANE MA (2022) Evaluation of Sentinel-2/MSI atmospheric correction algorithms over two contrasted French coastal waters. Remote Sensing **14 (5)** 1099. https://doi:103390/rs14051099

BUMA WG and LEE SI (2020) Evaluation of Sentinel-2 and Landsat 8 images for estimating Chlorophyll-a concentrations in Lake Chad, Africa. Remote Sensing **12 (15)** 2437. https://doiorg/103390/rs12152437

CABALLERO I, ROMÁN A, TOVAR-SÁNCHEZ A and NAVARRO G (2022) Water quality monitoring with Sentinel-2 and Landsat-8 satellites during the 2021 volcanic eruption in La Palma (Canary Islands). Science of the Total Environment **822**. https://doiorg/101016/jscitotenv2022153433

CAMACHO F, FUSTER B, LI W, WEISS M, GANGULY S, LACAZE R and BARET F (2021) Crop Specific Algorithms Trained over Ground Measurements Provide the Best Performance for GAI and FAPAR Estimates from Landsat-8 Observations. Remote Sensing of Environment **260**. https://doi:101016/jrse2021112453

ČAMPULOVÁ M, ČAMPULA R and HOLEŠOVSKÝ J (2022) An R package for identification of outliers in environmental time series data. Environmental Modelling and Software **155** 105435. https://doiorg/101016/jenvsoft2022105435

CHAVULA G, BREZONIK P, THENKABAIL P, JOHNSON T and BAUER M (2009) Estimating chlorophyll concentration in Lake Malawi from MODIS satellite imagery. Physics and Chemistry of the Earth **34 (13-16)** 755-760. https://doiorg/101016/jpce200907015

CHEN C, SHI P and MAO Q (2003) Application of remote sensing techniques for monitoring the thermal pollution of cooling-water discharge from nuclear power plant. Journal of Environmental Science and Health **38 (8)** 1659-1668.

CHEN J, XU N, HE X, QUAN W, HE Q, HAN Q and PAN D (2022) Determining pseudo-invariant calibration sites for comparing inter-mission ocean color data. ISPRS Journal of Photogrammetry and Remote Sensing **192** 377-394. https://10.1016/j.isprsjprs.2022.08.004

COLVIN C, MURUVEN D, LINDLEY D, GORDON H and SCHACHTSCHNEIDER K (2016) *Water: Facts and futures* World Wildlife Fund Report, Cape Town, South Africa.

CONSTANTIN S, CONSTANTINESCU S and DOXARAN D (2017) Long-Term Analysis of Turbidity Patterns in Danube Delta Coastal Area Based on MODIS Satellite Data. Journal of Marine Systems **170** 10-21. https://doi:101016/jjmarsys201701016

DALL'OLMO G and GITELSON AA (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. Applied Optics **45 (15)** 3577. https://doiorg/101364/AO45003577

DE KEUKELAERE L, STERCKX S, ADRIAENSEN S, KNAEPS E, REUSEN I, GIARDINO C, BRESCIANI M, HUNTER P, NEIL C, VAN DER ZANDE D and VAICIUTE D (2018) Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. European Journal of Remote Sensing **51 (1)** 525-542.https://doiorg/101080/2279725420181457937

DOGLIOTTI AI, RUDDICK KG, NECHAD B, DOXARAN D and KNAEPS E (2015) A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters. Remote Sensing of Environment **156** 157-168. https://doiorg/101016/jrse201409020

DÖRNHÖFER K, GÖRITZ A, GEGE P, PFLUG B and OPPELT N (2016) Water Constituents and Water Depth Retrieval from Sentinel-2A-A First Evaluation in an Oligotrophic Lake. Remote Sensing **8 (11)** 941. https://doiorg/103390/rs8110941

DU PREEZ EA (2019) The contribution of geological features to visitor experiences: comparison between two geotourism attractions in South Africa. GeoJournal of Tourism and Geosites **26 (3)** 1006-1020. https://doi:1030892/gtg2

DUBE T, MUTANGA O, SEUTLOALI K, ADELABU S and SHOKO C (2015) Water quality monitoring in sub-Saharan African lakes: A review of remote sensing applications. African Journal of Aquatic Science **40 (1)** 1-7.

