A REMOTE SENSING AND MACHINE LEARNING BASED FRAMEWORK FOR THE ASSESSMENT OF SPATIOTEMPORAL WATER QUALITY ALONG THE MIDDLE GANGA BASIN

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

ASHWITHA S K

DEPARTMENT OF WATER RESOURCES AND OCEAN ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE – 575025

JULY, 2023

A REMOTE SENSING AND MACHINE LEARNING BASED FRAMEWORK FOR THE ASSESSMENT OF SPATIOTEMPORAL WATER QUALITY ALONG THE MIDDLE GANGA BASIN

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

ASHWITHA S K

177018AM001

Under the guidance of

Dr. RAMESH H

Associate Professor,

Dept. of Water Resources and Ocean Engineering

NITK, Surathkal

DEPARTMENT OF WATER RESOURCES AND OCEAN ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE – 575025

JULY, 2023

DECLARATION

By the Ph.D. Research Scholar

I hereby *declare* that the Research Thesis entitled **A Remote Sensing and Machine Learning Based Framework for the Assessment of Spatiotemporal Water Quality Along the Middle Ganga Basin** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Water Resources Engineering** is a *bonafide report of the research work* carried out by me. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

177018 AM001, ASHWITHA S K

(Register Number, Name & Signature of the Research Scholar)

Department of Water Resources and Ocean Engineering

Place: NITK-Surathkal

Date: 17/7/2023

CERTIFICATE

This is to certify that the Research Thesis entitled **A Remote Sensing and Machine Learning Based Framework for the Assessment of Spatiotemporal Water Quality Along the Middle Ganga Basin** submitted by **ASHWITHA S K** (Register Number: 177018AM001) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

17/07/2023

Dr. Ramesh H.

Research Guide

Dr. RAMESH H.
Associate Professor
Dept. of Water Resources & Ocean Engineering
NITK-Surathkal, Mangaluru-575 025

(Name and Signature with Date and Seal)

Dept. of Water Resources & Ocean Engineering
NITK
Surathkal

Varija K
17.7.23

Chairman - DRPC

(Signature with Date and Seal)

Chairman (DRPC)
Dept. of Water Resources & Ocean Engineering

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, **Dr. Ramesh H**, Associate Professor, Department of Water Resources and Ocean Engineering, NITK, Surathkal, for his guidance, support, and encouragement throughout the process of researching and writing this thesis. His expertise, knowledge in the field and willingness to share them with me greatly contributed to the success of this project.

It gives me immense pleasure and innate satisfaction to express my deep gratitude and indebtedness to my former advisor **Late Prof. Paresh Chandra Deka,** to whom I owe what I have learnt in this research field. His expert guidance, training and encouragement motivated me during my early stage of work.

I wish to thank **Prof. Amai Mahesha, Prof. Amba Shetty, Prof. Dodamani B M** and **Dr. K. Varija** the successive Heads of the Department of Water Resources and Ocean Engineering, NITK, for the kind support and encouragement extended by them. I will always be grateful to all faculty members of the Department of Water Resources and Ocean Engineering, for their kind support and encouragement, throughout my tenure as a research scholar in the Department. I wish to place on record my sincere thanks to **Dr. Debabrata Karmakar**, Assistant Professor, Department of Water Resources and Ocean Engineering and **Dr. Rajasekaran C**, Assistant Professor, Department of Civil Engineering, the members of my Research Progress Assessment Committee, for their valuable suggestions and the encouragement provided at various stages of this work.

I sincerely thank all other teaching and nonteaching staff for their extended support and unrelenting efforts in structuring this research.

I was fortunate to have the encouragement and support from my dear friends **Parthasarathy, Tom, Dinu, Alka, Rony, Binoy, Amala, Chaithanya, Arya** and **Nithya**. Special thanks to **Mr Vineet Kumar Porwal**, M-Tech batch 2015-2017, for the help provided at various stages of this research.

Last but not least, I would like to thank my husband, **Krishnaraj Koduvally,** who has been a constant source of support from every angle and for his unwavering support and encouragement throughout my research.

Finally, words are insufficient to express my gratitude to my sweetest children, **Arundhati Raj** and **Vaidehi Raj**. My sister and brother **Shamitha Shibu** and **Shibu Narayanan**; parents **Satheesh Kotekani** and **Vanaja**, in-laws **Ramani** and **Damodaran Koduvally** for the love and support they have offered me during this journey.

I want to thank all those who have helped me in any way, whether they know it or not. This has indeed been a team effort, and I am grateful for the contributions of everyone involved.

Ashwitha S K

ABSTRACT

Understanding the changes in surface water quality over time and space necessitates an examination of spatiotemporal water quality data. This data can be used to identify pollution sources, monitor changes in water quality, and assess the effectiveness of management and conservation efforts. Furthermore, spatiotemporal surface water quality assessment can forecast future water quality trends, allowing for precise decision-making and conservation. Overall, spatiotemporal water quality assessment is critical in protecting and managing water resources.

Various multivariate statistical and machine learning techniques are used in this study to determine the river water quality status and comprehend the spatiotemporal pattern along the Middle Ganga Basin in Uttar Pradesh. The study was carried out for 14 years (2005-2018), with 20 Water Quality Parameters (WQPs) collected monthly and covering spatially from up-stream to downstream Ankinghat to Chopan respectively (20 monitoring stations under Central Water Commission, Middle Ganga Basin). The temporal dissimilarity of river water quality is established by applying the Spearman non-parametric correlation coefficient test (Spearman *r*). A significant *p*-level (0.0000) is observed for temperature within the season with a Spearman *r* of -0.866. Besides that, the parameters EC, pH, TDS, T, Ca, Cl, HCO_3, Mg, NO_2+NO_3, SiO_2, and DO strongly correlated with the season ($p < 0.05$). The K-means clustering algorithm temporarily classified the 20 monitoring stations into four clusters based on the similarity and dissimilarity of WQPs. Box and Whisker plots were generated based on these clusters to study water quality trends along individual clusters in different seasons. PCA was applied to screen out the most dominating WQPs causing spatial and seasonal variations from a large data set. Seasonally, the three PCs chosen explained 75.69% and 75% of the variance in the data. With PCs >0.70, the variables EC, pH, Temp, TDS, NO_2+NO_3, P-Tot, BOD, COD, and DO have been identified as the dominant pollutants. The applied RDA analysis revealed that LULC has a moderate to strong contribution to WQPs during the monsoon season but not during the non-monsoon season. Furthermore, dense vegetation is critical for keeping water clean, whereas agriculture, barren land and build-up area degrade water quality. Besides that, the findings suggest the relationship between WQPs and LULC differs at different spatial scales. The

stacked ensemble regression model is applied to understand the model's predictive power across different clusters and scales. Overall, the results indicate that the riparian scale is more predictive than a watershed and reach scales.

As a further part of this work, an integrated use of remote sensing, *insitu* measurements, and machine learning modelling is used better to understand the water quality status along the study region. In this context, a remote sensing framework based on the Extreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP) regressor with optimized hyperparameters to quantify the concentrations of different WQPs from the Landsat-8 satellite imagery is developed. Six years of satellite data from upstream to downstream Ankinghat to Chopan (20 stations under Central Water Commission (CWC), Middle Ganga Basin) are analysed to characterise the trends of dominant physicochemical WQPs across the four identified clusters. A significant coefficient of determination (R^2) in the range of 0.88- 0.98 for XGBoost and 0.72-0.97 for MLP was generated using the developed XGBoost and MLP regression models. The bands B1-B4 and their ratios are found to be more consistent with the WQPs. Meanwhile, the performance matrix RMSE for the parameters SiO_2 and DO for all clusters for the XGBoost method is determined to be superior to MLP. Indeed, these findings show that a small number of *insitu* measurements is sufficient to develop reliable models for estimating the spatiotemporal variations of physicochemical and biological WQPs. As a result, Landsat-8 models could aid in the environmental, economic, and social management of any body of water.

Keywords: Surface water quality, CA, PCA, LULC, Multi- spatial scale, RDA, SEM, RS of water quality, XGBoost, MLP.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	Definition
Adam	Adaptive learning rate optimization algorithm
ANN	Artificial Neural Network
ANOVA	Analysis of variance
AOP	Apparent optical properties
APIs.	Application Programming Interface
BOD	Biological Oxygen Demand
BPNN	Back-propagation Neural Network
CA	Cluster Analysis
Ca	Calcium
CART	Classification And Regression Tree
CASI	Compact Airborne Spectrographic Imager
CDOM	Coloured dissolved organic matter
Chl-a	Chlorophyll-a
Cl	Chloride
CNN	Convolutional Neural Networks
CO_3	Carbon Trioxide
COD	Chemical Oxygen Demand
CWC	Central Water Commission
DEM	Digital Elevation Mo
DO	Dissolved Oxygen
DT	Decision tree
EC	Electrical conductivity
ELR	Enhanced Logistic Regression
EM	Electromagnetic
ENVI	Environment for Visualising Images
ERT	Extremely Randomized Trees
ETM	Enhanced Thematic Mapper
F	Fluoride
FA	Factor analysis
FC	Faecal Coliforms
FLAASH	Fast Line-of-sight Atmospheric Analysis
GBM	Gradient Boosting Machine
GCPs	Ground Control Points
GEE	Google Earth Engine
GIS	Geographical Information System
GP	Genetic Programming
GRB	Ganga River Basin
HCO_3	Hydrogen Carbonate
HPO	Hyper parameter Optimization

IOP	Inherent Optical Properties
IR	Infra-Red
ISRO	Indian Space Research Organization
K	Potassium
lbfgs	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LDA	Linear Discriminant Analysis
LULC	Land Use Land Cover
Mg	Magnesium
MGB	Middle Ganga Basin
ML	Machine Learning
MLD	millions of liter per day
MLR	Multiple linear regression
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	Mean square error
MSME	Micro, Medium, and Small Enterprises
MSS	Multi-Spectral Scanner
MW	Microwave
Na	Sodium
NDCI	Normalized difference chlorophyll index
NDTI	Normalized difference turbidity index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NH_3-N	Ammonical Nitrogen
NIR	Near Infrared
NO_2+NO_3	Total Nitrogen
NPS	Non-Point Source
NRSC	National Remote Sensing Centre
OLI	Operational Land Imager
PCA	Principal Component Analysis
PCs	Principal Components
pDNN	Progressively decreasing deep neural network
pH	Potential of Hydrogen
PLS-SEM	Partial Least Squares Structural Equation Modelling
PSO	Particle Swam Optimization
P-Tot	total phosphorus
Q-Q plot	quantile-quantile plot
R^2	Coefficient of Determination
RDA	Redundancy Analysis
Relu	Rectified Linear Unit
RF	Random Forest
RMSE	Root Mean Square Error
ROI	Region Of Interest

RRMSE	Relative Root Mean Square Error
R*rs*	Remote sensing reflectance
SEM	stacked ensemble modelling
SGB	Stochastic Gradient Boosting
SiO_2	Silicon Dioxide
SO_4	Sulphate
SSR	Sum of Squared Regression
SST	Sum of Squared Total
STAR-FM	Spatial and Temporal Adaptive Reflectance Fusion Model
SVM	Support Vector Machine
SVR	Support Vector Regression
S-W test	Shapiro-Wilk
SWAT	Soil and Water Assessment Tool
SWIR	Short-wave infrared
T	Turbidity
TDS	Total dissolved solids
Temp	Temperature
TKN	Total Kjeldahl Nitrogen
TM	Thematic Mapper
TOA	Top of Atmosphere
TP	Total phosphorus
TS	Total Solids
TSS	Total Suspended Solids
UV	Ultra Violet
VIS	Visible
WQI	Water Quality Indices
WQPs	Water Quality Parameters
WSS	Within-sum-of-squares
WT	Water temperature
XGBoost	eXtreme Gradient Boosting

CHAPTER 1

INTRODUCTION

1.1 GENERAL

The sustainability of any natural resource necessitates a thorough understanding of the changing environment and socioeconomic issues. Investigating qualitative measurement is just as important as quantitative analysis for the long-term viability of water resources. Concerns about water quality are growing in Asia due to increased population and urbanization, which will aggravate the situation when combined with climate change. Water scarcity and water quality are expected to be the most severe barriers of the twenty-first century, particularly in developing countries. According to International Water Management Institute reports, approximately 30% of the world's population suffers from a lack of clean water. The quality of water in freshwater ecosystems is influenced by various natural and human-caused factors and can be highly complex. The specific characteristics of the catchment area and the impact of human activities play a significant role in determining surface water quality. Therefore, adequate water quality monitoring is critical for better water resource management programs. Acknowledging the concentration of different Water Quality Parameters (WQPs) present in any waterbody will provide quantitative information concerning water quality (Sudheer et al. 2007). Many researchers have discussed the influence of anthropogenic pollutants from rural (Tibebe et al. 2019), urban (Zhang et al. 2013; Miller and Hutchins 2017; Carstens and Amer 2019; Gu et al. 2019) and agricultural land use (Amato et al. 2018; Cheng et al. 2018). As a result, changes in land use could be directly linked to changes in water quality, i.e., the higher the percentage of agriculture and urban land, the higher the concentration of nitrate and phosphates in the freshwater system (Álvarez-cabria et al. 2016). Besides that, variations in precipitation, runoff, groundwater flow and interflow can all affect water quality. The orientation of a river and its associated drainage basin can also play a role, as can natural phenomena such as floods, droughts, and storms.

All these factors and human activities make water quality highly dynamic, varying throughout the year and across different locations. Maintaining water quality can be

challenging due to a variety of pollutants that can be present in freshwater ecosystems. Point source pollutants come from a specific, identifiable source, such as a factory or sewage treatment plant. Non-point source pollutants (NPS), on the other hand, are not from a specific, identifiable source but rather from diffuse sources such as agricultural runoff or urban storm water runoff (Shi et al. 2017). These NPS pollutants can be more challenging to control and mitigate because they are not coming from a single, identifiable location (Zhou et al. 2016). Both point source and non-point source pollutants can have negative effects on water quality, and both types of pollution need to be addressed to effectively maintain water quality. Further, the correlation between rainfall and landscape characteristics complicates NPS identification (Abdul-Aziz and Al-Amin 2016). Over the last few decades, there has been an increased demand for regular monitoring of rivers, which has led to the accumulation of a large amount of data on water quality and has raised the need for tools to process and analyse such massive amounts of data. Advances in computing technology have made it possible to analyse these large databases in ways that were previously impractical. This has allowed for the development of more sophisticated models and analytical tools that can help to better understand and predict water quality trends and patterns. This technology can also be used for real-time monitoring and early warning systems to identify and respond to potential water quality issues more quickly and effectively (Antonopoulos et al. 2001). Therefore, long-term data are required to address the state of the world's water resources. In contrast, field measurements for water quality measurements include expensive, time-consuming, labour-intensive sampling on-site and transport to land-based or shipboard laboratories for evaluation. Furthermore, there is a possibility of data compromise due to poor quality-control protocols, as well as quality assurance due to the extended holding of samples before analysis. As a result, improved comprehension of water quality spatiotemporal patterns at large scales is only possible using remote sensing techniques. To address the shortcomings of the traditional data collection method, the use of remote sensing data for water quality assessment has been investigated (Haji Gholizadeh et al. 2016). In other words, an integrated use of remote sensing, *insitu* measurements, and computer water quality modelling may result in a greater understanding of the water quality.

1.2 NATURAL AND ANTHROPOGENIC DRIVERS OF WATER QUALITY

Various natural and anthropogenic factors influence water quality in fluvial ecosystems. In the absence of human influences, water quality would be controlled solely by natural processes such as wind deposition of dust and salt, natural leaching of organic matter, nutrients from the soil, hydrological factors leading to runoff, and biological processes in the aquatic environment, which could result in changes in the physical and chemical composition of water. As an outcome, water in nature may comprise both dissolved and non-dissolved particulate matter. Human changes to catchments have changed the quantity, quality, and balance of natural sources and introduced new water flow paths such as irrigation transfers, dam flow releases, and point and diffuse sources. The natural flow and associated water quality characteristics are altered by anthropogenic activities such as river regulation, catchment land use, and water extraction. Some water quality indicators respond almost instantly to environmental changes, particularly flow or water volume changes. However, it can take decades in some cases, such as salinity responses to land clearing (Arora et al. 2017). Sources of pollution, categories and their influences are presented in Table 1.1.

Table 1.1 Sources, categories and factors influencing pollution

Main source	**Subcategories**	**Factors of influence**
Natural	Topography, e.g., Slope	Transportation of organic and inorganic compounds such as Phosphorus and Nitrogen.
	Orientation of the river	Controls the solar radiation over the water surface, the temperature of the river and the atmospheric temperature
	Precipitation	Controls the catchment runoff processes.
	Natural disasters (e.g., Drought, Floods, landslides etc.), Lithology	Large amounts of earth, rock, or mudflow quickly down the mountain sides and have a huge impact on water resources.
		Water alkalinity (pH), conductivity, and the concentration of various ions essential in many biogeochemical processes

Anthropogenic	Construction	Construction of dams causes changes in the natural hydrologic regime and the hydraulic characteristics of fluvial ecosystems, affecting the natural distribution of aquatic organisms and the export ratios of various organic and inorganic compounds. Direct dumping and site clearance activities are leading to deforestation, etc.
	Industrial	Waste is produced in chemical manufacturing units, printing, petroleum, leather, paper, metal etc.
	Agricultural	Mining process, Pesticides and Fertilizers- Runoff from the agriculture fields- Increase nitrogen concentration.
	Urban activities	Transportation, construction, domestic and municipal sewage disposal, over usage of water, deforestation, rise in impervious surfaces, increased urban runoff etc.

Source: Álvarez-Cabria et al (2016)

1.3 WATER QUALITY EXPLANATORY VARIABLE AT MULTI-SCALE

To better manage the effects of LULC on water resources, it is necessary to think of streams as complex ecosystems that operate at different spatial and temporal scales (Mello et al. 2020). To relate landscape variables to stream water quality, three spatial scales, including reach, catchment, and riparian, have been widely used (Ding et al. 2016; Shi et al. 2017; Mello et al. 2020).

Catchment scale: This scale refers to the entire drainage area that contributes water to a particular stream or river. At the catchment scale, LULC can affect water quality through changes in precipitation, evapotranspiration, and runoff, which can affect the amount and timing of water that enters the stream. Land use practices such as urbanization, agriculture, and forestry can also affect water quality by releasing pollutants such as nutrients, sediment, and chemicals.

Riparian scale: This scale refers to the area immediately adjacent to a stream or river, often referred to as the "riparian zone." At the riparian scale, LULC can affect water quality through changes in vegetation, which can alter the amount and timing of water that enters the stream, as well as through changes in the types and amounts of pollutants that are released into the stream.

Reach scale: This scale refers to a specific segment of a stream or river, and it is the most localized scale. At the reach scale, LULC can affect water quality through changes in stream flow, temperature, and water chemistry, which can affect the growth and survival of aquatic organisms.

By relating landscape variables to stream water quality at these different spatial scales, it is possible to understand how LULC affects water quality and to identify the specific land use practices that are having the most significant impact on water quality. This information can then be used to develop management strategies to improve water quality and protect aquatic ecosystems. The past results demonstrate that anthropogenic activities at various scales affect water quality. Also, the dimensions of these variables may vary between studies based on the concentration of water quality and the density of LULC at different scales. Different catchment areas can also be used, ranging from local-segment catchments to small-stream catchments to entire river basins (Mello et al. 2018). The relative importance of those three scales on water quality, hydrology, and biology is determined by the extent and intensity of each scale's land use pressures determines the relative importance of those three scales on water quality and hydrology. Graphics of different spatial scale is presented in Figure 1.1.

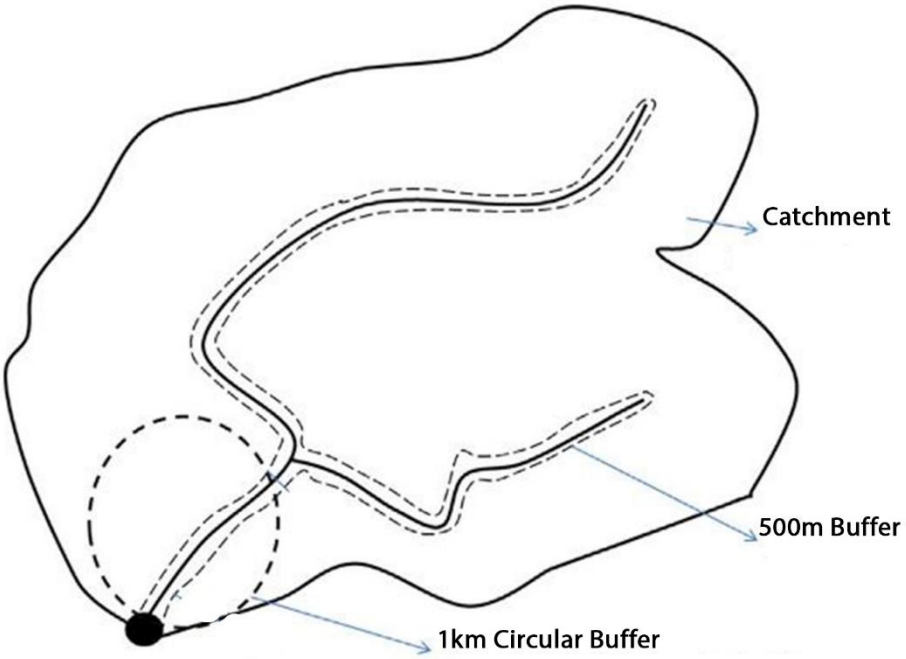


Figure 1.1 Graphics showing different spatial scales

Source: Mainali and Chang (2018)

1.4 REMOTE SENSING OF WATER QUALITY

Remote Sensing techniques have emerged as a widely accepted technology for carrying out research in complex water resource systems. Due to traditional laboratory sampling methods' coverage, efficiency, and cost-effectiveness, the fast-developing environmental information technology and remote sensing techniques have played an eloquent role in water quality monitoring. Numerous satellite of multi-sensors has been launched since the 1970s, continuously providing data. Satellite remote sensing is indeed a promising method for evaluating water quality variations in space and time (Glasgow et al. 2004; El Saadi et al. 2014; Bonansea et al. 2015; Liu et al. 2015; Zhou et al. 2017) along coastal, inland and estuarine water bodies (Vander Woerd and Pasterkamp 2004; González-Márquez et al. 2018; Yepez et al. 2018). Optically active substances in water, such as dissolved organic matter and algae, can interact with light and cause changes in the electromagnetic spectrum of reflected radiation. These changes can be analysed using techniques such as spectrophotometry and remote sensing to study the water's optical properties and detect the presence of specific substances (Koponen et al. 2002; Teodoro et al. 2007; Wen and Yang 2009; Haji

Gholizadeh et al. 2016). Remote sensing can be used to determine the relationship between the reflectance of water at specific wavelength bands (measured using sensors on satellites or aircraft) and *insitu* WQPs such as dissolved oxygen, chlorophyll, suspended sediment etc. This relationship, known as a "spectral signature," can be used to infer the water quality at a specific location based on the reflectance measurements. However, it is important to note that the relationship between remote sensing reflectance (R_{rs}) and WQPs can be affected by many factors, such as water body types, weather conditions, and atmospheric correction methods. Therefore, the accuracy of water quality estimation by remote sensing may vary depending on the specific case. Specifically, these studies will focus on identifying the various wavebands or band combinations (Panda et al. 2004; Sharaf El Din et al. 2017) with the highest correlation with different WQPs. Many studies have developed similar theories (Baba 1993; Allee & Johnson 1999; Andrzej Urbanski et al. 2016; Abdelmalik, 2018; Bonansea et al. 2018) and described the usefulness of remote sensing applications in continuous water quality monitoring programmes. Although, the best band or combinations suggested to predict WQPs differ from one study to another due to the optical complexity of turbid, productive waters. Besides, the transferability of algorithms developed in one study to other environments remains unknown from the papers reviewed. Visible (VIS), infrared (IR), and microwave (MW) are indeed considered to be some of the most critical spectral bands of interest for remote sensing of water bodies. VIS bands are sensitive to water depth, turbidity and dissolved substances like chlorophyll and suspended sediment. IR bands are sensitive to water temperature and dissolved gases like oxygen and can detect sediment and organic matter in the water. MW bands are sensitive to water surface roughness and water-land boundaries and can detect the presence of surface vegetation and shallow waterbodies. It's worth noting that depending on the specific application, other spectral bands can also be considered important, such as ultraviolet (UV) and short-wave infrared (SWIR) bands. Also, the importance of each band may vary depending on the specific water body type and the WQPs being studied (Chang et al. 2015a). Spectral absorption by pure water does follow an approximately parabolic trend. In the UV region, pure water absorbs strongly due to the presence of dissolved oxygen, dissolved carbon dioxide, and other dissolved gases. In the visible spectrum, water has a peak absorption at a wavelength of around 440 nm, which is in

the blue-green region. This is due to the presence of dissolved substances such as dissolved organic matter, algae, and minerals, which can also cause a slight greenish-blue colouration of the water. In the red-infrared (IR) region, pure water also absorbs strongly, this is due to the presence of dissolved gases and dissolved minerals. Therefore, pure water is usually a blue-green colour in transmission, although this can vary depending on the specific waterbody, and the presence of other dissolved substances (Julian et al.2013; Abdelmalik 2018) (Electromagnetic spectrum is illustrated in Figure 1.2).