DWS SAMPLING GUIDE (2000) *Quality of Domestic Water Supplies Volume 2: Sampling Guide.* https://www.dws.gov.za/iwqs/AssessmentGuides/SamplingGuide/SamplingGuide.pdf (Accessed 01 November 2024)

DYBA K, ERMIDA S, PTAK M, PIEKARCZYK J and SOJKA M (2022) Evaluation of Methods for Estimating Lake Surface Water Temperature Using Landsat 8. Remote Sensing **14 (15)** 3839.

EVANS AEV, MATEO-SAGASTA J, QADIR M, BOELEE E and IPPOLITO A (2019) Agricultural Water Pollution: Key Knowledge Gaps and Research Needs. Current Opinion in Environmental Sustainability **36** 20-27. https://doi:101016/jcosust201810003

FU B, LAO Z, LIANG Y, SUN J, HE X, DENG T, HE W, FAN D, GAO E and HOU Q (2022) Evaluating optically and non-optically active water quality and its response relationship to hydro-meteorology using multi-source data in Poyang Lake, China. Ecological Indicators **(145)** 109675.

FULLER LM, AICHELE SS and MINNERIC RJ (2004) Predicting water quality by relating secchi-disk transparency and chlorophyll a measurements to satellite imagery for Michigan inland lakes. Research Report no. 2004-5086, US Geological Survey Scientific Investigations Report.

GARCÍA-NIETO PJ, GARCÍA-GONZALO E, ALONSO FERNÁNDEZ JR and DÍAZ MUÑIZ C (2020) A New Predictive Model for Evaluating Chlorophyll-a Concentration in Tanes Reservoir by Using a Gaussian Process Regression. Water Resources Management **34 (15)** 4921-4941. https://doiorg/101007/s11269-020-02699-x

GELSEY K, CHANG H and RAMIREZ D (2023) Effects of landscape characteristics, anthropogenic factors, and seasonality on water quality in Portland, Oregon. Environmental Monitoring Assessment **195 (219)**.

GHOLIZADEH M, MELESSE A and REDDI L (2016) A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. Sensors **16 (8)** 1298. https://doiorg/103390/s16081298

GIARDINO C, PEPE M, BRIVIO PA, GHEZZI P and ZILIOLI E (2001) Detecting chlorophyll, Secchi disk depth and surface temperature in a sub-alpine lake using Landsat imagery. Science of the Total Environment ***268* (1-3)** 19-29.

GITELSON AA, DALL'OLMO G, MOSES W, RUNDQUIST DC, BARROW T, FISHER TR, GURLIN D and HOLZ J (2008) A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. Remote Sensing of Environment **112 (9)** 3582-3593. https://doiorg/101016/jrse200804015

HASSAN N and WOO CS (2021) Machine Learning Application in Water Quality Using Satellite Data. In: *IOP Conference Series: Earth and Environmental Science*. IOP Publishing Ltd.

HOBBS PJ (2011) *Situation assessment of the surface water and groundwater resource environments in the Cradle of Humankind World Heritage Site*. Report prepared for the Management Authority Department of Economic Development, Gauteng Province, South Africa.

HOPPE H, DIETRICH P, MARZAHN P, WEIß T, NITZSCHE C, VON LUKAS FU, WENGEREK T and BORG E (2024) Transferability of Machine Learning Models for Crop Classification in Remote Sensing Imagery Using a New Test Methodology: A Study on Phenological, Temporal, and Spatial Influences. Remote Sensing **16 (9)** 1493. https://doiorg/103390/rs16091493

In: *Recent Trends in Wastewater Treatment*. Springer International Publishing. https://doiorg/101007/978-3-030-99858-5_1

IOCCG (2018) Earth Observations in Support of Global Water Quality Monitoring Greb, S, Dekker, A and Binding, C (eds). Research report no. 17, International Ocean Colour Coordinating Group, Dartmouth, Canada.

JAJI MO, BAMGBOSE O, ODUKOYA OO and AROWOLO TA (2007) Water quality assessment of Ogun river, South West Nigeria. Environmental Monitoring and Assessment **133 (1-3)** 473-482. https://doiorg/101007/s10661-006-9602-1

JI C, LI X, WEI H and LI S (2020) Comparison of Different Multispectral Sensors for Photosynthetic and Non-Photosynthetic Vegetation-Fraction Retrieval. Remote Sensing **12 (1)**. https://doiorg/103390/rs12010115

JOSHI A, PRADHAN B, GITE S and CHAKRABORTY S (2023) Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review. Remote Sensing **15 (8)**. https://doiorg/103390/rs15082014

JOSHI N, PARK J, ZHAO K, LONDO A and KHANAL S (2024) Monitoring Harmful Algal Blooms and Water Quality Using Sentinel-3 OLCI Satellite Imagery with Machine Learning. Remote Sensing **16 (13)** 2444. https://doiorg/103390/rs16132444

KARIMI B, HASHEMI SH and AGHIGHI H (2023) Development of the best retrieval models of non-optically active parameters for an artificial shallow lake by random forest algorithm. Remote Sensing Applications: Society and Environment **29**. https://doiorg/101016/jrsase2023100926

KNAEPS E, RUDDICK KG, DOXARAN D, DOGLIOTTI AI, NECHAD B, RAYMAEKERS D and STERCKX S (2015) A SWIR based algorithm to retrieve total suspended matter in extremely turbid waters. Remote Sensing of Environment **168** 66-79. https://doiorg/101016/jrse201506022

KUHN C, DE MATOS VA, WARD N, LOKEN L, SAWAKUCHI HO, KAMPEL M, RICHEY J, STADLER P, CRAWFORD J, STRIEGL R, VERMOTE E, PAHLEVAN N and BUTMAN D (2019) Performance of Landsat-8 and Sentinel-2 surface reflectance products for river remote sensing retrievals of chlorophyll-a and turbidity. Remote Sensing of Environment **224** 104-118. https://doiorg/101016/jrse201901023

KUPSSINSKÜ LS, GUIMARÃES TT, DE SOUZA EM, ZANOTTA DC, VERONEZ MR, GONZAGA L and MAUAD FF (2020) A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. Sensors (Switzerland) **20 (7)**. https://doiorg/103390/s20072125

KURIHARA J, TAKAHASHI Y, SAKAMOTO Y, KUWAHARA T and YOSHIDA K (2018) HPT: A high spatial resolution multispectral sensor for microsatellite remote sensing. Sensors (Switzerland) **18 (2)**. https://doiorg/103390/s18020619

LATHROP JR (1992) Landsat Thematic Mapper monitoring of turbid inland water quality. Photogrammetric Engineering and Remote Sensing **58**.

LEE Z, SHANG S, HU C, DU K, WEIDEMANN A, HOU W, LIN J and LIN G (2015) Secchi disk depth: A new theory and mechanistic model for underwater visibility. Remote sensing of environment **169** 139-149.

LEGGESSE ES, ZIMALE FA, SULTAN D, ENKU T, SRINIVASAN R and TILAHUN SA (2023) Predicting Optical Water Quality Indicators from Remote Sensing Using Machine Learning Algorithms in Tropical Highlands of Ethiopia. Hydrology **10 (5)**. https://doiorg/103390/hydrology10050110

LI L, GU M, GONG C, HU Y, WANG X, YANG Z and HE Z (2022) An advanced remote sensing retrieval method for urban non-optically active water quality parameters: An example from Shanghai. Science of The Total Environment **(880)** 163389.