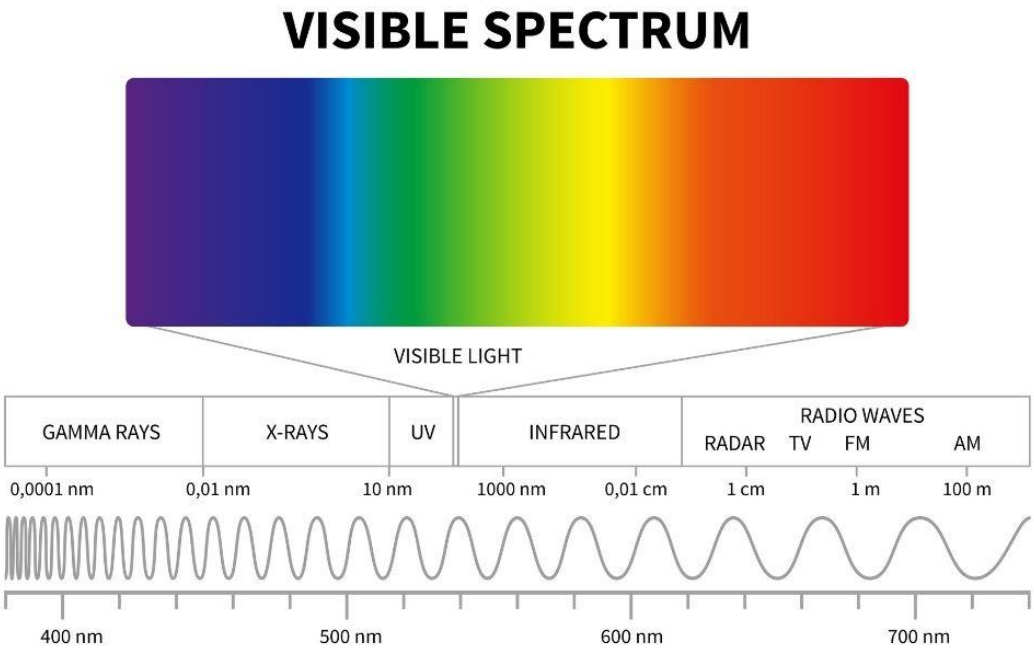


Figure 1.2 Electromagnetic spectrum

Source: Hajigholizadeh and Melesse (2017a)

A list of commonly used space-borne sensors in the field of remote sensing of water quality is listed in Table 1.2. The relationship between satellite spectral signatures and WQPs in inland water bodies, such as lakes, can be more complex and non-linear than in other types of water bodies. This is because the water dynamics in lakes are influenced by a wide range of factors such as water depth, sediment and nutrient content, algae growth, and meteorological conditions, all of which can affect the reflectance of the water at different wavelength bands. For example, in lakes with high sediment and chlorophyll content, there can be a considerable scattering of light in all VIS and NIR bands. This scattering can cause changes in the reflectance of the water,

making it more difficult to infer specific WQPs from satellite data. Additionally, the non-linear relationship between the WQPs and the spectral signatures can be caused by the presence of multiple substances with different spectral properties in the water, such as dissolved organic matter and mineral particles, making it difficult to estimate the concentrations of a single WQP. Therefore, the use of advanced techniques such as multi-temporal data analysis, atmospheric correction, and inversion algorithms can help to improve the accuracy of water quality estimation from remote sensing data in inland water bodies (Panda et al. 2004; Chang et al. 2015c; a; Gholizadeh et al. 2016; Ritchie et al. 2003)

Table 1.2 List of commonly used space-borne sensors in water quality

Category	Types of Satellite-sensors	Date of Launch	Spectral bands (nm)	Spatial resolution (m)	Revisit in (Days)	Mission Status
Very high resolution	Geo Eye-IKONOS	24-Sep-1999	4 MS (445–853), 1 Pan (526–929)	3.2-0.82	~3	Decommissioned
	SPOT 5-HRG	4-May-2002	3 MS (500–890), 1 Pan (480–710), 1 SWIR (1580–1750)	2.5 and 5–10–20	2-3	Decommissioned
	Digital Globe Quickbird	18-Oct-2001	4 MS (430–918), 1 Pan (450–900)	0.65	1-3.5	Decommissioned
	CARTOSAT-1	5-May-2005	Pan (500–850)	2.5	5	Active/ Elapsed Life of 13 years, six months, 29 days
	CARTOSAT-2	10-Jan-2007	Pan (450-850)	≤ 1	-	Active
	Digital Globe Worldview-1	18-Sep-2007	Pan	0.46	1.7	Active/ Expected end at the fourth quarter of 2020
	Geo Eye-1	6-Sep-2008	4 MS (450-920), 1 Pan (450-800)	1.84-0.46	1-3	Active
	Digital Globe world view-2	8-Oct-2009	8 MS (400–1040), 1 Pan- (450–800)	1.85-0.46	1.1	Active/ Expected end in the fourth quarter of 2022.
	NOAA World view-3	13-Aug-2014	8 MS (400–1040), 1 Pan- (450–800), 8 SWIR (1195-2365)	1.24-3.7-0.31	1-4.5	Active/ Estimated life 10-12 years

	Digital Globe Worldview-4	11-Nov-2016	4 MS (655-920), Pan-(450-800)	0.31	1-4.5	Active/10-12 years of Estimated Life
High resolution	Landsat- 5 MSS	1-Mar-1984	4 MS (450-1750)	80	18	Decommissioned
	Landsat- 5 TM	1-Mar-1984	5 MS (450-1750), 2 SWIR (1550-2350), 1 Thermal (10400-12500)	30-120	16	Decommissioned
	IRS-1A	29-Aug-1988	1 LISS-I (450-520), 3 LISS-II A/B (520-860)	72.5, 36	22	Decommissioned
	IRS-1B	29-Aug-1991	1 LISS-I (450-520), 3 LISS-II A/B (520-860)	72.5, 36	22	Decommissioned
	IRS-1C	28-Dec-1995	3 LISS-III (520-1700), 1 PAN (500-750), 2 WiFS (620-860)	23.5, 70, 5.8 (PAN), 188.	24, 5(WiFS)	Decommissioned
	Landsat-7 ETM+	15-Apr-1999	4 MS (450-900), 2 SWIR (1550-2350), 1 Thermal (10400-12500), 1 pan (520-900)	30-15-60	16	Active
	IPS-P6 ResourceSat-1	17-Oct-2003	3 LISS-IV (520-860), 4 LISS-III (520-1700), 3 AWiFS (620-1700)	5.8, 23.5, 70	24, 5(WiFS)	Active
	Landsat-8- OLI	11-Feb-2013	5 MS (430-880), 2 SWIR (1570-2290), 1 Pan (500-680), 1 Cirrus (1360-1380), 2 Thermal (10600-12510)	30-15-100	16	Active

	ALOS AVNIR-2	24-Jan-2006	4 MS (420-890), 1 Pan (520-770), 1 L-Band (1.3GHz)	10	46	Decommissioned
	EO-1 ALI	21-Nov-2000	9(433-2350)- 1Pan(480-690)	10-30	16	Decommissioned
	EO-I Hyperion	21-Nov-2000	242(350-2570)	30	16	Decommissioned
	HICO	10-Sep-2009	128(350-1080)	100	10	Decommissioned
	Sentinel-2A MSI	23-Jun-2015	10 MS (443-945), 3 SWIR (1375-2190)	10, 20, 60	5-10 days	Active
	Sentinel-2B MSI	07-Mar-2017	10 MS (442-943), 3 SWIR (1376-2185)	10, 20, 60	5-10 days	Active
	SPOT-1 HRV	22-Feb-1986	3 MS (500-890), 1 Pan (510-730)	20,10	26 days	Decommissioned
	SPOT-2 HRV	22-Jan-1990	3 MS (500-890), 1 Pan (510-730)	20,10	26 days	Decommissioned
	SPOT-3 HRV	26-Sep-1993	3 MS (500-890), 1 Pan (510-730)	20,10	26 days	Decommissioned
	SPOT-4 HRVIR	24-Mar-1998	3 MS (500-890), 1 SWIR (1530-1750), 1 Pan (610-680)	20,10	26 days	Decommissioned

Moderate resolution	IRS-P3	21-Mar-1996	3WiFS (620-1700), 3 MOS (750-1605), IXAE (Indian X-Ray Astronomy Experiment)	188, 1500, 520, 550	5	Decommissioned
	MERIS	1-March-2012	15 MS (390-1040)	300-1200	Daily	Decommissioned
	MODIS	18-Dec-1999	2(620-876)-5(459-2155)-29(405-877) and thermal	250-500-1000	1-2	Active
	GOCI	26-June-2010	8(400-865)	500	1 hr/ 8 image in day time	Active/7years

The advantage of the temporal and spatial coverage of optical remote sensing are:

- The best approach for overcoming the disadvantages of *insitu* water quality monitoring and reducing the difficulties in dealing with the complex heterogeneous and dynamic behaviour of coastal, inland, and estuaries (González-Márquez et al. 2018; Yepez et al. 2018).
- Remote sensing data generally have a relatively high temporal resolution. This allows for repeated analysis and facilitates time series analysis, which can provide information on how WQPs are changing over time. (Kong et al. 2015) and quantify water quality issues (Haji Gholizadeh et al. 2016). Additionally, time series analysis can be used to detect changes in water quality that may be caused by human activities, such as changes in land use or changes in water management practices.
- Remote sensing can be used to analyse water quality issues at different scales, such as at the regional, local, and watershed scales. By using remote sensing data with different spatial resolutions, it is possible to study the water quality of large areas, such as entire regions, as well as more localized areas, such as individual lakes or watersheds.
- Satellite imagery provides near-continuous spatial coverage over large areas, allowing synoptic estimates of water quality. The use of satellite imagery can provide a synoptic view of water quality across a region, which can be useful for identifying patterns and trends in water quality over time. Additionally, a long record of archived Landsat imagery can be used to estimate historical water quality when ground measurements cannot be performed.

1.5 MULTIVARIATE STATISTICAL TECHNIQUE FOR WATER QUALITY ANALYSIS

Multivariate statistical approaches are extensively used for classification, model construction, dimension reduction, and interpreting water quality data with a minimum loss of original information. These statistical techniques enable the determination of the factors that significantly impact the river water quality. Such methodological approaches have emerged as a critical method for developing effective water

management strategies and addressing pollution issues. Besides this, multivariate statistical analysis can aid in validating seasonal variations induced by natural and anthropogenic factors. Principal Component Analysis (PCA), Factor Analysis(FA), Cluster Analysis (CA), Discriminant Analysis (DA) and Redundancy Analysis (RDA) are all commonly used multivariate techniques in the field of water quality research. These techniques analyse large datasets and identify patterns or relationships between WQPs. They can also classify or group water samples based on their chemical or physical characteristics. A multivariate model can be broadly classified into three groups:

- Univariate analysis - which examines only one variable
- Bivariate analysis - examines two variables.
- Multivariate analysis - more than two variables.

1.6 INVERSION METHODS AND RETRIEVAL ALGORITHM

Remote sensing based water quality monitoring is a complex process that involves multiple steps, including image acquisition, pre-processing, data inversion, and data analysis. Selecting a suitable inversion method to find strong relationships between WQPs and remotely sensed data is one of the main challenges in this field. Inversion methods are used to infer WQPs from remotely sensed data, such as reflectance or radiance measurements. Different inversion methods have different strengths and weaknesses, and the choice of method will depend on the specific characteristics of the data and the research question. Some commonly used inversion methods include empirical and semi-analytical algorithms, such as the Beer-Lambert Law, and more complex methods, such as neural networks and machine learning algorithms. It is important to carefully evaluate the performance of different inversion methods and select the one that provides the most accurate and reliable results for a specific study. This often requires extensive testing and comparison of results using both remotely sensed data and *insitu* measurements of WQPs. The inversion model can be broadly classified into three

- Empirical Method
- Analytical Method
- Semi-empirical/semi-analytical method (Hybrid method)

Empirical methods are a type of inversion method that is used to infer WQPs from remotely sensed data by identifying and modelling the statistical relationships between the measured WQPs and the spectral values (Chang et al. 2015). These methods are simple to implement and can provide quick results. They are commonly used in water quality monitoring applications, such as the development of algorithms for retrieving WQPs from satellite data. Empirical methods are relatively inexpensive and easy to use, but their results depend on the data quality used to establish the empirical relationship, and their accuracy may be limited in certain cases. The most critical barrier of empirical methods approaches is uncovering the relationships among WQPs and spectral values (single bands, band combinations, or band ratio) through regression-related methods. The existence of a nonlinear relationship among WQPs, combined with the failure of linear regression techniques to detect genuine relationships, inspired scientists to devise novel approaches. As an outcome, machine learning methods such as ANN, RF, SVM, PSO, and GP, as well as combinations of these models, have been employed to retrieve WQPs from different bands of remote sensing data. The analytical method models the reflectance using inherent optical properties (such as the absorption coefficient, scattering coefficient, and volume scattering function) and intrinsic optical properties (such as the diffuse attenuation coefficient for downwelling irradiance and irradiance reflectance). The analytical approach uses a bio-optical approach to retrieve WQPs. Remotely sensed bio-optical models use the connection among inherent optical properties (IOPs) and apparent optical properties (AOPs) to evaluate WQPs based on sensor reflectance. IOPs are water's physical attributes that govern how light interacts with it. These include the water's absorption and scattering coefficients, which are determined by the concentration of dissolved and suspended substances. AOPs are water properties that can be analysed utilising remote sensing methods, including reflectance or radiance. These properties are influenced by the IOPs of the water, as well as the viewing and illumination constraints and sensor characteristics. These models use information from light-water interactions to measure the concentrations of WQPs. Although the analytical method is independent of *insitu* water quality data, it depends on the optical properties of water bodies. It thus is not a method that can be used across a wide range of spectral bands. Semi-empirical/semi-analytical methods, on the other hand, are the integration of empirical and analytical data.

1.7 SCOPE OF THE PRESENT STUDY

Continuous monitoring of water supplies for domestic, industrial, and irrigated agriculture, livestock, and mining activities is required to ensure the right norms and standards are met (Giri and Qiu 2016). The impacts of water pollution are not limited to the environment; it also affects human health, livelihoods, and economies, and it is crucial to mitigate these impacts. Dischargeable sewage, agricultural runoff, and industrial waste have led to increased levels of bacteria and other pathogens in water bodies, making them unsafe for drinking and recreation. Identifying the causes of water quality deterioration, measuring various water quality indices, determining appropriate explanatory variables, processing the data to capture the effect of these variables, and modelling them using identified explanatory variables are all essential steps in conducting a comprehensive water quality assessment. Because anthropogenic activities significantly impact most freshwater environments, dismissing human disturbance factors will limit the model's robustness and accuracy (Wang and Yang 2019). As a result, the relationship between landscape characteristics and water quality provides critical information for addressing specific NPS management challenges. Remote sensing and GIS techniques are the most effective, cost-effective, and reliable tools for continuously monitoring and interpreting spatiotemporal phenomena. Moreover, the availability of open-access software and scripting provides additional benefits for processing these images. Given the preceding discussion, the purpose of this research is to investigate,

- What is the geographic distribution of seasonal water quality changes at monitoring stations? Where may statistically significant trends be found? How seasonal are the monitoring stations? What are the most critical WQPs that favour spatiotemporal trends?
- What are the implications of various seasons and analysis scales on the relationship between water quality changes and landscape characteristics? The study hypothesises that the factors impacting water quality trends differ depending on the season. The researchers also hypothesize that predictor variables generated at smaller scales can explain water quality better than the full catchment approach.

- Aside from that, the study focused on describing the spatiotemporal variability of dominating WQPs on smaller scales utilising remote sensing and machine learning-based frameworks for improved management techniques.

1.8 ORGANIZATION OF THESIS

This thesis is divided into six chapters.

Chapter I: Provides a brief overview of the importance of understanding the concentration of different WQPs present in any waterbody and the water quality explanatory variable at multi spatial scale. Besides that, the study discussed the benefits of incorporating remote sensing data and *insitu*, using machine learning techniques. The commonly used sensors in previous studies and the scope of the study are also discussed here.

Chapter 2: Provides a comprehensive review of the literature on spatiotemporal variations of WQPs, feature selection, and dimension reduction to identify the dominant WQPs. The literature on assessing water pollution caused by anthropogenic activities is also reviewed. Finally, the study examined the literature using remote sensing and *insitu* data to overcome the disadvantage of understanding water quality problems at a finer scale. The research objectives are presented based on the identified research gaps.

Chapter 3: Describes the study area, the water quality problem along the Ganga River Basin, and the justification for choosing the Middle Ganga Basin for this study. There is also a brief description of the data collection and pre-processing steps.

Chapter 4: Summarises the theoretical aspects of various multivariate statistical and machine learning techniques used to achieve the multiple objectives in this study.

Chapter 5: Presents the results and discussions in separate sections in the exact order in which the objectives are derived. The results of various multivariate statistical techniques used to achieve the first and second objectives and the discussions based on the results are detailed. Finally, the spatiotemporal maps and predicted dominant WQPs generated using machine learning techniques are presented and discussed in this section.

Chapter 6: Highlights the results' conclusions, as well as the study's limits and future scope.

CHAPTER 2

LITERATURE REVIEW

2.1 BACKGROUND

Natural and anthropogenic components strongly influence the river water quality of any area; understanding these drivers is critical for the long term viability of the aquatic ecosystem. Over the last few decades, there has been an increase in the demand for regular monitoring of many rivers, which has increased the accumulation of reliable long-term water quality data. Advanced computing technology now allows processing and manipulating enormous databases in various ways that were impossible before (Antonopoulos et al. 2001). Water quality data are frequently collected at various stations will account for changing hydrogeological conditions. These flaws eventually result in non-normally distributed, noisy outliers and missing data, which cause significant deviations in modelled and monitored results (Fu and Gan Wang 2012; Liu et al. 2019). Numerous studies have been undertaken to analyse the influence of various contaminants on river water quality utilising water quality indices (WQI). Even though WQI is beneficial for predicting changes in water quality by considering various characteristics, it does not give evidence on pollution sources because it is derived after the normalisation of analytical results (Wunderlin et al. 2001).

On the other hand, a univariate strategy is a most often used method for analysing river water quality. Still, it is not a viable solution for environmental data involving many physiochemical variables. As a result, the application of multivariate statistical and machine learning models to learn spatial and temporal hydrological data has shown to be a valuable tool in dealing with uncertainties in water quality data (Sundaray et al. 2006; Azhar et al. 2015) caused by natural and anthropogenic factors (Singh et al. 2004; Bhat et al. 2014). As emphasized in the previous section, the advancements in communications and technologies caused quality assessment to become a new change norm. Incorporating remote sensing with traditional water quality programs has boosted the data's synoptic coverage and temporal consistency. Furthermore, their capacity to give vital information on inland and near-coastal transitional waters in locations where traditional water quality programmes are either missing or insufficient has expanded

the scope of evaluation in isolated areas. It also allowed a successful real-time evaluation of water quality as well as the quick detection of possible contamination, such as eutrophication and dangerous algal blooms.

Several study articles have been published in recent years that analyse WQPs to determine the most significant water quality variable and parameter that causes temporal and geographical fluctuations caused by natural and human seasonal variables, and so on. At the onset of the survey, the research articles on these topics were reviewed and documented. The overall methodology followed in this chapter is explained below.

- An overall methodological framework for the literature review will be explained in Sec. 2.2.
- Evaluation of spatiotemporal variations in WQPs Sec 2.3.
- Sec 2.4 Assessment of water pollution induced by anthropogenic activities.
- Mapping the concentration of WQPs using satellite data and machine learning algorithms is discussed in Sec 2.5.

2.2 THE METHODOLOGICAL FRAMEWORK FOR LITERATURE REVIEW

Good definitions are essential for most research studies, particularly in theory building, where concepts must be aptly defined, and in literature reviews, the research topic should be clearly outlined. As discussed in the review paper by Pourhabibi et al. (2020), this literature survey adopted a systematic approach to the literature review and followed the three-phase methodology employed by Ngai et al. (2011). "Research definition" was part of the first phase. It entails identifying the research area, developing review objectives, and defining the research scope. The river water quality assessment is the focus of this study, with three main goals: (1) identifying current research trends in the topic, (2) highlighting current challenges and providing directions for future research. The scope also covers studies conducted on inland waters using satellite data products in the remote sensing domain. The second stage is "research methodology," which begins by searching scientific databases for publications in peer-reviewed journals that are relevant to the study. The latest research papers were segregated based on the parameters studied, the scale of the study and the methodology applied. The last

phase was based on the set of research questions, similar to the procedure employed by (Chan et al. 2017; Pourhabibi et al. 2020; Snyder 2019), which included multiple levels of interpretation and analysis of previous research within the domain area, are addressed in the following sections.