LI S, QIAN Z and LIU J (2024) Multi-Output Regression Algorithm-Based Non-Dominated Sorting Genetic Algorithm II Optimization for L-Shaped Twisted Tape Insertions in Circular Heat Exchange Tubes. Energies **17 (4)**. https://doiorg/103390/en17040850

LIU H, LI Q, SHI T, HU S, WU G and ZHOU Q (2017) Application of Sentinel 2 MSI Images to Retrieve Suspended Particulate Matter Concentrations in Poyang Lake. Remote Sensing **9 (7)** 761. https://doi:103390/rs9070761

MAIER PM, KELLER S and HINZ S (2021) Deep learning with WASI simulation data for estimating chlorophyll a concentration of inland water bodies. Remote Sensing **13 (4)** 1127. https://doiorg/103390/rs13040718

MAJOZI NP, SALAMA MS, BERNARD S, HARPER DM and HABTE MG (2014) Remote sensing of euphotic depth in shallow tropical inland waters of Lake Naivasha using MERIS data. Remote Sensing of Environment **148** 178-189. https://doiorg/101016/jrse201403025

MALAHLELA OE, OLIPHANT T, TSOELENG LT and MHANGARA P (2018) Mapping Chlorophyll-a Concentrations in a Cyanobacteria- and Algae-Impacted Vaal Dam Using Landsat 8 OLI Data. South African Journal of Science **114 (9/10)**. https://doi:1017159/sajs2018/4841

MATTHEWS MW and STEWART B (2015) Eutrophication and Cyanobacteria in South Africa's Standing Water Bodies: A View from Space. South African Journal of Science **111 (5/6)** 1-8. https://doi:1017159/sajs2015/20140193

MISHRA S, and MISHRA DR (2012) Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. Remote Sensing of Environment **117** 394-406. https://doiorg/101016/jrse201110016

MOLEKOA MD, KUMAR P, CHOUDHARY BK, YUNUS AP, KHARRAZI A, KHEDHER KM and ALSHAYEB MJ (2022) Spatio-Temporal Variations in the Water Quality of the Doorndraai Dam, South Africa: An Assessment of Sustainable Water Resource Management. Current Research in Environmental Sustainability **4** 100187. https://doi:101016/jcrsust2022100187

MPAKAIRI KS, MUTHIVHI FF, DONDOFEMA F, MUNYAI LF and DALU T (2024) Chlorophyll-a unveiled: unlocking reservoir insights through remote sensing in a subtropical reservoir. Environmental Monitoring and Assessment **196(4)**. https://doiorg/101007/s10661-024-12554-w

MURRAY C, LARSON A, GOODWILL J, WANG Y, CARDACE D and AKANDA AS (2022) Water Quality Observations from Space: A Review of Critical Issues and Challenges. Environments **9(10)** 125.

MVANDABA V and SHADUNG J (2023) *Surface water and groundwater resources monitoring, Cradle of Humankind World Heritage Site, Gauteng Province, South Africa: Water resources status report for the period April 2023 to September 2023*. Council for Scientific and Industrial Research, Pretoria, South Africa.

MYINT SW and WALKER ND (2002) Quantification of surface suspended sediments along a river dominated coast with NOAA AVHRR and Sea WiFS measurements: Louisiana, USA. International Journal of Remote Sensing **23** 3229-3249. https://10.1080/01431160110104700

NECHAD B, RUDDICK KG and NEUKERMANS G (2009) *Calibration and validation of a generic multi-application: Society mapping of turbidity in coastal waters*. BOSTATER JR, MERTIKAS SP, NEYT X and M VELEZ-REYES M. https://doiorg/101117/12830700

NECHAD B, RUDDICK KG and PARK Y (2010) Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters. Remote Sensing of Environment **114 (4)** 854-866.

NEIL C, SPYRAKOS E, HUNTER PD and TYLER AN (2019) A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types. Remote Sensing of Environment **229** 159-178. https://doiorg/101016/jrse201904027

NGUYEN HQ, HA NT, NGUYEN NL and PHAM TL (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 Environment Research **93 (12)** 2941-2957. https://doiorg/101002/wer1643

NOVO EMM, HANSOM JD and CURRAN PJ (1989) The effect of viewing geometry and wavelength on the relationship between reflectance and suspended sediment concentration. International Journal of Remote Sensing **10 (8)** 1357-1372.