2.3 EVALUATION OF SPATIOTEMPORAL VARIATIONS IN WQPS

2.3.1 Spatial pattern of seasonal water quality trends in monitoring stations

Investigating any study's spatiotemporal changes of dominant WQPs is critical for proposing the appropriate treatment for the water bodies. River water quality data at numerous spatiotemporal scales are necessary for environmental pollution control and policy planning for contaminated site management. Evaluating river water's physical, chemical, and biological condition at finer spatiotemporal scales is critical for operationalising wastewater facilities and sectoral water supply from riverine sources (Swain and Sahoo 2017a). However, they were identified as not being unique for different studies. Mainali and Chang (2018) observed that the significant factors differ across scales but not across seasons on the same scale. This part of the literature survey discussed the various multivariate statistical techniques applied over the year to understand the monitoring stations causing the spatiotemporal changes and the significant WQPs.

In real-world problems, the dataset collected from the respective authorities is unlabelled. In machine learning, there are various methods for labelling these datasets, and clustering is one of them. Clustering is an unsupervised learning technique to predict the groups from an unlabelled dataset. Over the past years, clustering techniques have been widely used in engineering, economics, geology, electronics, statistics, and psychology. It divides the input space into regions based on predesigned criteria without training data (Ay and Kisi 2014; Li et al. 2016). Since it does not require any prerequisite knowledge about the data, it is also known as unsupervised learning (Kamble and Vijay 2011; Hajigholizadeh and Melesse 2017). Clustering identifies the subgroups in the data so that the data points in the same subgroups or clusters are similar, while data points in different clusters are different (Shamitha and Ilango 2019). Chang et al.(2012) used K-means clustering to divide the dataset into a user-specified

number of subsets called clusters to discover and evaluate the spatiotemporal patterns of WQ in Tampa Bay based on parameters of similar qualities. K-means clustering is one of the simplest unsupervised learning algorithms, which identifies the K number of centroids in the whole dataset and allocates every data point to the nearest cluster while keeping the centroids as small as possible. The K-means algorithm initially starts by randomly selecting a centroid value for each cluster. The K-means algorithm calculates the Euclidean distance within each data instance and centroids of all the clusters. Then it assigns the data instances to the cluster of the centroid with the nearest distance. After the clusters are assigned, a new centroid value is calculated based on the mean values of the coordinates of all the data instances from the corresponding cluster. The principle behind the K-means clustering is it calculates the Euclidean distance between two points, which can be measured with a measuring device or found using the Pythagorean formula (Chang et al. 2012a).

2.3.2 Feature selection and dimensionality reduction

PCA is a statistical unsupervised machine learning approach that uses an orthogonal transformation to convert a group of correlated variables to uncorrelated variables (Li and Liu 2018). Each of the principal components (PCs) is chosen to describe the majority of the available variance in the data and is orthogonal to each other. The first principal component has the most significant variance of all PCs. It is a necessary tool for ecological evaluation since most environmental studies are made up of multiple variables, making it difficult to identify relevant patterns in the data. The factor score is used to organise the data into groups of variables having the greatest relationships. In other words, the groups, also known as PCs, are organised or sorted in the order of the total variance explained. As a data reduction strategy that compresses a large number of variables into a smaller number of variables (Wunderlin et al. 2001) that makes ecological assessment more practicable. Feature selection and dimensionality reduction task is crucial for high dimensional machine learning analysis to select dominant features in training the dataset. Besides, these techniques can be further helpful in preventing overfitting, simplifying the model to improve the computational efficiency and reducing the algorithm's overall running time. It is critical to gather only relevant characteristics in the training dataset before executing any machine learning

method., i.e. minimising the dimensions of feature space is called dimensionality reduction. Besides, it further helps to prevent overfitting and makes the model simple and efficient with less running time. PCA and linear discriminant analysis (LDA) is the popular pre-processing linear transformation techniques often used for dimensionality reduction and to select relevant features. Table 2.1 lists the critical findings and WQPs used in various studies.

Table 2.1 Feature selection and dimension reduction of water quality data

Parameters	**Models**	**Critical Findings**	**Reference**
22 WQPs	Spearman *r* coefficient, FA/PCA, DA	FA/PCA gave a 40% data reduction, selected 13 out of 22 WQPs and explained 71% of spatiotemporal changes. DA observed considerable data reduction with 6 WQPs 73% reduction to differentiate samples from monsoon or non-monsoon season and 5 WQPs 77% reduction in data to differentiate spatially.	(Wunderlin et al. 2001)
24 WQPs	CA, FA/PCA	Hierarchical CA grouped the eight sampling sites into 3 clusters of similar characteristics based on their water quality characteristics and pollution load (Natural and Anthropogenic). Data reduction by PCA and DA and observed soluble salts (natural) and organic pollution load (anthropogenic) are the parameters responsible for spatiotemporal variations.	(Singh et al. 2004)

Chl-a, DO, TKN, TP and water temperature	CA and DA	Stepwise DA and Spatial DA identified the most critical discriminating WQPs responsible for temporal and spatial variations. Using the CA trend in environment pollution was found from the low-pollution region to high pollution region.	(Hajigholizadeh and Melesse 2017b)
DO, FC, BOD, pH, T, TP, TS, Nitrate	PCA, CA	PCA of all WQPs showed a sample separation based on seasonality. CA grouped the stations based on similarity and dissimilarity exist between the parameters.	(Ahmed et al. 2019)

2.4 WATER POLLUTION INDUCED BY ANTHROPOGENIC ACTIVITIES

LULC has emerged as an important research topic for determining the relationship between surface water quality and non-point source (NPS) pollutants, which are essential pollutant regulators in overland flow and interflow (Chen et al. 2016b). NPS pollution concerns are distinguished by complicated processes and occasional occurrences, attributed mostly to NPS, which is recognised to have a direct association with LULC (Abdulkareem et al., 2018). The point source is easily identifiable and governed by identifying its source, primarily industrial and domestic sewage load. However, owing to the complex and diffuse interaction between runoff and landscape, NPS pollution is challenging to identify (Ding et al. 2016; Giri and Qiu 2016). NPS usually involve urban and agricultural runoff, pollutant decomposition, and so on, whereas land use patterns such as forest cover tend to retain water quality conditions (Álvarez-cabria et al. 2016; Mello et al. 2020). Many previous studies have demonstrated a strong correlation between WQPs and agricultural land use due to fertilizer mixing into river water via agricultural runoff (Patra et al. 2018; Ahmad et al. 2021; Umwali et al. 2021). Urbanization alters the water quality via three primary mechanisms: i) pollutant discharge at the point source and mobilization of pollutants

from diffuse sources; ii) flow modification; and iii) changes in water temperature (Miller and Hutchins 2017). As a result, watersheds that include relevant water catchments used for domestic, agricultural, and industrial purposes necessitate well-balanced LULC planning to reduce the detrimental impacts of certain types of land use on river water quality (Meneses et al. 2015). Many researchers have previously examined the effect of LULC on water quality using various analysis techniques, particularly statistical tools. It can be established that the entire area upstream from the monitoring stations may be utilised as a predictor of WQPs at many scales and that comparing scale and seasonal impacts with LULC variations could be useful in exploring the complex dynamics of LULC on water quality (Tanaka et al. 2016; Shukla et al. 2017; Cheng et al. 2018; Mello et al. 2018). As a result, one critical question is which land use spatial extent has the most effect on water quality (Ding et al. 2016). Using scale correctly allows managers to make better decisions and use resources more efficiently. However, the outcomes are not always reliable. Some studies discovered that land use at the reach or riparian sizes predicted WQPs better than land use at the catchment scale (Wang et al. 2012; Sandoval et al. 2014; Pathak et al. 2018; Gu et al. 2019), whilst Others discovered that land use at the watershed scale accounted for better variability in water quality. These contradictory findings were most likely attributable to discrepancies in study designs and geographical areas. Multivariate statistical methods including cluster analysis (CA), correlation analysis, and Principal Component Analysis (PCA) are extensively utilised assessing temporal and spatial variations of WQPs, as well as Redundancy Analysis (RDA) to evaluate the global descriptions of the influencing factors of the LULC pattern on WQPs while accounting for the various scale effect. In the table below, we discussed spatial scale, models, and key findings from the most recent publications from 2017 to 2021 (Table 2.2).

Table 2.2 Recent literature on assessment of water pollution induced by anthropogenic factors

Spatial Scale , WQPs, Environmental Variables	Models	Critical Findings	Reference
Catchment, Riparian and Reach scale. BOD, COD, DO, EC, pH, TSS, nitrate nitrogen and ammonium nitrogen. Agriculture, Forest, Grass land, Urban. Water and Landscape metrics	One-way ANOVA, t-test, RDA	A strong relationship between land use and WQPs was observed during the monsoon season. The RDA results revealed that riparian scales explained more of the LULC patterns on WQPs than catchment and reach scales. Various land use measures generated different scale impacts, emphasising that multi-scale land use planning should be used in water quality management.	(Shi et al. 2017)
Sub catchments TSS, TP, TN and ammonia nitrogen	ANOVA	Changes in LULC condition increased NPS pollutant loads among different LULC changes. Urbanisation was the dominant LULC change with the highest pollutant load. LULC changes have given rise to high-level TSS in urbanised areas.	(Abdulkareem et al. 2018)
WQPs LULC Elevation Slope Soil Types Population density	Mann Kendall's trend analysis, Correlation, Spatial Autocorrelation, Regression	The explanatory power of the 100m buffer and one-kilometre upstream scale analyses was more significant than that of the sub-watershed size studies. Each regression model's significant factors vary between scales but not between seasons on the same scale. By reducing residual spatial autocorrelation, the	(Mainali and Chang 2018)

		spatial filtering strategy considerably boosted the explanatory power of water quality trend models.	
Watershed and Riparian	Correlation, RDA	The RDA model explained 82% of the variation for the whole watershed and 75% for the riparian zone composition. Forest cover is significant in keeping water clean, urban and agricultural areas degrade water quality. Also identified the importance of streamflow and temperature as important predictors that explain some variations in WQPs.	(Mello et al. 2018)
Sub-watersheds TOC, COD, TP, NO_3, EC Farmland, forest land, grassland, bodies of water, barren land, urban land, and rural land, altitude, slope, NDVI and landscape metrics	Source-Sink landscape theory, RDA and PLS-SEM	Watersheds containing urban landscapes significantly impacted water quality ($\beta = 0.835$, $p<0.001$), indicating water quality degradation due to point source pollution in the non-monsoon season. Rainy season - Agriculture and urban areas serve as the "source" landscape. Agricultural development exacerbates inorganic pollution, and urban development exacerbates organic pollution. Natural vegetation landscape - serves as a "sink" for inorganic pollution and a "source" for organic pollution.	(Wang et al. 2021)
DO, FC, pH, BOD, Temp, TP, Nitrate, Turbidity, TS	Pearson Correlation, PCA and CA Partial Least Squares Path Modelling	During the non-monsoon season, a strong positive correlation was observed between DO, TP, and cropland; turbidity and forest; FC and built-up areas; and FC and waterbodies. BOD and cropland had a strong positive correlation; turbidity and forest; BOD and grassland had a strong	(Umwali et al. 2021)

		negative correlation; temperature and built-up had a strong negative correlation. Nitrates and wetlands had a strong positive correlation, whereas TP and built-up had a negative correlation.	

Although traditional monitoring techniques are more accurate, they are time- and labour-intensive. Many studies incorporating remote sensing and *insitu* WQPs have been published over the last few decades. Remote sensing techniques have evolved into valuable tools for achieving the goal of continuous water resource monitoring. With advancements in space sciences, computer applications and computing power have established a new standard for water quality assessment. Based on the literature study, a collection of sensors used to measure different WQPs, bands of interest, modelling methodologies, and critical discoveries were found and are provided in separate tables.

2.5 MAPPING THE WQPS USING SATELLITE DATA AND MACHINE LEARNING ALGORITHMS

Long term data requirement is vital to address the state of the world's water resources. Field measurements for water quality assessment include costly, time-consuming, labour-intensive sampling on site and transport to land-based or shipboard laboratories for their evaluation. Besides, there are possibilities to compromise on data due to poor quality-control protocols and quality assurance due to the extended holding of samples before analysis (Gholizadeh et al. 2016). The fast developing environmental information technology and remote sensing techniques have played a significant role in water quality monitoring due to the coverage, efficiency and cost saving. It depends on the spectral response or scattering reflected from the water. Remote sensing techniques employ different optical properties of surface water by changing the reflected energy spectra or emitting thermal radiation from it (Ritchie et al. 2003; Swain and Sahoo 2017b). The concentrations of optically active water constituents can be estimated from satellite images by interpreting the received radiance at the sensor at different wavelengths (Keiner and Yan 1998). Sensors on board satellites with a wide range of spectral (the ability of a satellite sensor to measure specific wavelengths of the electromagnetic spectrum), temporal (time between images), and spatial resolutions (size of the smallest feature that a satellite sensor can detect) have been proposed to access several WQPs (Liu et al. 2015). These processed data have the potential to analyse the trends in many instances of time. With the incorporation of traditional sampling data, the assessments of WQPs have been improved (Steissberg et al. 2010).

2.5.1 Satellite Sensors and Platforms

Technological advances in the late twentieth century improved satellite sensor capabilities, such as finer spatiotemporal and spectral resolutions. Sensors with short revisit times, such as MODIS Aqua, enable more satellites to be used for real-time monitoring applications (Chang et al. 2015a). This has created a huge opportunity for event-based water quality assessment, such as surface water quality assessment after the flood. Passive remote sensing devices that detect light within visible NIR regions of the electromagnetic spectrum (400-1000 nm), whether portable or mounted on aeroplanes or satellites, are most typically employed for water-related applications. The most widely used data resources are the Landsat series satellites. Because of its early launch, open access, and longest operational period in orbit, Landsat-5 TM is used significantly more than any other multispectral sensor (Wang and Yang 2019). The thermal band fitted to Landsat-7 and 8 extends the parameter range measured by the ETM sensor. Multispectral sensors give more integrated spectral information than hyperspectral sensors but have lower spectrum resolution. Hyperspectral remote sensing has shown high potential in detecting water quality conditions and their parameters (Peneva et al. 2008; Olmanson et al. 2013; Antonini et al. 2017; Chen et al. 2016a). The narrow intervals of hyperspectral channels enable a wide range of reflectance that is useful in assessing the water quality conditions of many open water aquatic ecosystems (El-Magd and El-Zeiny 2014). Remote sensing of shallow waters may produce images characterized by limited image coverage, strong uneven background, and high noise/speckle levels, which contribute to the challenges of extracting spatial information. Phinn et al. (2008) compared Quickbird-2 multi-spectral, Landsat-5 Thematic Mapper multi-spectral data and Airborne hyperspectral image CASI-2 sensor using a pixel size of 4.0 m. The study observed the highest overall accuracies (46%) for airborne hyper-spectral data produced, followed by Quickbird-2 and Landsat-5 Thematic Mapper. The low accuracy levels were attributed to the mapping methods and difficulties in matching locations on image and field datasets. The applications of hyperspectral remote sensing have been applied in riverine (Lepistö et al. 2010; Rostom et al. 2017; Zhang et al. 2020), coastline (Bertels et al. 2008; El-Magd and El-Zeiny 2014; Keith et al. 2014) and both riverine and costline water quality analysis (Brando and Dekker 2003). However, the major drawbacks of airborne

hyperspectral data for routine monitoring are the limited coverage during each flight and the relatively high costs involved (Bertels et al. 2008).

Morel and Prieur (1977) grouped the surface water into Case I and Case II waters based on their optical properties. Case I generally refers to open ocean, whereas Case II inland waterbodies, estuaries and coastal waterbodies. In Case I, chlorophyll is the most optically active constituent, and the water contains little suspended sediment (SS). As a result, the algorithms that use empirical models to relate sensor radiances to surface concentrations have yielded promising results (Sudheer et al. 2007). Therefore, this group's optical properties can be modelled as a function of Chl-a.

However, in Case II waters, the relationship between sensor radiance and WQPs is more complicated due to the interaction of components, such as chlorophyll, suspended sediments, and yellow substances. Therefore, in the complex Case II waters, the optical properties cannot be modelled as a function of Chl-a but as an independent variable (Novoa et al. 2011). Some marine coastal waters will neither belong to Case I nor Case II waters are referred to as a Non-Case I waters (Kondratyev et al. 1998). Remote sensing of turbid Non-Case I water still poses numerous challenges due to poor interpretation of signals stemming from different constituents of such waters. Considering past research, it can be noticed that remote sensing has been widely used for water quality monitoring of optically active substances like Turbidity (T), Phytoplankton and Yellow substances or Coloured Dissolved Organic matter (CDOM) (Novoa et al., 2011). These studies determine the reliable relationship between R_{rs} at certain wavelength bands and *insitu* WQPs. The studies specifically sought to explain the current level of empirical remote sensing in inland waterways and to identify remotely sensed band(s), band ratios, and band arithmetic variables suitable for identifying specific characteristics in various water types using statistical methodologies (Matthews 2011).

Sensors involved in water resource problems cover a broad portion of the EM spectrum that indirectly measures variables of a water resource system using remote sensing techniques later, these hydrological variables can be solved by applying some empirical or transform functions. The VIS, IR, and MW bands are the essential spectral bands of interest for remote sensing of water bodies (Chang et al. 2015a). These bands allow for the detection and analysis of various water properties, such as water depth, temperature,

and turbidity. Each band has its advantages and limitations, and they are often used in combination to provide a complete picture of the water body being studied. In the case of inland water, the water dynamics are more complex to have a linear relationship between the satellite spectral signatures and WQPs. There is considerable scattering in all VIS and NIR bands from the lake waters with high sediments and chlorophyll content (Panda et al. 2004). Thus, attention has to be paid to identifying the methods capable of sensitively and continuously detecting and quantifying WQPs (Li and Liu 2018). In this chapter's coming sections, different modelling techniques are discussed for various WQPs using remote sensing techniques.

2.5.1.1 Water temperature

The advantage of satellite data in digital format is that it can be easily combined with other geographic information and could be used to create temperature models. Emitted thermal infrared radiation (TIR, λ = 8 to 14 µm) can be used to measure surface water temperature (top approximately 100 µm) (Haji Gholizadeh et al. 2016; Vanhellemont 2020). The wide range of TIR sensors provides many opportunities to measure water bodies' temperatures. The TIR technology to measure the water temperature of rivers are diverse and has been employed in a wide variety of fluvial environments (Handcock et al. 2012). Most of the works focused on monitoring the temperature of lakes primarily using Landsat TM and TM+ sensors with the perfect spatial and time resolution (Ramsey et al. 1992; Kay et al. 2005; Lamaro et al. 2013; Bonansea et al. 2015).

In most cases, the band ratio multi regression approach is used to obtain the relation between satellite signal and *insitu* measurements. The success of WQPs quantification in inland environments depends on water characteristics and the used sensor. The combination of good real-time coverage, spatial resolution and accessible data availability makes the Landsat system appropriate for studying these waterbodies.

2.5.1.2 Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD)

Remote sensing applications still present challenges in estimating DO, BOD, and COD. Landsat 8-based BPNN have shown an R^2 of 0.934 (Sharaf El Din et al. 2017), and a

similar accuracy is observed with the regression model (González-Márquez et al. 2018). Moreover, the study by Wang et al. (2011) using GA-SVR and SPOT-5 data observed an R^2 of 0.938. In another study conducted using Landsat 7, ETM+ identified the bands B1, B2 and B4, showing good agreement with the remote sensing data using multiple linear regressions (Sharma 2018). Their study has observed, the independent variables B1 and B1/B2, B2 and B2/B1 and B1/B4 and B2/B4 explain about 93% of the variance in BOD, COD and DO, respectively. It is noted that no research has given any recommendation on sensors or band/band ratio. Many studies in the past have proposed the application of Landsat comparatively promising in conjunction with *insitu* data and statistical techniques. However, the relationship between nonoptical parameters and satellite reflectance are very complicated to model using simple regression equations with Landsat. To address this fundamental issue, indirect estimation of COD, DO and BOD through other indicator variables may serve as a practical solution. The discussions on different band or their combinations correlations obtained in different study are poorly explained in most of the studies. However, the studies cleared the technicality on understanding the correlation between satellite data and *insitu* physicochemical parameters and their mapping. In fact, such studies are highly beneficial to understanding the dynamics of these parameters at a finer spatial and temporal scale, especially over a remote area.

2.5.1.3 Turbidity

Turbidity is an optical property of water that makes light scatter and absorb rather than transmit in straight lines, which is the opposite of clarity. Turbidity in any waterbody can be caused due to the presence of suspended and dissolved matter (Nazeer and Nichol 2015; González-Márquez et al. 2018). Highly turbid water will affect the physical appearance of water. Algae, CDOM and suspended matters are more dominant parameters in most lakes that obstruct the clarity and the satellite responses. The high concentration of turbidity is generally observed in the monsoon season due to the heavy surface runoff and transportation of sediments from soil to the stream (Maillard and Pinheiro Santos 2008; Yuan and Chen 2011). Their concentrations largely depend on catchment geology, climate, topography, vegetation, impoundment, and land use. Turbidity will make the light disperse and absorb rather than transmit in a straight line

in the water column. Its radiance is governed by the size of the particles individually present in the suspension and grain size distribution as well. It was suggested in many studies that the higher the concentration of TSS/turbidity, the higher will be the radiant emergent from the water surface along VIS and NIR of the EM spectrum (Garg et al. 2017; Espinoza-Villar et al. 2018; Santos et al. 2018; Yepez et al. 2018). Vanhellemont and Ruddick (2015) discussed the advantages of using Landsat 8-OLI for high resolution and quality monitoring of coastal sediments and explained the benefits of higher SNR (Signal to Noise Ratio) compared to other Landsat images. A recently published study conducted by Garg et al. (2020) examined the variability of turbidity along the Ganga river basin during the COVID-19 lockdown period without using the ground observation data. The variation in turbidity has been discussed in terms of alterations in reflectance values in the VIS and NIR regions of Sentinel-2A/B. The results have been validated using the normalised difference turbidity index (NDTI) band ratio approach. The red and near-infrared wavelengths were found to be the most sensitive for turbidity analysis. A summary of various WQPs, sensors, bands of interest and methods employed is listed in Table 2.3.

Table 2.3 Summary of some of the WQPs analysed in past research

Parameters	**Sensors used**	**Bands and Methods**	**Reference**
Chl-a and SS	Landsat TM Bands 1, 2,3,4,5 and their combinations	ANN, MLP. TM1, TM2, TM3, and TM4 performed best among all ANN SS models and TM1 and TM2 for ANN Chl-a models.	(Sudheer et al. 2007)
Chl-a and TSS	Landsat5 and ETM+	Based on the red spectral reflectance. Resampled Landsat data to MODIS in a GIS environment using ENVI Software. The statistical model was applied to model TSS and remote sensing images. A significant correlation ($p < 0.05$) for predicted TSS with predicted chlorophyll-a concentration ($R^2 = 0.969$).	(Dalu et al. 2015)
Turbidity and TSS	MODIS and Landsat	Temporal Adaptive Reflectance Fusion Model (STAR-FM). Surface reflectance values of all six bands of fused images. MATLAB ANN toolbox entitled NARXNET	(Imen et al. 2015)
Turbidity ,TSS, COD, BOD and DO	Landsat 8 OLI	ANN and SVM. Seven neurons using bands CB, B, G, R, NIR, SWIR1, and SWIR2. BPNN models, with $R^2 \geq 93\%$	(Sharaf El Din et al. 2017)

		at the network testing phase, are observed for WQPs turbidity, TSS, COD, BOD, and DO concentrations.	
Fe, TSS, Turbidity, Zn, Cu, Cr, Pb and Cd	MODIS and Landsat	STAR-FM-based algorithm to fuse MODIS and Landsat. Bands Red, NIR and combinations of these two. Development of regression models between turbidity concentration and Landsat surface reflectance.	(Swain and Sahoo 2017a)
Turbidity , EC, pH, DO, and depth maps	Landsat-8 OLI	R^2 = 0.6419 with the sum of bands b4 and b5 for T. EC linked to the B2-B3/B4-B6 with R^2 =0.6994. pH with B3, B4, B5, and B6 with R^2 of 0.8153. DO with B1, B3, B4, B5, and B7 presented R^2 = 0.930.	(González-Márquez et al. 2018)
Chl-a	33 years of Landsat from 1984-2017. Bands 1 to 5	33 years of data from 1984-2017. Bands 1 to 5. Normalized difference chlorophyll index (NDCI), cloud computing tool by Google Earth.	(Maeda et al. 2019)
Chl-a, DO, total SS, Secchi disk depth, TDS, and pH	Landsat 8 OLI Sentinel 2A and Göktürk-2 (GK2) satellite sensors.	Principal component analysis (PCA) data fusion and mining techniques. R^2=0.89 for PCA. The first five components represented about 90% of the data used for the band selection. The PCA-based method is superior to MLR, ANN, and SVM data.	(Batur and Maktav 2019)

Blue-green algae, Chl-a, Fluorescent dissolved organic matter DO, specific conductance, and Turbidity	Landsat-8 and Sentinel data fusion	Progressively decreasing deep neural network (pDNN), MLR, SVR and ELR. Fusion increased the temporal frequency required for dynamic systems such as inland water bodies. The pDNN approach significantly outperformed MLR, SVR, and ELR in terms of overall accuracy and error.	(Peterson et al. 2020)
TP, Total Kjeldahl Nitrogen (TKN), TSS, and Chl-a	Landsat TM and Landsat OLI	A strong correlation was observed for TP and TKN with Chl-a, TSS, and selected band ratios. Bands Blue, Green, Red, NIR and combinations. Medium to high R^2 values was observed for non-monsoon and monsoon seasons.	(Hajigholizadeh et al. 2021)

2.6 SUMMARY OF LITERATURE REVIEW

The review provides an overview of conventional river water quality assessment and incorporating remote sensing with traditional techniques. Most published research discusses the significance of having a continuous monitoring program to understand the spatiotemporal water quality of any aquatic water resource. Past literatures discussed about how unsustainable land use practices affect the availability of usable freshwater in terms of both water quality and quantity. Furthermore, the changing relationship between water quality and LULC at different spatial scales is well discussed. Previous research into the relationship between landscape characteristics and water quality used the reach, riparian, and catchment scales to predict water quality. Literature documents numerous natural and anthropogenic factors that influence stream water chemistry. Their impacts on hydrochemistry can be diffuse (for example, runoff from urban and crop cultivation, interflow through organic rich soils) or point pollutants (e.g., industrial effluents). From the research survey, it is found that these relationships are indeed site specific. Some certainly exercised the role of multivariate statistical approaches to identify the spatiotemporal trend in water quality, dominant WQPs causing spatiotemporal changes in water quality, grouping the monitoring stations based on similarity and dissimilarity among them. Different modelling approaches, like statistical and machine learning approaches to identify the relationship between water quality response variables and predictors are also well acknowledged. Many researchers, however, discussed the inability of *insitu* monitoring data to explain the long-term water quality trend on a finer scale. To address this issue, we reviewed articles that used remote sensing techniques in conjunction with *insitu* water quality data. Many researchers discussed the success of empirical methods in identifying the statistical relationships between *insitu* WQPs and remote sensing spectral values. These techniques include statistical regression, curve fitting, nonlinear regression and computational intelligence techniques such as ANN, SVM, Tree-based algorithms, Ensemble algorithms, PSO, and GP models, among others. According to the available literature, Landsat sensors, including the Thematic Mapper (TM), Multi-Spectral Scanner (MSS), Enhanced Thematic Mapper (ETM), and Operational Land Imager

(OLI), have been used fairly successfully to measure physico-chemical and biological WQPs to explain the near real time water quality status of any waterbody.

2.7 RESEARCH GAPS BASED ON LITERATURE REVIEW

Understanding spatiotemporal water quality characteristics and the relationships between (LULC) patterns and water quality is critical for investigating the causes of different source pollution and conducting scientific land-use planning. Many researchers have examined the water quality trend using a variety of WQPs and identified their relationships with various LULC classes. No study has identified the dominant WQPs from a large water quality dataset to find the relationship between the LULC. However, due to the high cost of data collection and laboratory work, the amount of possible *insitu* measurements of WQPs is usually limited, particularly in spatial and temporal domains. At this stage, a long term water quality assessment still lacks in this domain.

In the remote sensing part of the research survey, we observed that most studies estimated WQPs, especially optically active parameters, using remote sensing and regression-based modelling. Theoretically, water quality is too complicated to have a simple relationship with satellite spectral signatures (Yu et al. 2016). Furthermore, regression based approaches fail to describe the complex connection between satellite reflectance and concentrations of various WQPs, particularly non-optical constituents (Sharaf El Din et al. 2017). The best band or combinations suggested for predicting WQPs differ from one study to another due to the optical complexity of turbidity, suspended matter, and productive waters. Besides, the transferability of algorithms developed in one study to other environments remains unknown from the papers reviewed. It has also been found that classical machine learning algorithms are used in inland water remote sensing such as ANN (Teodoro et al., 2007; Liu et al., 2015; Sharaf El Din et al., 2017; Said & Khan, 2021), GA (Chang et al. 2012b; Lounis et al. 2013; Swain and Sahoo 2017a), SVM (Li et al. 2018; Wang et al. 2010), Random Forest (RF)/Boosted Regression Trees (Hafeez et al. 2019; Rubin et al. 2021) and Convolution Neural Network (CNN) for wetland water area (Günen 2022) has shown potential in reliably calculating WQPs at several spatiotemporal scales. However, to the best of our

knowledge, the boosting ensemble algorithm and other deep learning algorithms are rarely used in this domain.

2.8 PROBLEM FORMULATION

The Ganga River Basin (GRB) is India's largest river basin in terms of the catchment area, named India's "National River" in 2008. GRB accounts for 26% of the country's land mass (8,61,404 sq. km), holding approximately 43% of its population (448.3 million as of the 2001 census) and accounts for about 79% of the area. 11 states share the GRB: Uttarakhand, Madhya Pradesh, Rajasthan, Haryana, Himachal Pradesh, Chhattisgarh, Jharkhand, Bihar, West Bengal, and Delhi, which lies between longitudes 73°02' and 89°05' E and latitudes of 21°06' and 31°21' N. The flow characteristics of the river vary considerably throughout the year due to significant temporal variations in precipitation (Namami Gange 2020).

In this study, we have chosen the part of the Middle Ganga Basin (MGB), a stretch from Haridwar to Varanasi of Uttara Pradesh (UP) (Dutta et al., 2020), which in terms of the catchment area is the most significant contributor of pollution. High and low flows mainly cause the basin's water-related problems. Concurrently, increased effluent discharge from industries and urban areas has also contributed to water quality issues in many reaches of the Ganga and Yamuna upstream of Allahabad (India-WRIS). Numerous cities along the Ganga basin yield and discharge substantial quantities of sewage water, the vast majority eventually transported to the river via the natural drainage system. Tanneries in Kanpur, distilleries, paper mills, and sugar mills in the catchments of the Yamuna, Kosi, Ramganga, and Kali rivers are identified as significant polluters (Dutta et al., 2020). The risk of contamination from urban and industrial areas has caused a drastic decline in the quality of Ganga water. The Ganga and its tributaries have become industrial effluent transport channels in addition to city wastewater drain over the years. The usage of river water is classified as follows: Class A water is best suited for drinking without conventional treatment, Class B for outdoor bathing, Class C for drinking with traditional treatment, Class D for wildlife and fisheries, and Class E for recreation and aesthetics, irrigation, or industrial cooling (ENVIS-UP). With a current treatment capacity of about 4,000 MLD, approximately 12,000 MLD of sewage is produced (Consortium of 7 IITs 2013a). Currently, roughly

3000 MLD of sewage is released into the Ganga from Class I and II towns along its banks, necessitating a 1000 MLD more treatment capacity installation (Namami Gange 2020). Furthermore, industrial pollution appears to contribute about 20% of total pollution, but its impact is much more significant because it's toxic and non-biodegradable nature.

2.9 RESEARCH OBJECTIVES

- Evaluation of spatial and temporal variations in WQPs.
  - To establish a correlation between complex *insitu* WQPs and to identify the statistically significant seasonal trend.
  - To identify the similarities or dissimilarities between the sampling sites and to identify the most significant water quality variables responsible for spatial and temporal variations in water quality.
- Assessment of water pollution induced by anthropogenic activities.
  - To model the relationship between LULC pattern and water quality at different spatiotemporal scales.
  - To identify which LULC has the strongest influence on water quality.
- Mapping the concentration of WQPs using Landsat-8 and machine learning algorithms.
  - To propose an appropriate learning-based algorithm to model *insitu* water quality data and satellite-derived reflectance data using Landsat-8 for water quality prediction over inland waterbody at a finer scale.
  - To produce a spatial distribution map for each water quality parameter over each pixel of the selected study area.

CHAPTER 3

STUDY AREA AND DATA COLLECTION

3.1 GENERAL

The study focuses on a portion of the MGB, a stretch of the GRB from Balrampur to Chopan in the Indian state of Uttar Pradesh. The analysis in this study was based on *insitu* data collected from the Middle Ganga Division (MGD I & II) by Central Water Commission (CWC), India, for 20 water quality monitoring stations located in Uttar Pradesh (UP).

3.2 GANGA RIVER BASIN

The GRB is India's largest river basin in terms of catchment area and was named India's "National River" in 2008. The Ganga originates as Bhagirathi from the Gangotri glaciers in the Himalayas at an elevation of about 7010 m in the Uttarkashi district of Uttrakhand and flows for a total length of about 2525 km and empties into the Bay of Bengal via the former main course Bhagirathi-Hooghly. The five Ganges headstreams are the Bhagirathi, Alaknanda, Mandakini, Dhauliganga, and Pindar, which all rise in the mountainous region of northern Uttarakhand state. Tropical and subtropical temperature zones dominate the GRB climate. Summers are hot and humid, while winters are cool. In the Ganga plains, temperatures range from 5° to 25° C in winter and from 20° C to more than 40° C in summer. From July to October, the southwestern monsoon is responsible for most of the rain in the basin (Consortium of 7 IITs, 2013). Rainfall along the basin ranges from 390 to 2000 mm, with an average of 1100 mm, 80% of which falls during the monsoon season (Consortium of 7 IITs 2012, 2013c, 2014a). Agriculture is the most common land use (51%), followed by forest (17%), uncultivated land (14%), and fallow land (8%) (Consortium of 7 IITs 2014b). GRB drains about 26% of the country's land mass (8,61,404 sq. km), houses approximately 600 million populations (nearly half of the Indian population) and contributes 40% of the Indian GDP (Namami Gange 2020). The Ganga's main stem passes through 50 major Indian cities, almost all of which have populations over 50,000. The basin provides more than one-third of India's surface water, with 90% used for irrigation.

Eleven states share the GRB (Figure 3.1 & Figure 3.2) Uttarakhand, Madhya Pradesh, Rajasthan, Haryana, Himachal Pradesh, Chhattisgarh, Jharkhand, Bihar, West Bengal, and Delhi, which lies between longitudes 73°02'E and 89°05' E and latitudes of 21°06'N and 31°21' N.

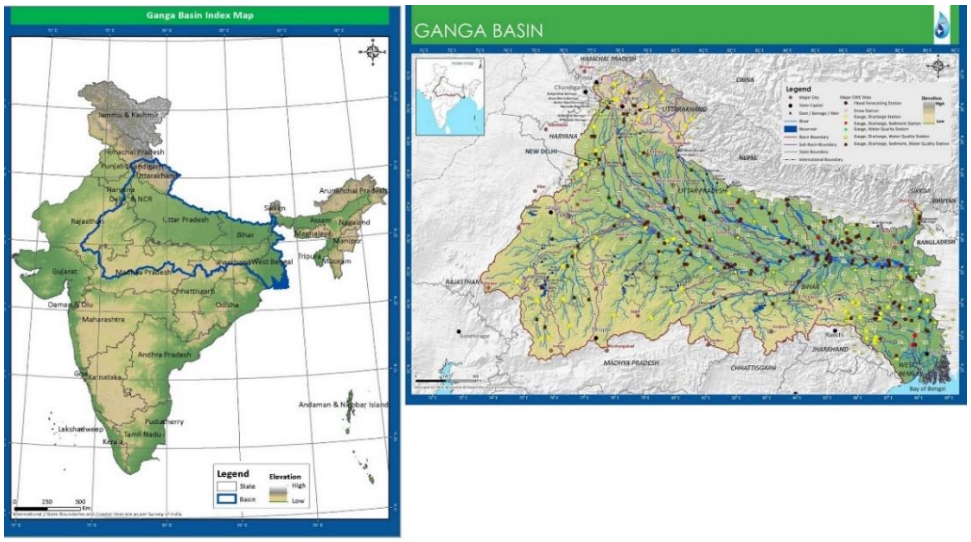


Figure 3.1 GRB Index map, Drainage and Sub-basin

Source: CWC and NRSC (2014)

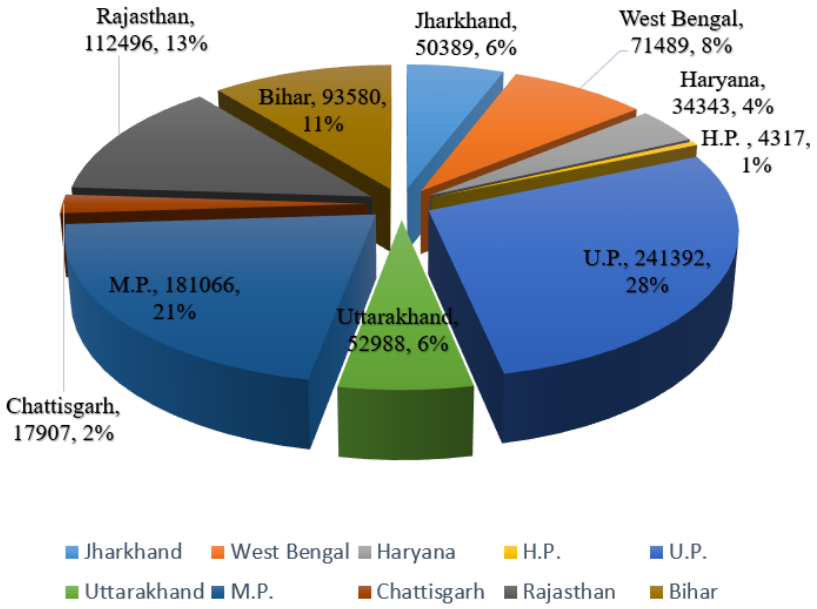


Figure 3.2 State-wise drainage area of Ganga basin (In Indian Territory)

Source: (Consortium of 7 IITs 2014b)

The 'Namami Gange Programme' is an Integrated Conservation Mission initiated as a 'Flagship Programme' by the Union Government in June 2014 with a budget of Rs. 20,000 Crore to achieve the twin goals of effective pollution abatement and conservation and rejuvenation of the National River Ganga. The basin's primary sources of river pollution are urban sprawl, industrialization, and agrarian chemicalisation (Consortium of 7 IITs 2013c). These sources can be classified as point and NPS. Some of the point and non-point source of pollution and their causes along the basin is presented in Figure 3.3. Domestic sewage accounts for 70-80 per cent of wastewater entering the Ganga, with industrial effluent accounting for the remaining 15 per cent. The productivity of the basin varies greatly. Parts of Uttar Pradesh have very high soil productivity, while certain parts of Madhya Pradesh, Rajasthan and Haryana have lower productivity (CWC and NRSC 2014).

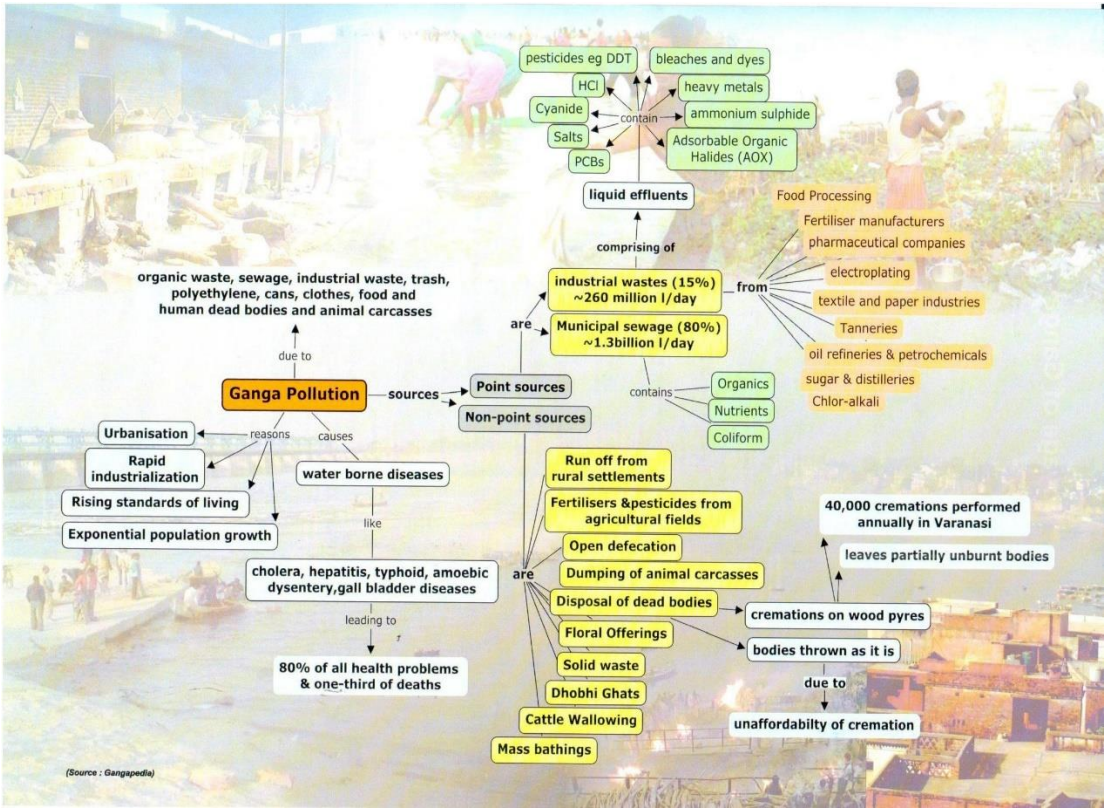


Figure 3.3 Various point and NPS of pollution and their causes along GRB

Source: Namami Gange (2020)

Rapid population expansion, improved living standards, growing urbanisation, and industrialisation all resulted in different types of degradation. The mighty Ganga is no

exception since the decline of water quality directly impacts humans in some places, particularly during the non-monsoon season; the Ganga has become unfit for bathing (Namami Gange 2020).

According to (Consortium of 7 IITs 2013c), the entire basin is divided into three sections: Upper Ganga Basin (Uttarakhand), Middle Ganga Basin (Uttar Pradesh), and Lower Ganga Basin (Bihar and West Bengal). The Himalayan region's upper Ganga Basin is usually considered pollution-free. The Ganga water in the upstream, midstream and downstream areas is used not only for drinking and irrigation but also for deity worship and holy bathing. The results revealed that the values of TDS, EC, alkalinity, calcium, and hardness were highest in downstream places due to this intervention (Dimri et al. 2021). Also, according to studies on groundwater draft in UP, there is a general drop in water level by 2-4m (CWC and NRSC 2014). Given the significance of pollution contribution, a portion of MGB covering 20 monitoring stations was chosen for the study (Figure 3.4 & Figure 3.5).