OBAID AA, ADAM EM, ALI AK and ABIYE TA (2024) Time Series Analysis of Water Quality Factors Enhancing Harmful Algal Blooms (HABs): A Study Integrating In-Situ and Satellite Data, Vaal Dam, South Africa. Water **16 (5)** 764. https://doi:103390/w16050764

OLIPHANT T, TSOELENG LT, MHANGARA P and MALAHLELA OE (2018) Mapping Chlorophyll-a Concentrations in a Cyanobacteria- and Algae-Impacted Vaal Dam Using Landsat 8 OLI Data. South African Journal of Science **114 (9-10)**. https://doi:1017159/sajs2018/4841

OLMANSON LG, BAUER ME and BREZONIK PL (2008) A 20-year Landsat water clarity census of Minnesota's 10 000 lakes. Remote Sensing of Environment **112 (11)** 4086-4097.

ORR HG, SIMPSON GL, DES CLERS S, WATTS G, HUGHES M, HANNAFORD J, DUNBAR MJ, LAIZÉ CLR, WILBY RL, BATTARBEE RW and EVANS R (2015) Detecting changing river temperatures in England and Wales. Hydrological Processes **29 (5)** 752-766.

PAHLEVAN N, SMITH B, SCHALLES J, BINDING C, CAO Z, MA R, ALIKAS K, KANGRO K, GURLIN D, HÀ N, OPPELT N and 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 Sensing of Environment **240**. https://doiorg/101016/jrse2019111604

PARRIS K (2011) Impact of agriculture on water pollution in OECD countries: Recent trends and future prospects. International Journal of Water Resources Development **27 (1)** 33-52. https://doiorg/101080/079006272010531898

PIZANI FM, MAILLARD P, FERREIRA AF and DE AMORIM CC (2020) Estimation of water quality in a reservoir from Sentinel-2 MSI and Landsat-8 OLI sensors. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **3** 401-408.

RASMUSSEN CE (2004) Gaussian Processes in Machine Learning Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3176. https://doi:101007/978-3-540-28650-9_4

REINART A and PIERSOON D (2005) *Water Quality Monitoring of Large European Lakes Using Meris Full Resolution Data.* http://envisatesaint/services/beam/

RITCHIE JC and SCHIEBE FR (1976) Remote sensing of suspended sediments in surface waters. Journal of American Society Photogrammetry and Remote Sensing **(42)** 1539-1545.

RODRÍGUEZ-LÓPEZ L, DURAN-LLACER I, ALVAREZ BL, LAMI A and URRUTIA R (2023) Recovery of Water Quality and Detection of Algal Blooms in Lake Villarrica through Landsat Satellite Images and Monitoring Data. Remote Sensing, **15 (7)**. https://doiorg/103390/rs15071929

RUBIN HJ, LUTZ DA, STEELE BG, COTTINGHAM KL, WEATHERS KC, DUCEY MJ, PALACE M, JOHNSON KM and CHIPMAN JW (2021) Remote sensing of lake water clarity: Performance and transferability of both historical algorithms and machine learning. Remote Sensing **13 (8)**. https://doiorg/103390/rs13081434

SAALIDONG BM, ARAM SA, OTU S and LARTEY PO (2022) Examining the dynamics of the relationship between water pH and other water quality parameters in ground and surface water systems. PLOS ONE **17 (1)** 1.

SABERIOON M, KHOSRAVI V, BROM J, GHOLIZADEH A and SEGL K (2023) Examining the sensitivity of simulated EnMAP data for estimating chlorophyll-a and total suspended solids in inland waters. Ecological Informatics **75**. https://doiorg/101016/jecoinf2023102058

SAGAN V, MAIMAITIJIANG M, SIDIKE P, EBLIMIT K, PETERSON KT, HARTLING S, ESPOSITO F, KHANAL K, NEWCOMB MA and PAULI D (2019) UAV-Based High Resolution Thermal Imaging for Vegetation Monitoring, and Plant Phenotyping Using ICI 8640 P, FLIR Vue Pro R 640, and thermoMap Cameras. Remote Sensing **11 (3)** 330.