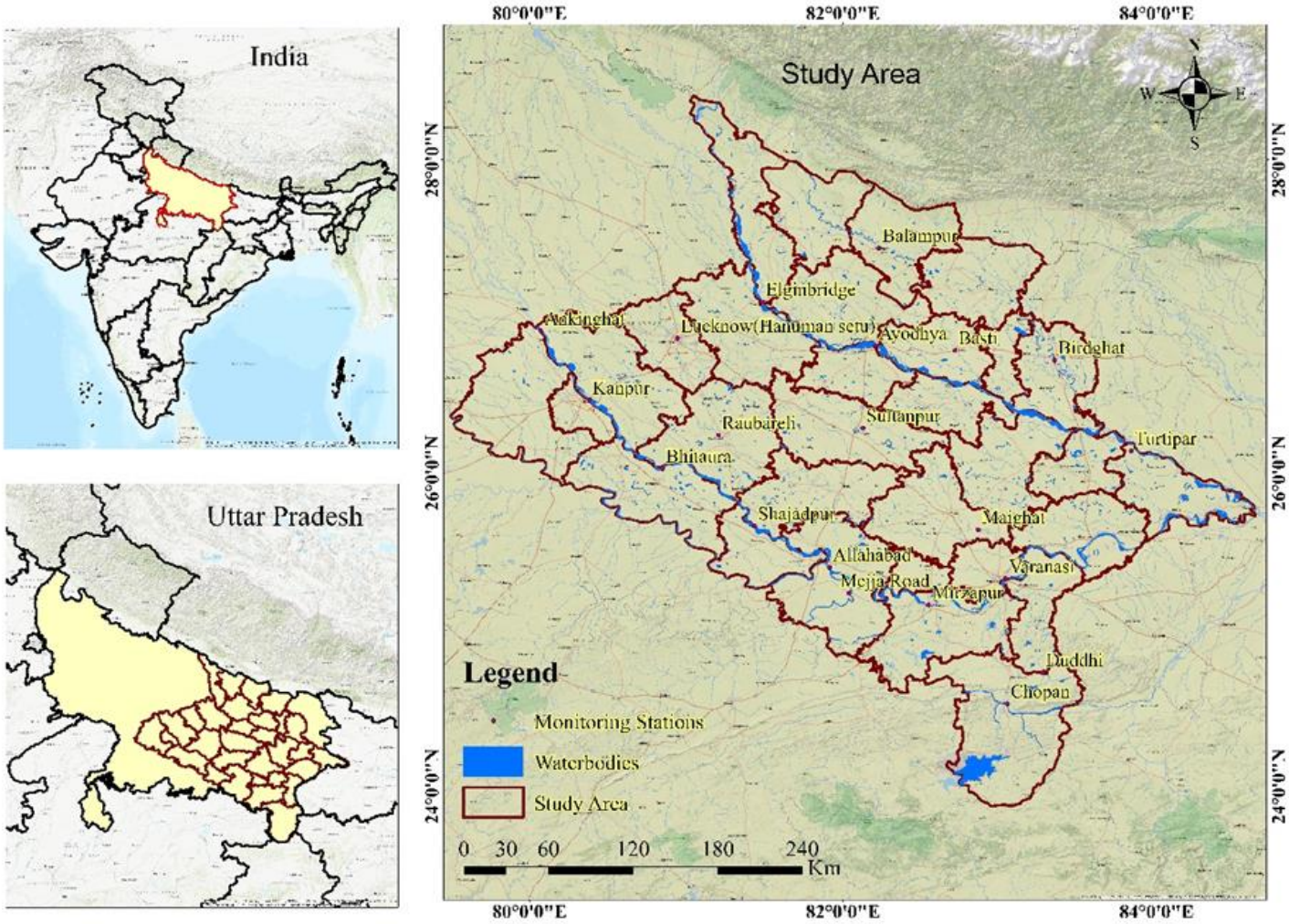


Figure 3.4 Geographical location of the study area

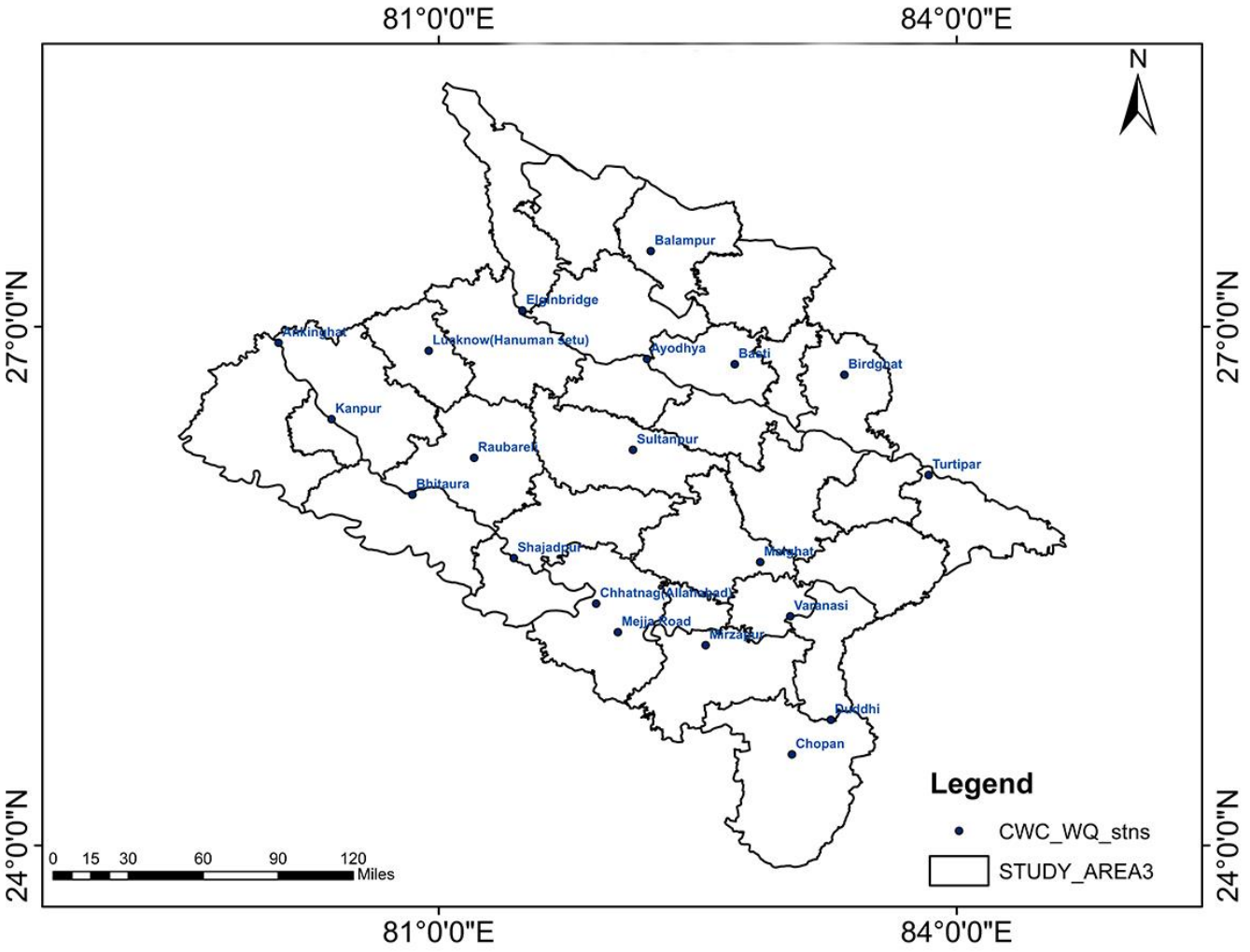


Figure 3.5 Monitoring stations along the parts of MGB

3.2.1 Middle Ganga Basin

The Middle Ganga plains have a transitional climate since they have positioned between the Himalayan area to the north and the peninsular foreland to the south. Winter cyclones sweep the central region, bringing cold waves and hot summer breezes from the west (CWC and NRSC 2014). The MGB, which includes the entire state of Uttar Pradesh, is one of India's most populous regions, with a population of around 200 million (as of the 2011 Census). Uttar Pradesh covers a massive portion of the very fertile and heavily inhabited upper and middle Gangetic plains, with an area of 236,286 square kilometres. (Consortium of 7 IITs, 2013).

As per Census 2011, the state has seven cities with populations higher than one million and 16 cities larger than five lakhs. The total urban population increased from 34.54 million in 2001 to 44.47 million in 2011 (Consortium of 7 IITs 2013c). Greater urbanization and river proximity to major cities such as Kanpur, Ghaziabad, Meerut, Gautam Budh Nagar, Agra, Aligarh, Allahabad, and Varanasi have a significant impact on river quality and quantity (Consortium of 7 IITs 2013c). The total of towns and cities had already risen from 704 in 2001 to 915 in 2011. The overall urban population has increased from 34.54 million in 2001 to 44.47 million in 2011, representing a compound annual growth rate of 2.56% (Consortium of 7 IITs 2013c). Furthermore, sewage quantities in certain religious and culturally significant cities and towns rise substantially throughout festivals (Refer Figure 3.6 Varanasi City along the banks of Ganga Course).

The state has the most Micro, Medium, and Small Enterprises (MSMEs) Table 3.1 in India, accounting for 14.20% of the country's total MSMEs.

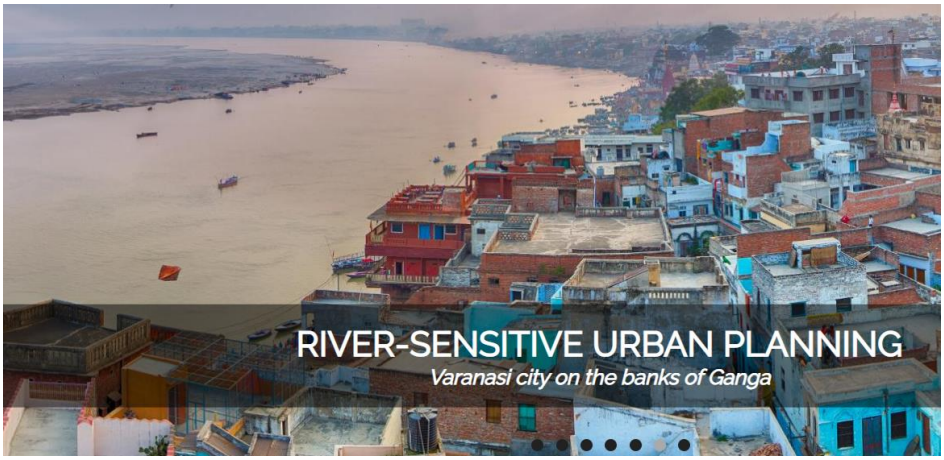


Figure 3.6 Varanasi City along the banks of Ganga

Source: Namami Gange, (2020)

Table 3.1 Key Industrial sectors at different zones under MSME.

Zone	**District**	**Key Sector**
Western	GB Nagar, Ghaziabad, Meerut, Saharanpur and Aligarh	Food processing, Electronics and metals
Northern	Amroha, Bijnor and Moradabad	Chemicals and basic metals
Eastern	Sonbhadra, Allahabad, Varanasi and Gorakhpur	Chemicals and basic metals
Southwestern	Agra and Firozabad	Leather and metal products
Bundelkhand Zone	Jhansi and Chitrakot	Aerospace and defence

Source: Misra (2015)

Municipal sewage is estimated to contribute approximately 80% by volume of the total wastewater disposed into the Ganga, while industries contribute around 15%. Over time, the urban population has risen considerably, while municipal sewage treatment facilities have remained insufficient (Table 3.2). Metropolitan, with a population of one million or more, are included in the Census 2011 for classifying the urban settlements. Class-I towns have a population ranging from 1 lakh to 10 lakhs; Class-II, 50,000 to 1 lakh; Class-III, 20,000 to 50,000; Class-IV, 10,000 to 20,000 and Class-V 5,000 to

10,000; Class-VI, 3,000 to 5,000 (Consortium of 7 IITs 2013c). Figure 3.7 depicts a list of Class I and Class II cities and their proximity to the river.

Table 3.2 Wastewater generation and treatment in Uttara Pradesh based on Population 2001

SL. No.	City/Town	Population 2001	Total Sewage generation (in MLD)	Treatment Capacity (in MLD)	Percentage covered
Class I					
1	Kanpur	3114530	339.3	171.1	50
2	Varanasi	1353920	187.1	141	75
3	Allahabad	1218070	208	89	43
4	Farrukhabad-cum-Fatehgarh	280290	30.5	8.3	27
5	Mirzapur-Vindhyachal	252470	27.5	14	51
6	Unnao	178250	23.9	19.4	81
7	BaIlia	125740	18	-	0
8	Dehradun	550800	76.1	-	0
9	Hardwar	215260	39.6	18	45
Class II					
10	Bijnor	79368	7.6	8.1	100
11	Mughalsarai	88386	16	-	0
12	Ghazipur	95243	10.7	-	0
13	Kannauj	71530	7	-	0
14	Deoband	81706	7.8	-	0
15	Gangaghat	70817	6.8	-	0
16	Rishikesh	59671	10.7	6.3	59
17	Roorkee	97064	11	-	0

Source: (Consortium of 7 IITs 2013c)

The study performed during the lockdown period reported a gradual transformation of water quality from the restoration point. However, this improvement in water quality is believed to be 'short-lived.' That quality will deteriorate once normal industrial activities resume, indicating a strong influence of untreated commercial-industrial wastewater (Dutta et al. 2020).

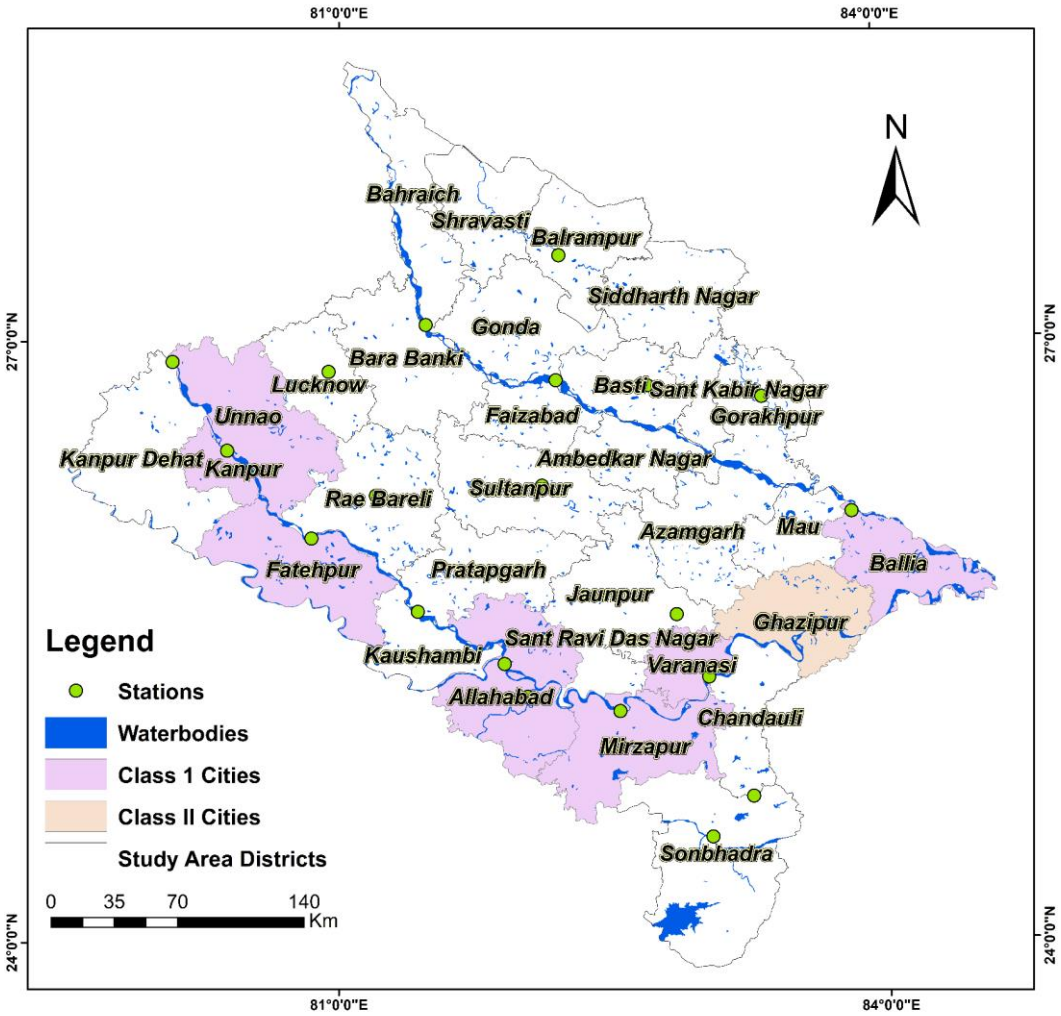


Figure 3.7 Class I and Class II cities along the study area

Source: Consortium of 7 IITs (2013a)

3.3 DATA COLLECTION AND PREPROCESSING

The data collected in this study are presented in two categories Table 3.3: Spatial and Non-Spatial datasets.

Table 3.3 Spatial and Non-spatial data

Data	Descriptions	Resolution	Year
Spatial Data			
DEM	Shuttle Radar Topography Mission (SRTM) Credits-USGS	30m	2005, 2009, 2015 and 2018
LULC	Published LULC - Bhuvan Portal, Indian Space Research Organisation		2005, 2009, 2015 and 2018
Non-Spatial Data			
***Insitu* WQPs**	Middle Ganga Division (MGD) (I & II), Central Water Commission (CWC)	Monthly data/20 stations	2005-2018

3.3.1 Spatial and Non-spatial dataset

Using ArcGIS© 10.2.1 geoprocessing tools, the global SRTM DEM was pre-processed for filling sinks in the dataset. For LULC analysis, Landsat images are pre-processed, cloud-removed, and mosaicked using Google Earth Engine (GEE). Published LULC maps from the Indian Space Research Organization (ISRO) on Bhuvan Portal were then used as a reference to improve the LULC classification (Kumar Shukla et al. 2018). For ground truthing of prepared LULC maps, ground control points (GCPs) were collected from Google Earth images. The SWAT module and ArcGIS© (10.2.1) are employed to delineate three geographical scales: watershed, riparian, and reach. The catchment scale encompasses the entire upstream of monitoring stations, a bandwidth of 1000m on each side extended upstream above all monitoring sites considered for riparian and 500m upstream of all monitoring sites considered for reach scale (Shi et al. 2017).

The non-spatial dataset is pre-processed using Anaconda (conda 4.7.10) and the Numpy, Pandas, Matplotlib, and Scikit Learn libraries. Because the different variables of this study are measured at distinct units, scales were treated by z-transformation in Python pre-processing tool after data scaling. The whole dataset used in the study is grouped into two seasons: non-monsoon (November-May) and monsoon (June-October) (CWC and NRSC 2014), and statistical analyses were run for each season (Hajigholizadeh and Melesse 2017b).

CHAPTER 4

METHODOLOGY

4.1 GENERAL

This chapter describes the statistical and machine learning methodologies for *insitu* WQPs integrated with remote sensing. A summary of descriptive and inferential statistics used in data analysis is provided. It also describes the various machine learning approaches used to analyse river water quality. The overall methodology chart (Figure 4.1) and methodology for each objective are explained separately under spatial and non-spatial categories.

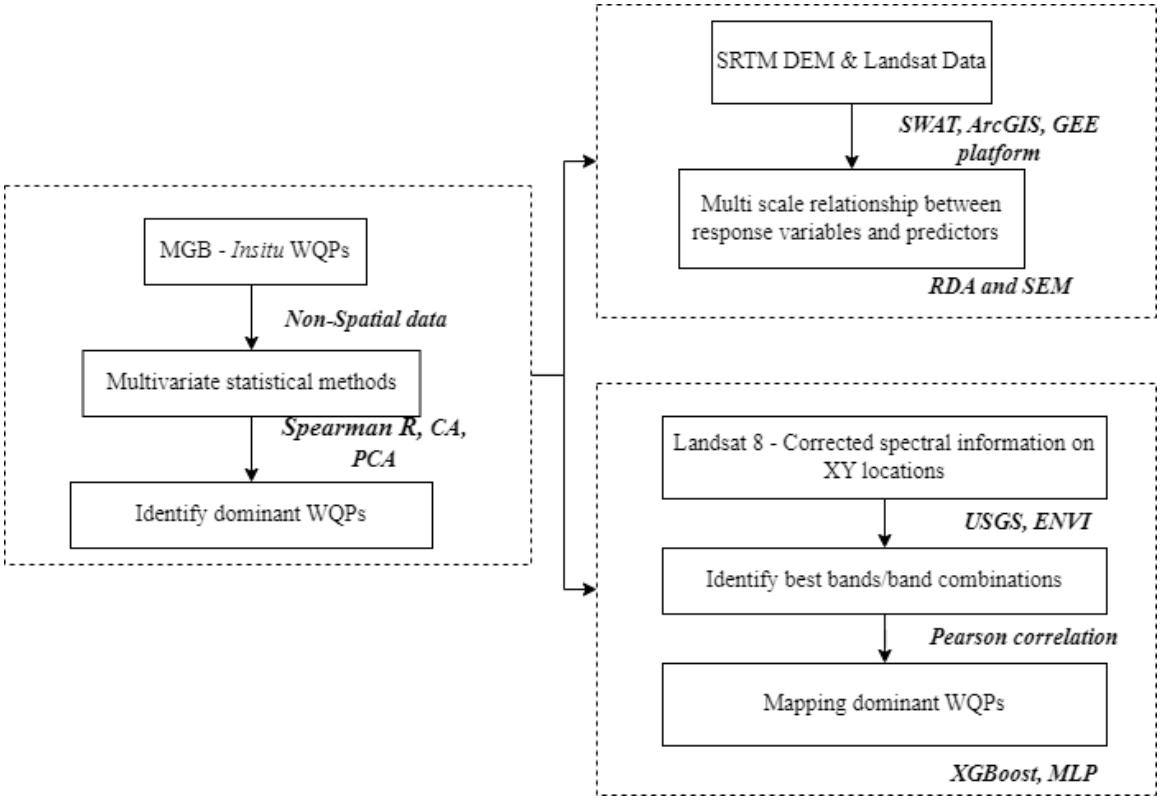


Figure 4.1 Overall methodology chart of the study

4.2 ASSESSMENT OF SPATIOTEMPORAL VARIATIONS IN WQPS

This section explains the part of 1st objective, which focuses on non-spatial analysis, explicitly using different multivariate statistical approaches. Multivariate data processing can be applied to evaluate temporal and spatial variations of water quality (Sundaray et al. 2006; Azhar et al. 2015) caused by natural and anthropogenic factors

(Singh et al. 2004; Bhat et al. 2014). The below section explains the type of multivariate techniques adopted in the present study (Figure 4.2).

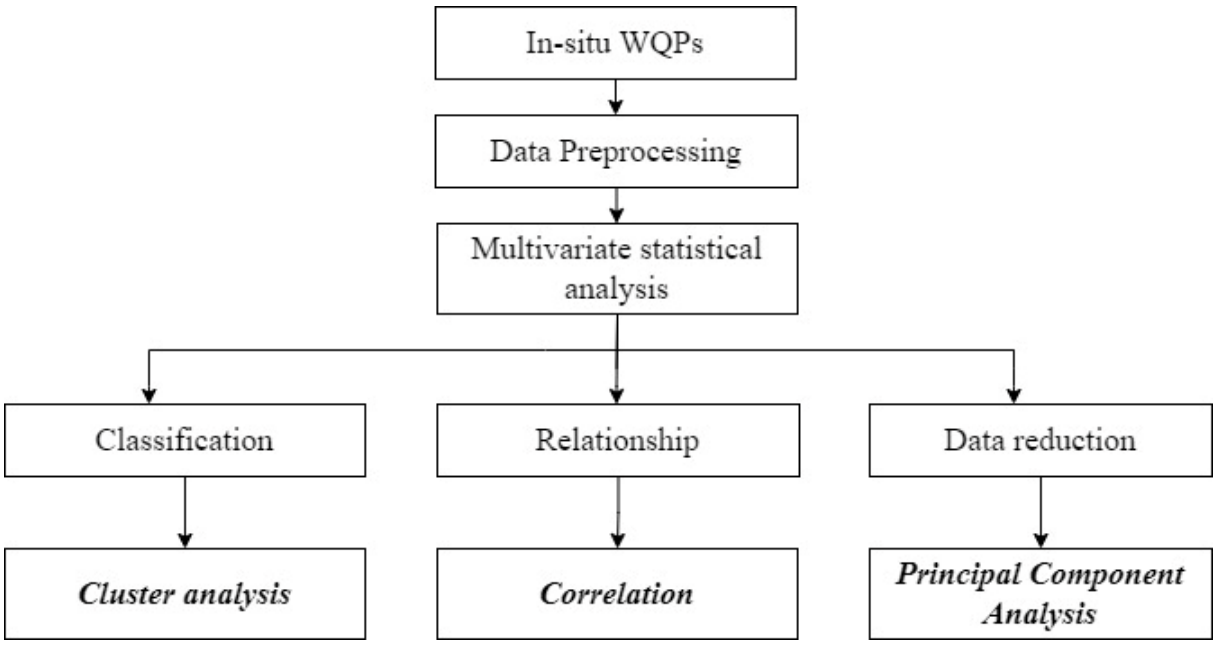


Figure 4.2 Concept map spatiotemporal variations in WQPs.

4.2.1 Data Pre-processing

4.2.1.1 Normality Test

A normality test evaluates whether or not a sample was drawn from a population with normal distribution. It is typically used to determine whether the data used in the study has a normal distribution. The concept of normal distribution supports many statistical methods, including parametric tests, correlation, regression, t-tests, and ANOVA. The mean, median, and mode values are the same in a perfectly normal distribution, and they clearly show the peak of the curve. There are two broad approaches for determining whether data are normally distributed or not. Graphical (histograms and Q-Q probability plots) and analytical (such as the Shapiro-Wilk, D'Agostino-Pearson test and Kolmogorov-Smirnov tests). In the present study, D'Agostino-Pearson and Shapiro-Wilk test was applied.

The D'Agostino-Pearson test is a highly effective and adaptable technique for identifying non-normality induced by skewness and kurtosis. It tests statistics by comparing the sample data's kurtosis and skewness coefficients with the moments of a normal distribution measured by Pearson's coefficients as defined in equation (4.1).

These statistical tests share the null hypothesis that the *insitu* water quality data was drawn randomly from a normal distribution. A statistically significant *p-value* (usually 0.05 or 5%) gives strong proof against the null hypothesis, showing a non-normal sample distribution.

$$K^2 = Z^2\left(\sqrt{b_1}\right) + Z^2\left(\sqrt{b_2}\right) \qquad (4.1)$$

Where,

$Z\left(\sqrt{b_1}\right)$ and $Z\left(\sqrt{b_2}\right)$ - normal approximations to test skewness.

$\sqrt{b_1}$ and $\sqrt{b_2}$ - Test of kurtosis

K^2 - Statistic has a chi-squared distribution with two degrees of freedom when the population is normally distributed.

The D'Agostino-Pearson test statistic combines the benefits of skewness and kurtosis tests to provide an omnibus normality test (Omnibus refers to the capacity to identify variations from normalcy caused by skewness or kurtosis).

Shapiro-Wilk (S-W) test will reject the normality hypothesis if *p* is equal to or less than 0.05. Unless the test fails, the test can state with 95% confidence that the data will not fit the distribution usually. However, if the test is passed, the test can state that there is no significant deviation from normality.

The S-W test was initially designed to assess the normality of univariate distributions. To begin, arrange the given univariate dataset *Y*= $y_i, \ldots y_n$, in ascending order and proceed as shown in equation (4.2).

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{nS^2} \qquad (4.2)$$

where the a_i - *S-W* coefficients, and S^2- statistical variance of the sample. W- a statistic which requires a sample size ≤ of 7 and ≥ 2,000 (Shapiro and Wilk 1965)

4.2.2 Classification

Cluster analysis is essential for comprehending various phenomena and investigating the characteristics of obtained data. Clustering identifies data groups that are similar to one another. It divides the data into similar groups ensuring that the distance between two instances is identical if they belong to the same cluster and far if they belong to

different clusters (Kotekani and Ilango 2022). Distance metrics are crucially significant in the clustering process. The greater similarity between the data throughout clusters, the more likely those specific data items will belong to that particular group. It is essentially an unsupervised learning method. Unsupervised learning is an approach that gathers references from dataset of input data without labelled responses. There are various types of clustering algorithms. The centroid-based process is one of the iterative clustering algorithms in which clusters are formed based on the proximity of the dataset to the cluster centroid. The cluster centre or centroid is included in this scenario so that the distance between data points and the centre should be as minimal as possible. K-means is the most popular clustering algorithm among all. This algorithm minimises the sum of squared errors, which is the objective function. The algorithm tries to identify 'K' clusters by satisfying specific clustering criteria (Kulluk et al., 2023). The Euclidean distance metric was used in the present study for implementation, which consists of positive real values.

Let (p_1, p_2) be the Cartesian coordinate of p in a Euclidean plane, and q have coordinates (q_1, q_2). Then Euclidean distance between these two can be written as given by equation (4.3)

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \qquad (4.3)$$

The above equation's accuracy depends entirely on the chosen initial seeds. The number of clusters is fixed based on the cluster quality using intrinsic methods like silhouette score, which is based on the silhouette coefficient. This method generates a concise, pictorial depiction of how perfectly each object fits into its cluster. A flowchart on the working of K- means clustering is presented in Figure 4.3.

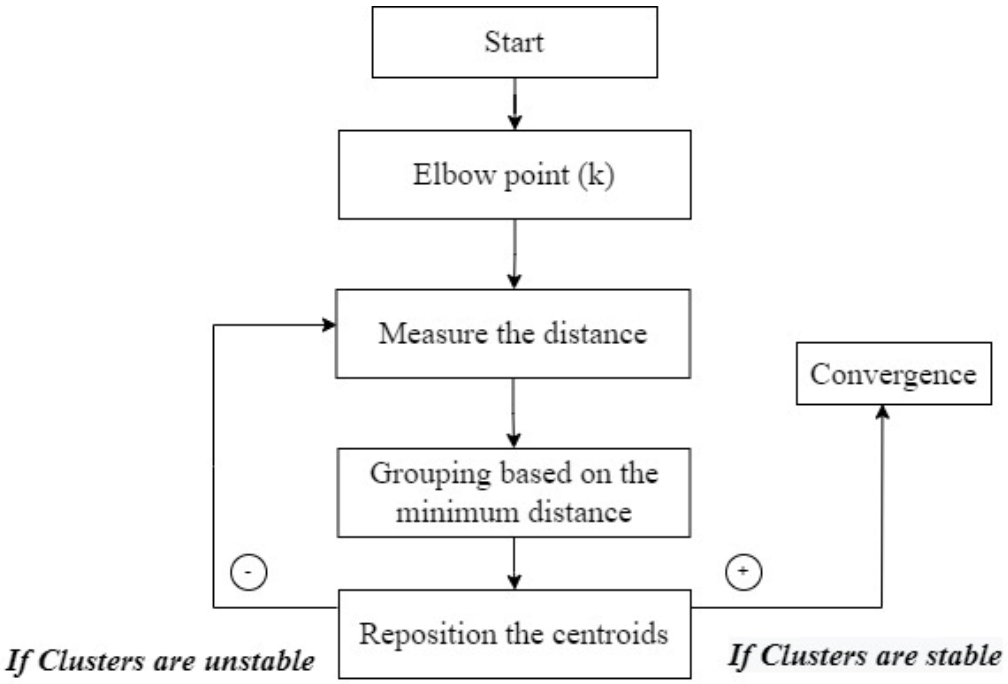


Figure 4.3 Working of K- means clustering

The Elbow method is the most effective way to determine the number of clusters. Within-sum-of-squares (WSS) is applied as a metric to determine the optimal number of clusters for the given dataset as defined in equation (4.4). *WSS* is calculated by adding the squared distances between each cluster member and its centroid.

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2 \qquad (4.4)$$

$x_i -$ Data points and $c_i -$ Closest point to the centroid

The algorithm begins with K initial seeds. The Euclidean distance is used to compare all *n* data to each seed, and the closest cluster seed is assigned. The method is then repeated until convergence is reached.

4.2.3 Relationship

Correlation is a statistical measure that expresses how closely two variables are related linearly. Because the measured WQPs are not normally distributed, the Spearman *r* coefficient is used to investigate the correlation structure between the variables. It is a non-parametric measure used to evaluate the correlation between variables derived from ranked data (Wunderlin et al. 2001; Singh et al. 2004). Spearman *r* is defined as the Pearson correlation coefficient but calculated over ranks (the values of variables

arranged from the smallest to the largest). The test identifies the steady decrease or increase in the importance of one random variable with the same changes in another random variable, called monotonically increasing or decreasing. When both variables are proportionally increasing/decreasing, a positive (+) correlation exists. Suppose one falls when the other increases or vice versa, implying that the variables are negatively (-) correlated. The result lies between -1 and +1.

The Spearman rank correlation coefficient can be calculated as discussed in equation (4.5) (Antonopoulos et al. 2001).

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(DiDi)}{n(n2-1)} \qquad (4.5)$$

$\rho = \text{Spearman}'\text{rank correlation coefficient}$

$Di = \text{Difference between the two ranks of each observation}$

$n = \text{number of observations}$

The WQPs were categorized into two seasons (monsoon and non-monsoon) and assigned a numerical value in the data file (monsoon = 1, non-monsoon = 2), which was correlated (pair by pair) with each of the measured parameters (Wunderlin et al. 2001; Singh et al. 2004).

4.2.4 Data reduction

Before implementing any machine learning algorithm, it is crucial to obtain only relevant features in the training dataset, i.e., reducing the dimensions of feature space is called dimensionality reduction. Besides, it further helps prevent overfitting, making the model simple and efficient with less running time. PCA and linear discriminant analysis (LDA) are two data pre-processing linear transformation techniques frequently used for dimensionality reduction and feature selection. PCA is a statistical unsupervised machine learning approach that employs an orthogonal transformation to convert a set of correlated variables to uncorrelated variables (Li and Liu 2018). PCA is the most widely used unsupervised dimension reduction tool in exploratory data analysis in machine learning techniques for predictive models.

Moreover, as an unsupervised statistical technique, it examines the interrelations among variables. Each of the principal components (PCs) is chosen to describe most of the available variance, and these PCs are orthogonal to each other. Out of all PCs, the first

principal component has a maximum variance. PCA sorts out the data into groups of variables with the strongest correlations as reflected by the factor score. In other words, the groups, otherwise called PCs, are grouped or ranked in the order of total variance explained.

PCA is implemented in the Scikit Learn library in Python. Before using PCA in Scikit Learn, it is necessary to standardize/normalize the data. PC can be mathematically derived (Dash et al. 2018) using the following steps

Step 1: Calculate the covariance matrix for the normalized data as in equation (4.6).

Covariance between feature vectors $\sigma_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_j^i - \mu_j\right)\left(x_k^i \mu_k\right)$ (4.6)

x_j and x_k are the two feature vectors, and σ_{jk} is the covariance. Covariance quantifies how two features differ from one another. A positive covariance indicates that features change together. A negative covariance indicates that the two characteristics vary in opposite directions.

Step 2: Eigen values and Eigen vectors.

The eigenvectors are the principal components with the directions of maximum variance of a covariance matrix. The eigenvalues are the magnitudes that correspond to them. The eigenvector with the largest corresponding eigenvalue represents the direction of the highest variance.

An eigenvector v satisfies the condition $\Sigma v = \lambda v$ where λ is a scalar, also known as eigenvalues.

Step 3: Selecting the Principal Components (PCs)

The calculated eigenpairs are arranged based on the magnitude of their eigenvalues. Plotting the cumulative sum of the eigenvalues can be used to determine the number of PCs to be chosen for dimensionality reduction.

The cumulative sum is computed using the equation following (4.7):

$$\text{Cumulative sum} = \frac{\lambda_j}{\sum_{j=1}^{d} \lambda_j} \quad (4.7)$$

Feature selection and dimensionality reduction task is crucial for high dimensional machine learning analysis to select dominant features in training the dataset. Besides, these techniques can be further helpful in preventing overfitting, simplifying the model to improve the computational efficiency and reducing the algorithm's overall running time.

4.3 WATER POLLUTION INDUCED BY ANTHROPOGENIC ACTIVITIES

4.3.1 General

The dominant WQPs causing spatiotemporal variability identified in previous sections were used to determine water pollution caused by anthropogenic activities. The anthropogenic activities in this study are limited to LULC changes. The methodology applied for the various models involved is discussed separately for spatial and non-spatial analysis Figure 4.4.

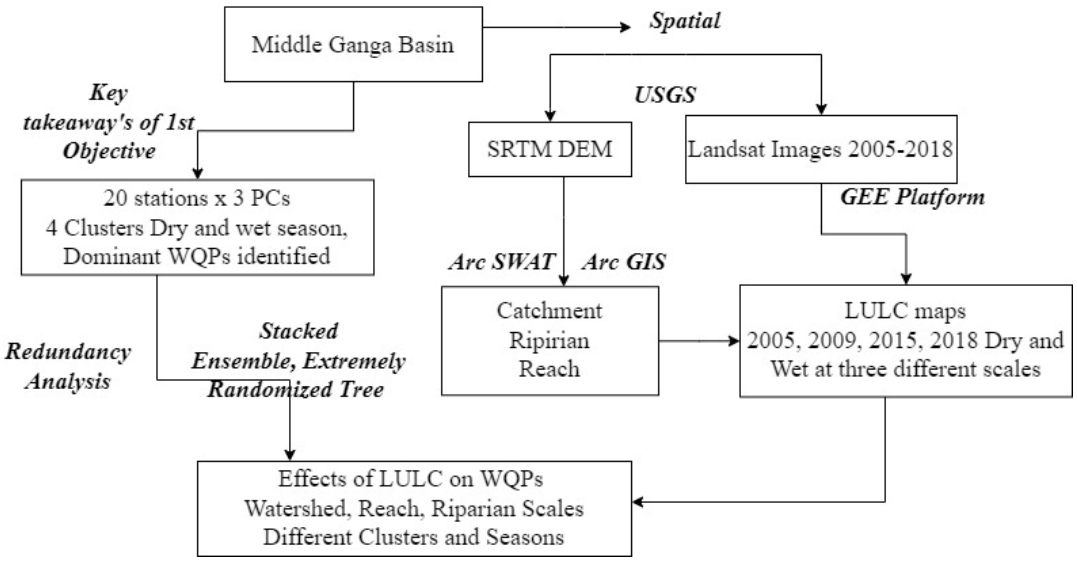


Figure 4.4 Concept map of water pollution induced by anthropogenic activities

4.3.2 Satellite data and LULC classification in the GEE platform

Understanding the changing LULC geographical distribution over large areas is extremely important for many environmental and monitoring tasks, including climate change, ecosystem dynamics analysis, food security, and others (Shelestov et al. 2017; Wang et al. 2020). With the availability of cloud-based platforms such as Google Earth Engine (GEE) (https://code.earthengine.google.com/), it is now feasible to monitor LULC spatiotemporally to study the global dynamics (Liu et al. 2020; Tamiminia et al. 2020). GEE is a cloud-based platform for the scientific analysis and visualization of petabyte-scale geospatial dataset. It stores several decades of historical images and scientific dataset and enables parallel computing and feasible big data processing in a large study area. In addition to the availability of an extensive repository of raw

remotely sensed imagery, users have access to pre-processed, cloud-removed and mosaicked images in the GEE data catalogue (Tamiminia et al., 2020).

Furthermore, Google provides a cloud-based platform called Google Cloud Platform (GCP) that offers a wide range of services for high-speed parallel processing and machine learning. GCP provides access to robust computational infrastructure, including virtual machines and clusters, as well as a library of APIs for machine learning and other tasks. The GEE platform supports several programming languages, including JavaScript and Python, users can access these tools and data through the GEE API and use them to develop their own applications. It has an extensive archive of satellite imagery and other earth observation data, including Landsat and Sentinel-2, which can be used for LULC classification. GEE provides a variety of pre-trained machine learning models for LULC classification and the ability to train custom models using the platform's built-in tools and APIs. Users can also access a library of pre-built scripts and algorithms for LULC classification tasks.

A graphical engine interface can acquire high-quality Landsat images from the GEE. For this study, two Landsat time series images, namely, Landsat-7 (2005, 2009) and Landsat-8 (2015, 2018) Top of Atmosphere (TOA) for non-monsoon and monsoon seasons, were used from GEE Landsat collection and pre-processed. The presence of the cloud is identified and removed using the *Fmask* cloud and shadow matching algorithm. All indices were calculated with the GEE Code Editor, and all images were clipped to a typical geographic extent before being assembled into a ready-to-use time series dataset.

GEE's Classifier package handles supervised classification using conventional machine learning algorithms in Earth Engine. The general classification workflow is as follows:

- Gather training data. Assemble features with a property that stores the known class label and properties that store numeric values for the predictors.
- Create a classifier object. If necessary, configure its parameters.
- Using the training data, train the classifier.
- Classify a feature collection
- Estimate classification error using data from independent validation.

Three classifiers, namely CART, SVM, and RF, were tested for the classification in GEE. These classifiers have a solid methodological foundation and are commonly used

in land cover and forest mapping applications (Koskinen et al., 2019). In this study, RF, CART and SVM classifiers algorithms were applied using the following code in GEE.

- RF:*ee.Classifier.smileRandomForest(numberOfTrees,variablesPerSplit,minLeafPopulation,bag-Fraction,maxNodes,seed*),
- CART: *ee.Classifier.smileCart(maxNodes,minLeafPopulation*)
- SVM:*ee.Classifier.libsvm(decisionProcedure,svmType,kernelType,shrinking,degree,gamma,coef0,cost,nu,terminationEpsilon,lossEpsilon,oneClass)* (Kulithalai Shiyam Sundar and Deka 2021).

The parameters for RF (Breiman 2001), CART (Breiman et al. 1984) and SVM (Hsu et al. 2003) are selected based on this literature. To train and validate the dataset, *sample* () syntax is applied. Each class is trained with 80-100 ROIs for classification and validated with 30-45 ROIs. It was additionally ensured that the data was distributed in a normal and spectrally pure manner (Kulithalai Shiyam Sundar and Deka 2021). The error matrix was created for the selected years during the non-monsoon and monsoon seasons to determine classification accuracy. The error matrix identifies various matrices such as overall accuracy, consumer accuracy, producer accuracy, and kappa statistics. The *errorMatrix* () function is called on the classified Feature Collection to obtain the confusion matrix representing the validation accuracy. Finally, based on the classification accuracies chosen for conceptual modelling, the best-performing classifier is determined. Finally, the best-performing classifier is determined based on classification accuracies selected for the conceptual modelling.

4.3.3 Delineation of Watershed, riparian and reach zone

Geographic Information System (ArcGIS© 10.2.1) was used to assess the spatial spread of various LULC classes at three distinct spatial scales: 1) Catchment, which includes the whole upstream section above the monitoring station 2) The riparian scale contained a band width of 1000 m on both sides that extended upstream above all sample locations (Shi et al. 2017) and 3) Reach buffer (reach) is a 500-m-wide area on either side that extends upstream above the monitoring site (Figure 4.5). These parameters were chosen based on past research and data resolution. Automatic watershed delineation using ArcSWAT/ArcGIS© was employed to delineate sub-watersheds based on an automated procedure using Digital Elevation Model (DEM). The Soil and Water Assessment Tool

(SWAT) outlines the catchment areas in the ArcGIS© 10.2.1 platform. The catchment's riparian and reach regions were digitised using a GIS buffering tool.

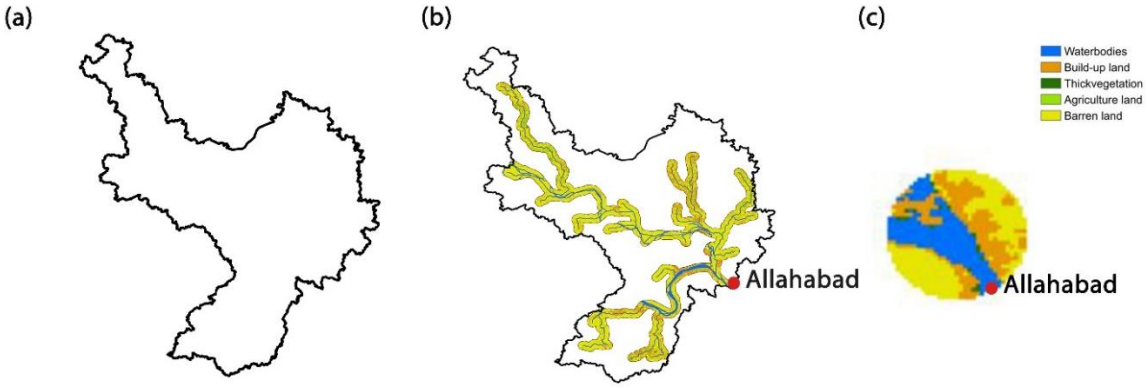


Figure 4.5 Spatial distribution of catchment (a), riparian (b), and reach (c) at Allahabad station.

4.3.4 Redundancy Analysis (RDA)

As a constrained ordination, RDA can determine how much variation in one set of variables can be explained by variation in another set of variables. It is a direct extension of multiple regression, which models the effect of an explanatory matrix *X (n* x *p*) on a response matrix *Y (n* x *m),* i.e., modelling the impact of an explanatory matrix on a response matrix rather than a single response variable. The Pearson correlation coefficient was used to evaluate the statistical significance of the association between LULC classes and water quality variables at $p < 0.01$ and $p < 0.05$ levels (2- tailed), respectively (Mello et al., 2018). To accurately perform the correlation test, the one-sample Kolmogorov-Smirnov test was used to check the variables' homogeneity and normal distribution. In this study, RDA was implemented to assess the influences of LULC patterns on WQPs while accounting for the watershed, reach, and riparian zone. The RDA function from R's vegan package was used in the present study. It consists primarily of two steps. The first step is multiple regression analysis. Each *Y* parameter is registered on an explanatory variable *X* (Figure 4.6). This produces a matrix of fitted values known as $Y_{fit,}$ which can be calculated using the below steps.

X- Matrix of explanatory variables

Y- Matrix of response variables

The following steps were performed as in (Borcard et al. 2011)

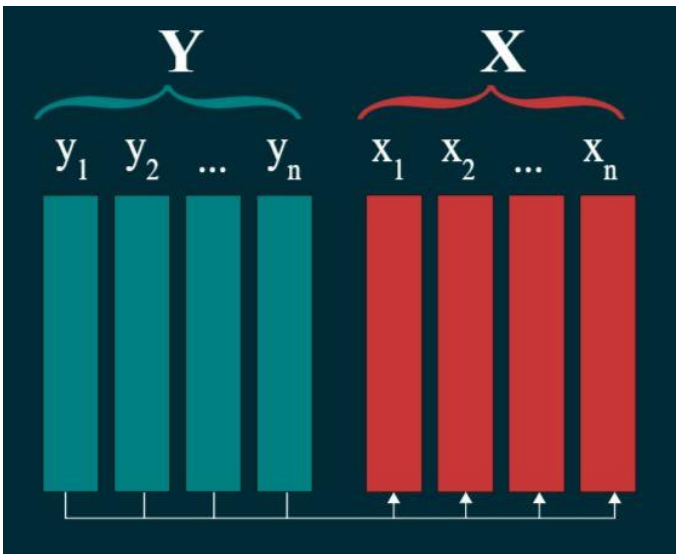


Figure 4.6 RDA Explanatory and Response Variables

Step 1: Each (centred) *y* variable is regressed on the explanatory matrix *X*, and the fitted ($\hat{y}$) and residuals (y_{res}) vectors are computed.