SENT G, BIGUINO B, FAVARETO L, CRUZ J, SÁ C, DOGLIOTTI AI, PALMA C, BROTAS V and BRITO AC (2021) Deriving water quality parameters using sentinel-2 imagery: A case study in the Sado Estuary, Portugal. Remote Sensing **13 (5)** 1-30.

SHADRIN D, NIKITIN A, TREGUBOVA P, TEREKHOVA V, JANA R, MATVEEV S and PUKALCHIK M (2021) An Automated Approach to Groundwater Quality Monitoring-Geospatial Mapping Based on Combined Application of Gaussian Process Regression and Bayesian Information Criterion. Water **13 (4)** 400. https://doi:103390/w13040400

SHAHZAD MI, MERAJ M, NAZEER M, ZIA I, INAM A, MEHMOOD K and ZAFAR H (2018) Empirical estimation of suspended solids concentration in the Indus Delta Region using Landsat-7 ETM+ imagery. Journal of Environmental Management **209** 254-261. https://doiorg/101016/jjenvman201712070

SHEN LQ, AMATULLI G, SETHI T, RAYMOND P and DOMISCH S (2020) Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Scientific Data **7 (1)** 161. https://doiorg/101038/s41597-020-0478-7

SHI K, ZHANG Y, ZHU G, LIU X, ZHOU Y, XU H, QIN B, LIU G and LI Y (2015) Long-term remote monitoring of total suspended matter concentration in Lake Taihu using 250m MODIS-Aqua data. Remote Sensing of Environment **164** 43-56. https://doiorg/101016/jrse201502029

SINGH MR and GUPTA A (2016) *Water pollution-sources, effects control*. HTTPS://WWWRESEARCHGATENET/PUBLICATION/321289637

SMANIA G and JONSSON EN (2021) Conditional distribution modeling as an alternative method for covariates simulation: Comparison with joint multivariate normal and bootstrap techniques. CPT: Pharmacometrics and Systems Pharmacology **10 (4)** 330-339. https://doiorg/101002/psp412613

SWAIN R and SAHOO B (2017) Improving river water quality monitoring using satellite data products and a genetic algorithm processing approach. Sustainability of Water Quality and Ecology **9-10** 88-114. https://doiorg/101016/jswaqe201709001

TAKIAR R (2023) A new method to identify the outliers based on the inter quartile range. https://doiorg/1033329/bomsr114103

TEUTONICO D, MUSUAMBA F, MAAS HJ, FACIUS A, YANG S, DANHOF M and DELLA PO (2015) Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques. Pharmaceutical Research **32 (10)** 3228-3237. https://doiorg/101007/s11095-015-1699-x

TIAN S, GUO H, HUANG JJ, ZHU X and ZHANG Z (2022) Comprehensive comparison performances of Landsat-8 atmospheric correction methods for inland and coastal waters. Geocarto International **37 (27)** 15302-15323. https://doiorg/101080/1010604920222097320

TOMING K, KUTSER T, LAAS A, SEPP M, PAAVEL B and NÕGES T (2016) First experiences in mapping lake water quality parameters with Sentinel-2 MSI Imagery. Remote Sensing **8 (8)** 640. https://doiorg/103390/rs8080640

TSAPANOU A, OIKONOMOU E, DRAKOPOULOS P, POULOS S and SYLAIOS G (2020) Coupling remote sensing data with in-situ optical measurements to estimate suspended particulate matter under the Evros river influence (North-East Aegean sea, Greece). International Journal of Remote Sensing **41 (6)** 2062-2080. https://doiorg/101080/0143116120191685713

USALI N and ISMAIL HM (2010) Use of Remote Sensing and GIS in Monitoring Water Quality. Journal of Sustainable Development **3 (3)**. www.ccsenet.org/jsd

VAN DEVENTER H and CHO MA (2014) Assessing leaf spectral properties of phragmites australis impacted by acid mine drainage. South African Journal of Science **110 (7-8)**. https://doiorg/101590/sajs2014/20130184

VAN NGUYEN M, LIN C-H, CHU H-J, MUHAMAD JL and SYARIZ AM (2019) Spectral Feature Selection Optimization for Water Quality Estimation International Journal of Environmental Research and Public Health **17 (1)** 272. https://doiorg/103390/ijerph17010272