Step 2: Make a new matrix ($\hat{Y}$) that contains all of the fitted vectors ($\hat{y}$).

Step 3: Perform a PCA on $\hat{Y}$. It produces a vector of canonical eigenvalues as well as a matrix *U* of canonical eigen vectors (PC).

Multiple linear regression between *X* and each y_i is used to generate $\hat{Y}$ (Figure 4.7).

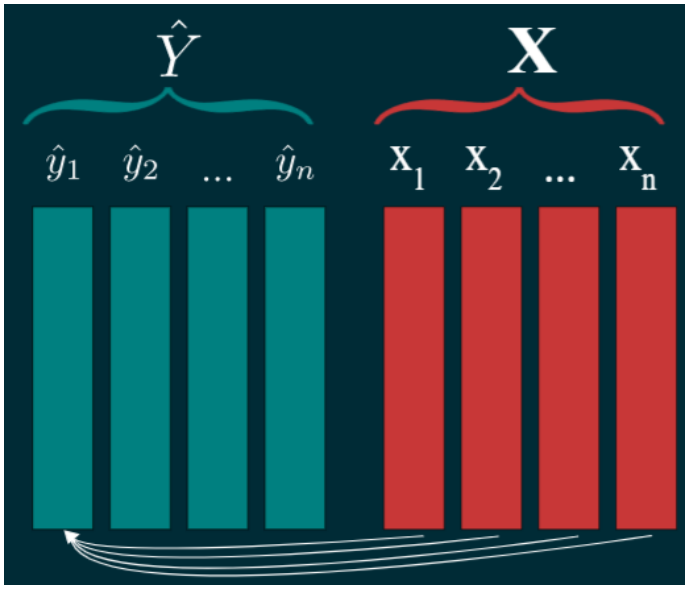


Figure 4.7 Multiple Regression between *X* and each $\hat{Y}$

Step 4: A PCA is run on $\hat{Y}$, yielding a set of principal component vectors *U* (Figure 4.8).

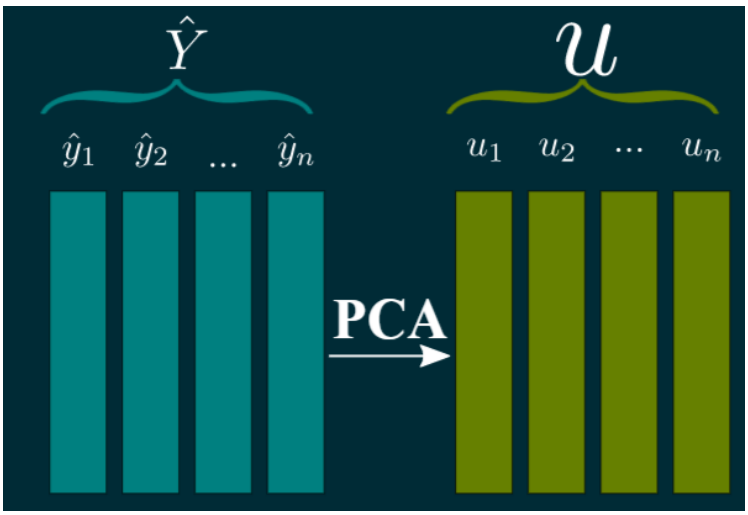


Figure 4.8 Generation of principal components

The primary distinction between PCA and RDA is that PCA is used to a matrix of explanatory variables, whereas RDA is applied to a set of predicted explanatory variables.

4.3.4.1 Evaluation matrix for RDA

The evaluation matrix R^2 is unadjusted and thus biased as the relative contribution of each eigen vector. Therefore, it is preferable to use adjusted R^2 to compute the fair values of R^2 using the function R^2 *Adjusted ()* as given by equation (4.8).

$$R^2_{adj} = 1 - \frac{n-1}{n-m-1}(1 - R^2) \qquad (4.8)$$

Where R^2- is sample, *n*- Number of samples, *m*- Number of predictors

Because ecological data is often non normally distributed, parametric tests are inefficient. In canonical studies such as RDA, permutation tests establish model significance. A permutation test is designed to establish a reference distribution of the selected statistic under the null hypothesis H_0 by randomly permuting suitable data items multiple times and recalculating the statistic each time. The statistic's actual value is then compared to the reference distribution. For a one-tailed test in the upper tail, such as the F test used in RDA, the *p-value* is determined as the proportion of permuted values equal to or larger than the actual (unpermuted) value of the statistic (Borcard et al. 2011). A permutation test is requested in R's Vegan package. The results are first displayed to determine whether or not the relationship between them is significant. The

statistical validity of the RDA was determined using a Monte Carlo permutation test in this study (999 permutations) (Mello et al. 2018). The degree of correlation in RDA is represented by two arrows pointing in the same direction. The angle formed by these arrows is inversely proportional to the degree of correlation. Moreover, the length of the arrow represents the degree of similarity between contributions (Ding et al. 2016; Shi et al. 2017; Günen 2022). RDA can also reveal the proportion of the difference in water quality attributed to different LULCs.

4.4 ENSEMBLE AND STACKED ENSEMBLE MODELLING (SEM)

Stacking is an ensemble machine learning algorithm that learns to combine prediction performance from multiple high-performing machine learning models effectively. To boost efficiency, this integrated algorithm employs a higher-level model to integrate lower-level models, ultimately increasing the predictive power of the classifier (Moradkhani and Fathi, 2022). Furthermore, by reducing bias and variance, this approach aims to minimize overfitting errors (Martín et al. 2021; Wu et al. 2021; Zounemat-Kermani et al. 2021). The most basic stacking model has two levels: level 0 (Base-Model) contains basic models, and level 1 (Meta-Model) contains the meta-learner (Kotekani and Ilango 2022). Base Model (Level 0): Models that fit the training data and predict data from outside the sample. Meta Model (Level 1): Model that fits on base-model predictions and learns how to best combine the predictions. Bagging and boosting are the most popular among the other methods.

Bagging: Also known as boot strap aggregation, combines the predictions from different decision trees (DT) via majority voting (democracy) (Figure 4.9). This approach, which employs DT, aids in variance reduction and model robustness enhancement; thus, it is well-suited for original weak models with high variance. RF, Extremely Randomized Tree (ERT) and rotation forest methods are some examples of bagging methods.

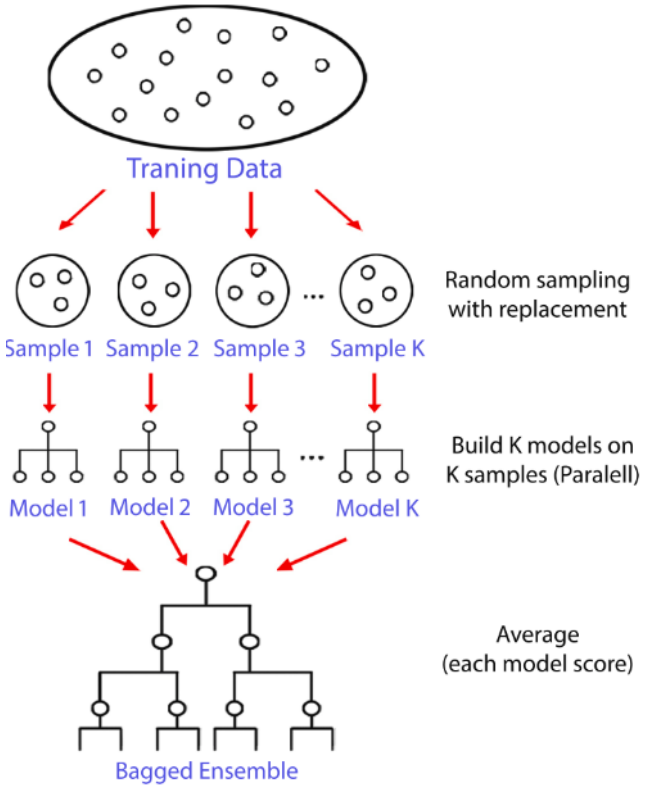


Figure 4.9 Schematic illustration of bagging ensemble modelling

Source: Zounemat-Kermani et al (2021)

Boosting: Builds model sequentially by minimizing the error from the previous models and boosting the influence of high-performance models (Figure 4.10). The boosting technique works by sequentially adding new models to the ensemble. Weak learners (base learners) are effectively boosted into strong learners in this ensemble. Boosting prevents DT from overfitting (Zounemat-Kermani et al. 2021). As a result, it aids in reducing variance and bias in ensemble machine learning and increasing prediction accuracy. Stochastic Gradient Boosting (SGB), AdaBoost, and eXtreme Gradient Boosting (XGBoost) are famous examples of this category

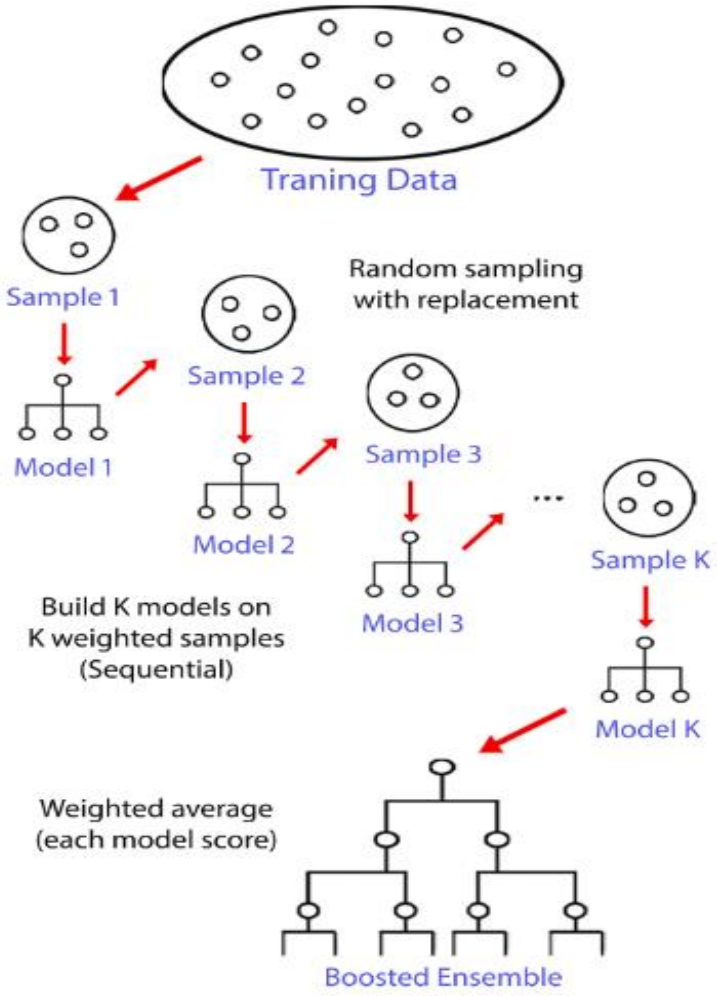


Figure 4.10 Schematic illustration of boosting ensemble modelling

Source: Zounemat-Kermani et al (2021)

The base model's training dataset can be subjected to k-fold cross-validation to avoid overfitting. At Level 0, four algorithms are considered Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM) and XGBoost. The below-explained steps and parameter range for the stacking algorithm (Table 4.1) are adapted from (Martín et al. 2021; Wu et al. 2021) is applied in this study.

The general steps involved in stacking model

1. Divide the original dataset into training and testing sets.
2. Implement a k-fold cross-validation to separate the dataset into k-folds.
3. Reserve one fold and train the other folds with multiple independent base models.
4. Predict the reserved fold using the base models

5. Repeat these steps until to get them out of sample predictions for all the defined k-folds.
6. Feed all out-of-sample predictions to the meta-model as features (training data).
7. Using the meta-model, predict the final result.

Table 4.1 Parameters selected for the stacking algorithm

Step 1	**Method**	**Parameter range**
Level-0 algorithms	SVM	tolerance in [0.0001, 0.01]; regularization in [2, 14];
	RF	trees in [350, 600]; min. Size of terminal nodes in [5, 15];
	GBM	trees in [550, 650];
	XGBoost	iterations in [40, 50]; instance weight in [5, 7]; trees in [50, 200]; depth in [3, 5]
Level-1 algorithms	XGBoost	iterations in [40, 50]; instance weight in [5, 7]; trees in [50, 200]; depth in [3, 5]

Source: Martín et al (2021)

4.5 RETRIVAL OF WQPS USING LANDSAT-8 AND MACHINE LEARNING ALGORITHMS

4.5.1 General

The goal is to propose a suitable learning-based algorithm for estimating optically active and non-active WQPs using an appropriate set of input variables (remote sensing band data). This could help to mitigate the drawbacks of limited *insitu* measurements for understanding spatiotemporal domains, as well as the cost of data collection and laboratory analysis. To accomplish this, we proposed ANN-based MLP and XGBoost regression with hyperparameter optimization, *insitu* water quality data and satellite-derived reflectance data (Landsat-8) for water quality prediction over inland waterbody. Finally, to delineate spatial maps for the predicted WQPs (methodology as discussed in Figure 4.11).

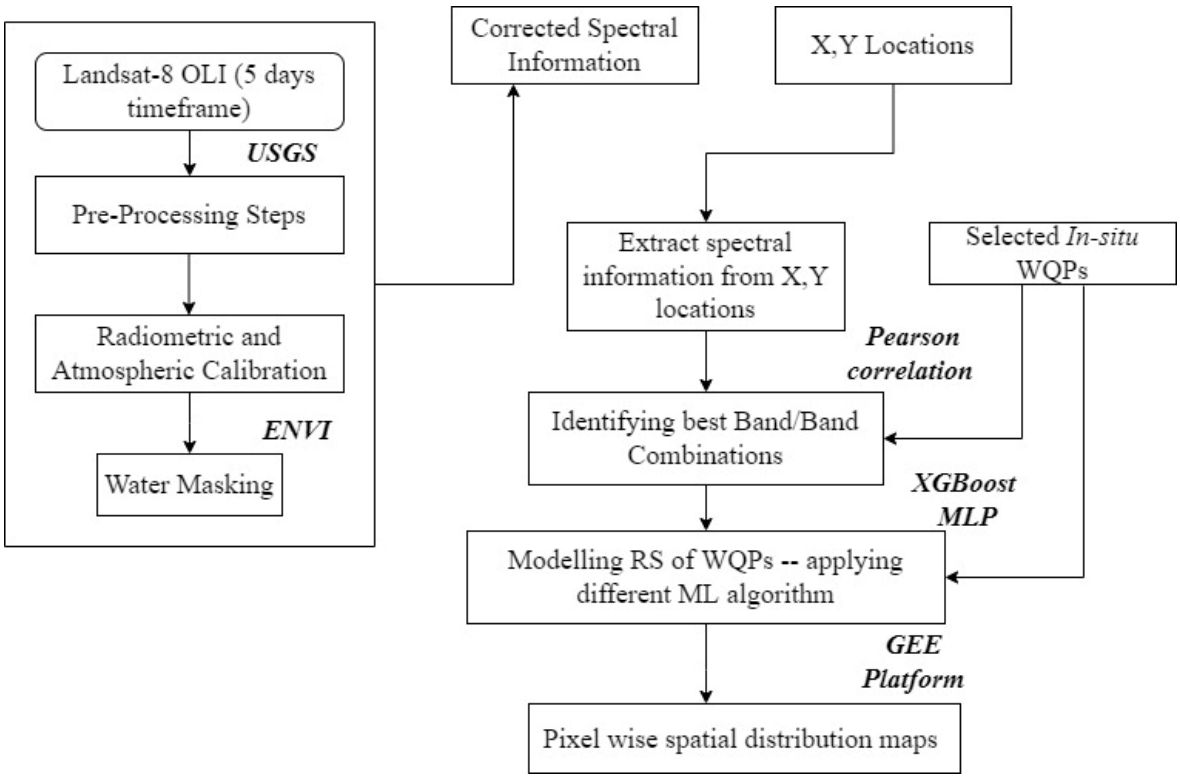


Figure 4.11 Concept map for remote sensing of WQPs using machine learning techniques

This section describes the methodologies in detail with four main Sections 1st Explains the study area selected and the list of clusters and WQPs chosen from the previous study. 2nd Satellite data, 3rd Modelling of WQPs using different machine learning algorithms: The feature selection method for the modelling, ANN-MLP and XGBoost regression algorithms, and hyperparameter optimization adopted for the study are discussed.

4.5.2 Satellite Data

Research suggests that imagery must be acquired within a day of an *insitu* collection event when selecting images for water quality correlation studies. This conservative window limits the availability of cloud-free Landsat images. Many researchers have even suggested a window of up to seven days as acceptable (Barrett & Frazier, 2016; Song et al., 2020) and 3-10 days as indicated by (Andrzej Urbanski et al., 2016). For our case study, we elected to include images collected within five days of (before or after) an *insitu* sampling event due to the highly variable weather pattern along the study area. Landsat-8 OLI data for cloud-free dates were acquired from 2014-2019 for the

study. Preprocessing the data, such as radiometric and atmospheric corrections, is essential for retrieving qualitative data using remote sensing imageries (Abdelmalik 2018). The images were precision corrected by radiometric and geometric means. Atmospheric correction was then conducted on the images acquired using the FLAASH module in ENVI (Olmanson et al. 2013; Abdelmalik 2018; Yepez et al. 2018) (Exelis garg Visual Information Solutions, Inc. Boulder, USA). FLAASH is a first-principle atmospheric correction tool that corrects wavelengths visible to near-infrared and shortwave infrared regions up to 3μm (Garg et al. 2017). It incorporates the MODTRAN 4 radiation transfer code. FLAASH is widely used to eliminate the effects of the atmosphere and convert spectral radiance to the surface reflectance of water. The geometric accuracy of the multispectral imagery was then checked by ground control points (Liu et al. 2015). The water area was then calculated using the Normalized Difference Water Index (NDWI) equation (4.9) in ENVI© 5.3.

$$NDWI = \frac{RG - R\ NIR}{RG + R\ NIR} \qquad (4.9)$$

Where RG and $R\ NIR$ are the reflectance in the green and NIR bands respectively

It is based on the difference in reflectance between NIR and SWIR bands, with higher NDWI values indicating higher probability of water. The NDWI values can range from -1 to 1, with values greater than 0 typically indicating the presence of water. However, the exact threshold value for water detection can vary depending on factors such as water body size, turbidity, and sensor characteristics, so a trial and error approach may be needed to find the optimal threshold value (Garg et al. 2020). After determining the water pixels, the visible VNIR bands were masked for water pixels for the respective dates.

4.5.3 Modelling of WQPs Using Different Machine Learning Algorithms

4.5.3.1 XGBoost modelling

Extreme Gradient Boosting (XGBoost) is a popular and powerful machine learning algorithm that is used for supervised learning tasks, such as classification and regression. It is an implementation of Gradient Boosting Machines (GBM) which is an ensemble method that combines multiple weak models to create a strong model

(Ibrahem et al. 2021). XGBoost uses the boosting ensemble learning algorithm principle better to predict performance (Kiangala and Wang 2021). A detailed explanation is carried out in this research paper. It is known for its high performance and ability to handle missing data and categorical features. Additionally, it is equipped with a number of built-in regularization options such as L1 and L2, which help to prevent overfitting. Ensembles are constructed from DT models, wherein trees are added one at a time to the ensemble. These will then correct the prediction errors made by prior models. XGBoost has the appealing properties of limited sample learning, fast model training, few parameters to adjust, mathematical solid explanation ability, tabular data processing, and data feature invariance when compared to popular neural network-based deep learning models. The procedure of constructing XGBoost consist of assembling a base model for an existing model, i.e., training an initial tree then it will construct a second tree combined with the initial tree, and repeat the second step until the expected number of trees is reached (Zhang et al. 2020).

The principle behind XGBoost is gradient boosting, which is a method of ensemble learning that combines the predictions of multiple weak models to create a more accurate and robust final prediction (Natekin and Knoll 2013). The basic idea is to train a sequence of simple models, such as decision trees, on the data and iteratively add new models that correct the errors made by the previous models. XGBoost is an implementation of gradient boosting that uses decision trees as the base model. It is designed to handle large datasets and high-dimensional features by using a technique called "gradient tree boosting" which is efficient and scalable. XGBoost uses a technique called SGB which is a variation of gradient boosting that randomly subsamples the training data before fitting the model, which makes the algorithm more robust to overfitting and improves the speed of the training process. Additionally, XGBoost also has a built-in regularization term called "Lambda" to prevent overfitting. XGBoost also has a number of advanced features such as handling missing data, parallel and distributed computing, and built-in evaluation metrics. To summarize, XGBoost is an efficient and powerful implementation of gradient boosting that uses decision trees as the base model, and it handles large datasets and high-dimensional features by using techniques such as SGB, regularization, and advanced features such

as handling missing data and parallel computing. A detailed workflow is explained in Figure 4.12.

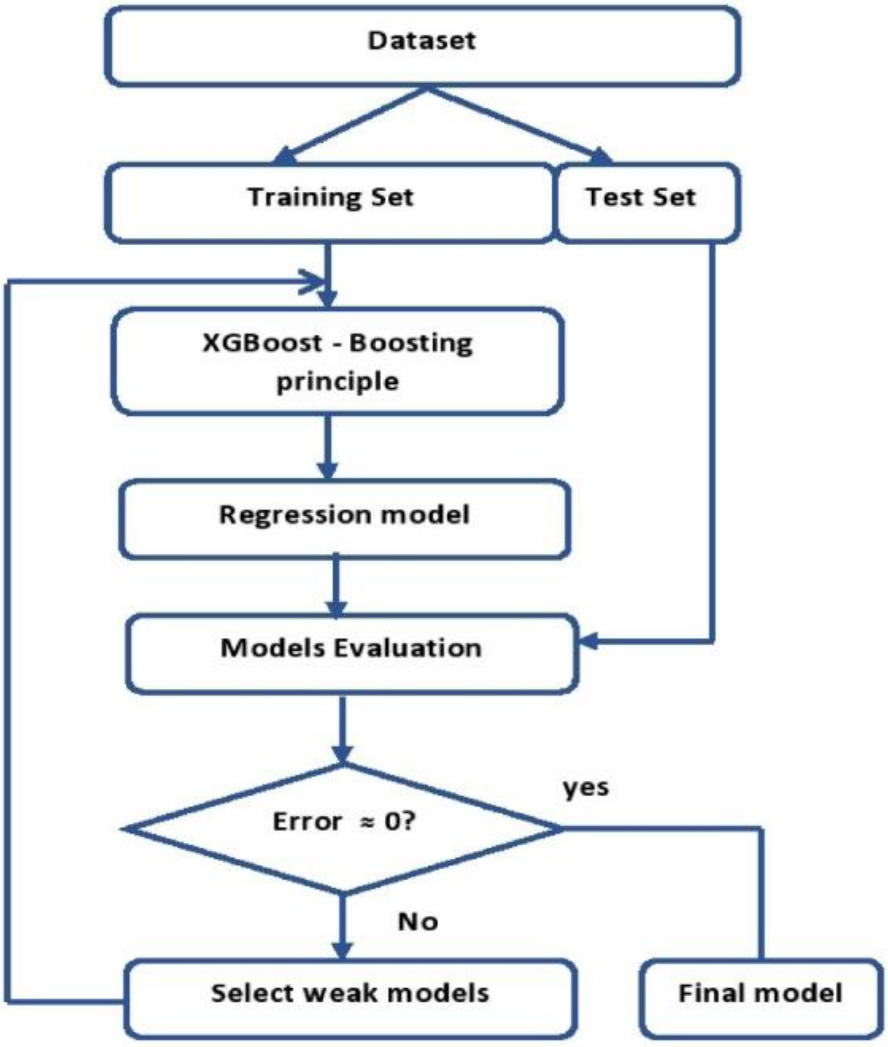


Figure 4.12 XGBoost Algorithm workflow

Source: Kiangala and Wang(2021)

It is designed to be highly efficient and scalable, and can handle large datasets with ease. Most XGBoost enhancements focus on fully maximizing computing and memory capacities to speed up the learning process to the maximum (González et al., 2020). Besides, XGBoost also includes necessary adaptations to reduce over-fitting and extend its use to all problems. The main feature against over-fitting is its regularized model formalization. XGBoost is a powerful gradient boosting algorithm that can perform three different boosting techniques: Regularized Boosting, Stochastic Boosting, and

Gradient Boosting. Regularized Boosting is a technique that adds a regularization term to the loss function, which helps to prevent overfitting. Stochastic Boosting is a technique that randomly subsamples the data before each iteration, which can improve the model's generalization performance. To approximate the goals, XGBoost adopts the Taylor expansion (González et al. 2020). XGBoost objective function includes a regularization term that controls the complexity of the model. This addition allows to learn simple and predictive models and find a good bias-variance trade-off.