VANHELLEMONT Q and RUDDICK K (2014) Turbid wakes associated with offshore wind turbines observed with Landsat 8. Remote Sensing of Environment **145** 105-115. https://doiorg/101016/jrse201401009

VANHELLEMONT Q and RUDDICK K (2021) Atmospheric correction of Sentinel-3/OLCI data for mapping of suspended particulate matter and chlorophyll-a concentration in Belgian turbid coastal waters. Remote Sensing of Environment **256** 112284. https://doiorg/101016/jrse2021112284

VARVIA P, RÄTY J and PACKALEN P (2023) MGPR: An R package for multivariate Gaussian process regression. SoftwareX **24** 101563. https://doiorg/101016/jsoftx2023101563

WANDA EMM, MAMBA BB and MSAGATI TAM (2016) Determination of the water quality index ratings of water in the Mpumalanga and North West provinces, South Africa. Physics and Chemistry of the Earth **(92)** 70-78.

WATANABE F, ALCANTARA E, RODRIGUE T, ROTTA L, BERNARDO N and IMAI N (2018) Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). Anais Da Academia Brasileira de Ciências **90 (2-1)** 1987-2000. https://doiorg/101590/0001-3765201720170125

WOLTERS E, TOTÉ C, STERCKX S, ADRIAENSEN S, HENOCQ C, BRUNIQUEL J, SCIFONI S and DRANSFELD S (2021) iCOR Atmospheric Correction on Sentinel-3/OLCI over Land: Intercomparison with AERONET, RadCalNet, and SYN Level-2. Remote Sensing **13 (4)** 654. https://doiorg/103390/rs13040654

XU K, LIN C, LEE SY, MAO L and MENG K (2022) Comparative analysis of complete Ilex (Aquifoliaceae) chloroplast genomes: insights into evolutionary dynamics and phylogenetic relationships. BMC Genomics **23 (1)**. https://doiorg/101186/s12864-022-08397-9

YANG H, KONG J, HU H, DU Y, GAO M and CHEN F (2022) A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges. Remote Sensing **14 (8)** 1770.

ZAINURIN SN, WAN ISMAIL WZ, MAHAMUD SNI, ISMAIL I, JAMALUDIN J, ARIFFIN KNZ and WAN KAMIL WM (2022) Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review. International Journal of Environmental Research and Public Health **19 (21)** 14080.

ZEHRA MS, ESRA E and GULSEN TK (2018) Multi-Output Regressions for Estimating Canola Biophysical Parameters From PolSAR Data. IEEE Geoscience and Remote Sensing Society.

ZHANG H, YAN D, ZHANG B, FU Z, LI B and ZHANG S (2022) An Operational Atmospheric Correction Framework for Multi-Source Medium-High-Resolution Remote Sensing Data of China. Remote Sensing **14 (21)**. https://doiorg/103390/rs14215590

ZHAO Y, YU T, HU B, ZHANG Z, LIU Y, LIU X, LIU H, LIU J, WANG X and SONG S (2022) Retrieval of Water Quality Parameters Based on Near-Surface Remote Sensing and Machine Learning Algorithm. Remote Sensing **14 (21)** 5305.

APPENDIX A: FIELD DATA SHEET

Project Name: MONITORING SURFACE WATER QUALITY USING REMOTE SENSING TECHNOLOGY (C2023/2024-01241)

General Information

Date & Time:______________________________ Site NO:______________________________

Spectrometer Operator (s): ______________________________ Photo NO:______________________________

GPS coordinates (Lat/Long):______________________________ River/Tributary: ______________________________

Environmental Information

Vegetation Type:____________________ Slope:____________________

Soil Type:____________________ Land use:____________________

Sun Azimuth: ____________________ Sun Zenith: ____________________

Other comments:____

Measurements	Spectra ID	Water Sample	DO	pH	Secchi disk depth	Turbidity (TSM)	Colour	Comments
1								
2								
3								
4								
5								