The theoretical background of the same is explained in the below section (Kiangala and Wang 2021). Dataset *B* contains a collection of independent attributes or observations (g_i) and dependent variable or labels (h_i) can be represented as equation (4.10)

$$B = \{g_i, h_i\} \qquad (4.10)$$

$i \in \{0 \dots k\}$, k denotes the total number of samples in the dataset

B and *i* is the i^{th} sample of the dataset. If the total number of DT in the model is *G* and the expected values are $\tilde{h}_i$, which can be computed by following equation (4.11)

$$\tilde{h}_i = \sum_p^G f_p(g_i) \qquad (4.11)$$

Where, $f_p(g_i)$ is the expected count of sample *i* for the p^{th} tree

To obtain excellent prediction outcomes, the XGBoost algorithm develops a "objective function" to optimize the loss function using the previous level's results and it incorporates a "regularization" mechanism to improve its output.

The initial objective function can be represented mathematically as in equation (4.12)

$$K = \sum_{i=1}^n l\left(h_i, \tilde{h}_i\right) + \sum_{p=1}^G \Omega\left(f_p\right) \dots\dots (4.12)$$

Where *n* - is the total amount of data processed at the p^{th} tree, *G* - the number of trees in total, *i* - i^{th} sample of the dataset, $l\left(h_i, \tilde{h}_i\right)$ - loss function computes the difference between the dependent variable and the independent variable (*hi*) and its predicted value $\tilde{h}_i$.

Equation (4.13) and (4.14) represents loss functions and regularization terms for overfitting issues, respectively.

$$l\left(h_i, \tilde{h}_i\right) = \sum_{i=1}^n \left(h_i - \tilde{h}_i\right)^2 \qquad (4.13)$$

$$\Omega(f) = \theta T + \frac{1}{2}\eta \sum_{b=1}^{T} (X_b^2) \qquad (4.14)$$

where *T*- sum of regression lead nodes in a DT, b - each leaf in a node has its own identification index, *X*- a single leaf node's weight or score, θ and η weight parameters higher the values simple DT structure and less overfitting risks. The objective function at a particular iteration *s* can be rewritten as given in (4.15). The second-order Taylor expansion, i.e., the optimized version of (4.15) is given in (4.16).

$$K^{(s)} = \sum_{i=1}^{n} l\left(h_i, \tilde{h}_i^{s-1} + f_s(g_i)\right) + \Omega(f_s) \qquad (4.15)$$

$$K^{(s)} = \sum_{i=1}^{n} \left[l\left(h_i, \tilde{h}_i^{s-1}\right) + m_i f_s(g_i) + \frac{1}{2} n_i f^2(g_i)\right] + \Omega(f_s) \qquad (4.16)$$

Where $m_i \; and \; n_i$ The gradient statistics of the loss function can be defined in equation (4.17) and (4.18).

$$m_i = \partial_{\tilde{h}_i^{(s-1)}} l\left(h_i, \tilde{h}_i^{(s-1)}\right) \qquad (4.17)$$

$$n_i = \partial^2_{\tilde{h}_i^{(s-1)}} l\left(h_i, \tilde{h}_i^{(s-1)}\right) \qquad (4.18)$$

Substituting (4.14), (4.16), (4.17) and (4.18) extracting the derivative and computing the optimal value of loss function score as in equation 4.19 at a specific leaf *t.*

$$K^{(s)} = \sum_{i=1}^{n} \left[l\left(h_i, \tilde{h}_i^{s-1}\right) + m_i f_s(g_i) + \frac{1}{2} n_i f_s^2(g_i)\right] + \dot{\Omega}(f_s)(7) K^{(s)} =$$

$$\sum_{i=1}^{n} \left[l\left(h_i, \tilde{h}_i^{s-1}\right) + m_i f_s(g_i) + \frac{1}{2} n_i f_s^2(g_i)\right] + \dot{\Omega}(f_s)$$

$$K^* = -\frac{1}{2} \sum_{t=0}^{T} \frac{(\sum m_i)^2}{\sum n_i + \eta} + \theta T \qquad (4.19)$$

$$X_t^* = -\frac{\sum m_i}{\sum n_i + \eta} \qquad (4.20)$$

A low value of *K** indicates a superior DT composition. The ideal weight of a leaf *t* is represented by equation (4.20).

4.5.3.2 Multi-Layer Perceptron (MLP)

The multi-layer perceptron (MLP) is a type of feedforward artificial neural network that is commonly used for supervised learning tasks such as regression and classification. MLPs can be used to model complex non-linear relationships between input and output variables, making them well-suited for applications in water resource management,

such as predicting water levels, flows, and other hydrological variables. The ability of MLPs to capture non-linear relationships is due to the presence of multiple layers of interconnected nodes, or neurons, that are trained to learn the underlying patterns in the data. MLP network is an extension of perception consisting of three layers: input, hidden, and output (Figure 4.13). It includes a set of artificial neurons that are information processing units. MLP uses the biological nervous system principle, which comprises a massive parallel system composed of many processing elements connected by links of variable weights. An elementary neuron consists of *R* inputs (different Band/Band combinations), each input is weighted with an appropriate weight of *W*. The sum of these weighted inputs and the bias forms the input to the transfer function *f*. Neurons can use any differentiable transfer function *f* to generate their outputs. The initial assigned weights are progressively corrected during the training process. Here, the outputs predicted by MLP are compared with known outputs, and errors are back propagated (from right to left) to determine the appropriate weight adjustments necessary to minimize errors. It is difficult to choose the best algorithm that can accurately predict the target while optimizing many factors such as processing speed, numerical precision, and memory requirements. As a result, choosing a training algorithm is the most important step in ANN. A training algorithm that works well for one problem but fails in another. In our study, we used the Levenberg–Marquardt backpropagation learning rule, which is a variant of Newton's method that incrementally adjusts the weight and bias terms to minimize the network's mean square error (MSE) (Naganna and Deka 2019) equation (4.21).

The output of a neuron can be expressed as *f (e*).

Where,

$$e = \sum_{j=1}^{R} W_j X_j + B; X_1, X_2, X_3 \ldots X_R \qquad (4.21)$$

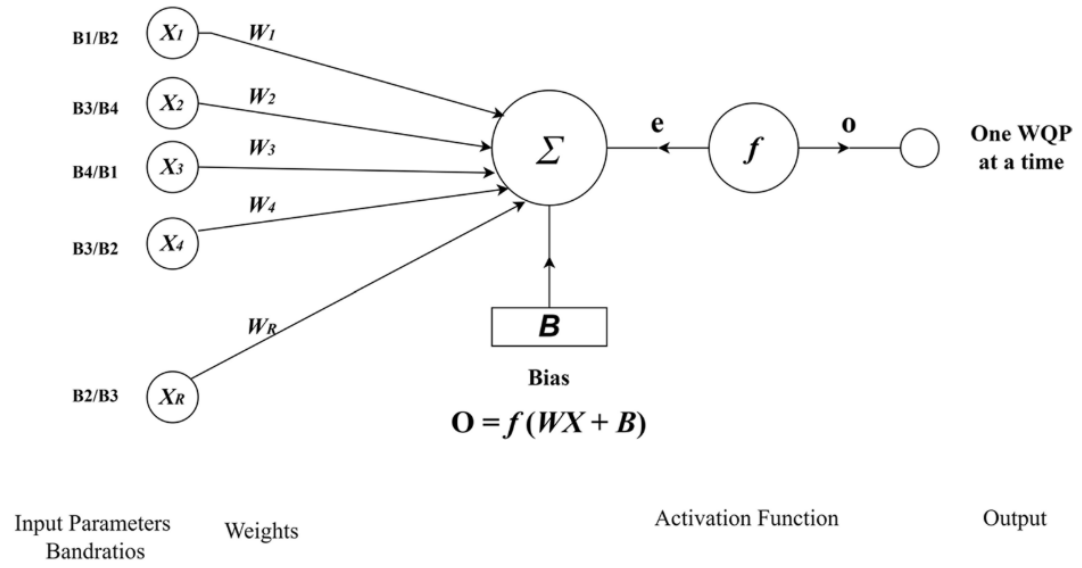


Figure 4.13 The basic structure of MLP

X_1, X_2, X_3…. X_R are the input signals, and W_1, W_2, …… W_R are the weights of neurons. B is the bias value, f is the activation function, and R is the number of elements in the input vector (Ay and Kisi, 2014; Naganna and Deka, 2019; Günen et al. 2020). At first, allocated weights are gradually adjusted over the training procedure. In addition to the input nodes, each node is a neuron with a non-linear activation function and stringent feature selection and data normalisation constraints (Tian et al. 2020). This methodology works by learning the problem-solving process and determining the implicit link. However, MLPs still suffer from problems (Zhan et al. 2003). First, the training algorithm may be trapped in a local minimum, and objective functions are frequently extremely complex. Traditional training algorithms are easily trapped in a local minimum and will never converge to an acceptable error. Even the training dataset cannot be adequately fit in that case. Second, it is difficult to determine the best MLP architecture, such as the number of hidden layers and nodes within them.

4.5.3.3 Feature Selection for XGBoost and MLP

Feature selection is the primary focus of any machine learning model to remove the non-informative or redundant predictors from the model. Because adding unwanted variables to the model will reduce the generalization and overall accuracy of any model, moreover, increase the complexity. Pearson's correlation coefficient method was used to analyse the correlation between WQPs, remote sensing bands, and band

combinations. Pearson's correlation coefficient is a method for analysing whether there is a close correlation between two variables, defined as the covariance quotient and the standard deviation between two variables. The correlation of two random variables can be well measured based on covariance.

4.5.3.4 Hyperparameter Optimization (HPO)

Model hyperparameters are the parameters that cannot be estimated by the model using the given data. Although, the model from the data can estimate a model parameter. HPO is the process of identifying the right combination of hyperparameters that makes the model maximize its performance. Conceptually, hyperparameters tuning is just an optimization loop on top of machine learning model to find the set of hyperparameters leading to the lowest error on the validation set. An hyperparameters is a parameter whose value controls the learning process. Model tuning is carried out for hyperparameters in order to determine the parameters that result in the most accurate predictions. These parameters directly influence the behavior of the training algorithm. These have a significant impact on model performance. Nevertheless, choosing the right combination of hyperparameters is not an easy task. Hyperparameters can be adjusted by manual tuning or by automated tuning, and the former is time-consuming. Automated tuning approaches such as Optuna and GridsearchCV were used in this work for the XGBoost and MLP regressors, respectively.

4.5.3.5 Hyperparameter optimization for XGBoost using Optuna

Optuna is a software framework for automating the optimization process of these hyperparameters. It automatically finds optimal hyperparameters values using different samplers such as Grid search, Random, Bayesian, and Evolutionary algorithms. Optuna uses a historical record of trails details to determine the promising area to search for optimizing the hyperparameters and hence finds the optimal hyperparameters in a minimum amount of time. It has a pruning feature that automatically stops the unpromising trails in the early stages of training to save computing time.

4.5.3.6 Hyperparameter optimization for MLP regressor using GridsearchCV

The traditional way of performing HPO is by exhaustive searching within a specified subset of the hyperparameters space of a learning algorithm. A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set. Since the parameter space of a machine learning may include real-valued or unbounded value spaces for specific parameters, manually set bounds and discretization may be necessary before applying grid search. Grid search is arguably the most basic hyperparameters tuning method. With this technique, we build a model for each possible combination of the hyperparameters values provided, evaluate each model, and select the architecture that produces the best results.

Below is a list of hyperparameters applied for tuning the model:

Hidden layer sizes - A tuple of numbers defining the sizes of hidden layers in multi-layer perceptions is accepted. Many perceptron will be generated per hidden layer based on the size of the tuple, default = (100,).

- activation - It defines the function for activating hidden layers, default=relu
    - 'identity' – Number of Activation. $f(x) = x$
    - 'logistic' - Logistic Sigmoid Function. $f(x) = 1 / (1 + exp(-x))$
    - 'tanh' - Hyperbolic tangent function. $f(x) = tanh(x)$
    - 'relu' - Rectified Linear Unit function. $f(x) = max(0, x)$
- solver - It accepts one of the following strings to select the optimization solver to use for updating neural network hidden layer perceptron weights, default='adam'
    - 'lbfgs'
    - 'sgd'
    - 'adam'

- learning_rate - The learning rate controls how much to update the weight at the end of each batch, and the momentum controls how much to let the previous update influence the current weight update. The learning rate controls how quickly the model is adapted to the problem.
- 'constant' - Maintains a consistent learning rate using a learning method specified in learning_rate_init.
- invscaling' - It steadily reduces the learning rate.
- 'adaptive' - It keeps the learning rate constant as long as the loss is reducing, or the score improves
- early_stopping - It allows a boolean indicating if training should be stopped if the training score/loss is not improving.

Grid Search tries the list of all combinations of values given for a list of hyperparameters with the model, records the model's performance based on evaluation metrics, and keeps track of the best model and hyperparameters.

4.5.3.7 Performance Evaluation Matrices

The Table 4.2 below explains the different performance matrices applied in remote sensing of water quality.

Table 4.2 Performance matrices

Criteria for Statistics	**Range**	**Inference**
Root Mean Square Error $\boldsymbol{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$	Value below half of standard deviation	Satisfactory
Relative RMSE $\boldsymbol{RRMSE} = \frac{RMSE}{\sigma_{obs}}$	0.00 <= RRMSE <= 0 .10	Very Good
	0.10 <= RRMSE <= 0.30	Good
	0.30 <= RRMSE <= 0.50	Satisfactory
	RRMSE > 0.70	Poor

Coefficient of determination $\boldsymbol{R^2} = \frac{SSR}{SST}$ $\boldsymbol{SSR} = \sum_i (\hat{y}_i - \bar{y})^2$ $\boldsymbol{SST} = \sum (y_i - \bar{y})^2$	$R^2 > 0.85$	Very Good
	$0.75 < R^2 <= 0.85$	Good
	$0.60 < R^2 <= 0.75$	Satisfactory
	$R^2 <= 0.60$	Poor
Adjusted R^2 R^2_{adj} $= 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$	$R^2 > 0.85$	Very Good
	$0.75 < R^2 <= 0.85$	Good
	$0.60 < R^2 <= 0.75$	Satisfactory
	$R^2 <= 0.60$	Poor

SSR – Sum of Squared Regression or the variation explained by the model

SST – Sum of Squared Total or Total variation in the data

y_i – *y* value for observation *i*

$\bar{y}$ – Mean of *y* value

$\hat{y}$ – Predicted value of *y* for observation *i*

n- Number of data points

k - represents the number of independent variables

R^2 - represents the R-squared values determined by the model

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 GENERAL

The descriptive and inferential statistical findings of various machine learning algorithms and remote sensing applications evaluated in the research region are presented in this chapter. The resulting statistical metrics are thoroughly examined and assessed to give a significant theoretical breakthrough in our knowledge of the spatiotemporal condition of river water quality. The graphs and maps created using the methodologies covered here are examined briefly. This chapter also discusses the outcomes of spatial maps developed with remote sensing and machine learning methods

5.2 SPATIOTEMPORAL WATER QUALITY ASSESSMENT

5.2.1 Feature selection and Dimensionality reduction

Descriptive statistics showing the spatial and seasonal variations of WQPs for all 20 monitoring stations are analysed separately for non-monsoon and monsoon seasons. The correlation heatmap is plotted using the Seaborn visualization library in Python, which is based on the Matplotlib library to visualize and interpret the data based on the colour. Table 5.1 & Figure 5.1 for the non-monsoon season.

Table 5.1 Descriptive Statistics for non-monsoon season (2005-2018)

WQPs	count	mean	std	min	25%	50%	75%	max
EC(µmho/cm)	16	359.43	83.71	189.48	297.14	386.20	425.39	488.85
pH	16	8.14	0.13	7.87	8.06	8.15	8.21	8.50
TDS (mg/L)	16	235.36	41.88	172.79	200.35	225.34	276.41	306.07
Temp (deg C)	16	21.50	1.28	19.44	20.83	21.34	22.80	23.72
Ca (mg/L)	16	36.48	4.19	27.70	35.33	37.32	39.11	42.41
Cl (mg/L)	16	27.75	12.37	13.22	20.44	24.66	30.36	55.97
CO_3 (mg/L)	16	7.50	3.28	4.07	5.79	6.48	7.63	15.93
F (mg/L)	16	1.21	3.63	0.19	0.22	0.30	0.39	14.82

HCO_3 (mg/L)	16	167.60	26.59	113.60	160.39	170.54	177.92	212.47
K (mg/L)	16	7.16	1.84	4.51	5.55	7.69	8.40	10.32
Mg (mg/L)	16	19.06	2.81	14.68	17.78	18.70	20.59	25.24
Na (mg/L)	16	21.08	9.58	8.62	15.36	18.28	30.47	39.63
NH_3-N (mg/L)	16	0.70	0.98	0.04	0.05	0.29	1.22	3.60
NO_2+NO_3 (mg /L)	16	1.89	2.53	0.35	0.38	0.53	1.97	6.91
P-Tot (mg/L)	16	0.13	0.10	0.01	0.01	0.19	0.21	0.24
SiO_2 (mg/L)	16	8.27	0.80	6.33	7.82	8.46	8.82	9.28
SO_4 (mg/L)	16	24.35	6.28	14.74	19.69	23.71	27.64	35.41
BOD (mg/L)	16	3.48	2.48	0.99	2.19	2.94	4.10	10.83
COD (mg/L)	16	12.23	4.46	5.85	9.13	12.68	14.51	23.43
DO (mg/L)	16	6.90	1.18	2.82	6.87	7.27	7.49	7.88

This approach was applied to synthesize the complex water quality data matrix and understand the correlation pattern between different WQPs.

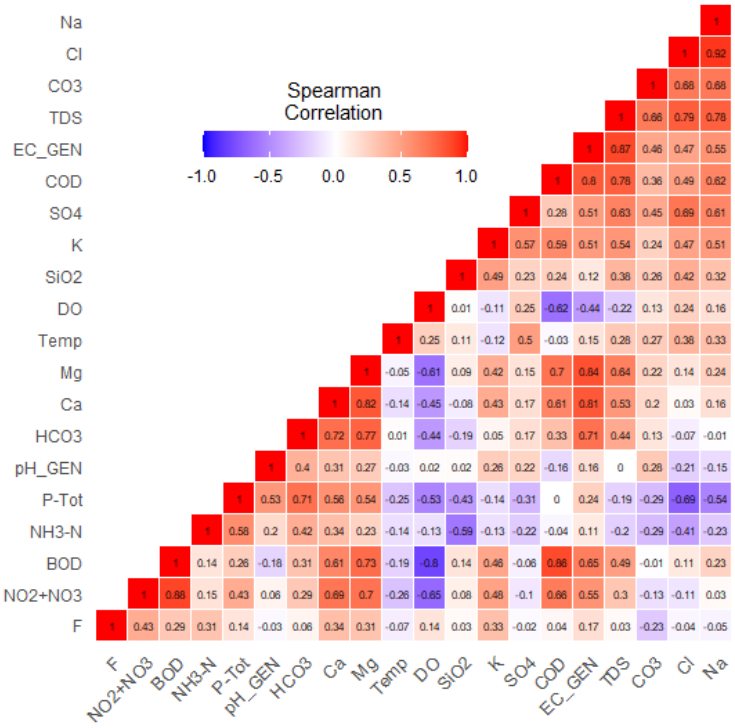


Figure 5.1 Correlation among WQPs non-monsoon season (2005-2018)

The Spearman non-parametric correlation coefficient test determines the temporal fluctuations in river water quality (Spearman *r*). Scientific library for Python (scipy), from scipy.stats package was utilized to perform Spearman *r* and significance of correlation (*p-value*). Before starting the test, two-season was assigned a numerical value in the data file (Monsoon=1 and Non-monsoon=2) as mentioned by (Wunderlin et al., 2001; Singh et al., 2004). The results of correlation and statistical significance against the season are discussed in Table 5.2.

Table 5.2 Correlation and statistical significance of WQPs against the season

WQPs	**Correlation with season *r***	**Statistical significance *p-value***
EC	0.3812	**0.0152****
pH	0.3942	**0.0118****
TDS	0.3725	**0.0179****
Temp	**-0.8663***	**0.0000****
Ca	0.5328	**0.0040****
Cl	0.1689	0.2974
CO_3	0.3075	0.0536
F	0.0260	0.8735
HCO_3	0.4765	**0.0019****
K	0.2469	0.1246
Mg	0.4895	**0.0013****
Na	0.2296	0.1542
NH_3-N	0.0780	0.6325
NO_2+NO_3	0.4158	**0.0076****
P-Tot	0.1083	0.5060
SiO_2	0.3119	**0.0501****
SO_4	0.2079	0.1980
BOD	0.0606	0.7101
COD	0.0823	0.6137
DO	0.3595	**0.0227****

**r*>0.8 and ** $p < 0.05$ are marked as bold.

Spearman *r* and *p-value* are calculated using this numerical value for 20 nos of WQPs. Temperature with the season has the greatest Spearman *r* (-0.866) with a *p-level* of 0.0000. Generally, many factors affect the increase or decrease in temperature, percentage of DO and other biological activities. Some parameters had a moderate temporal correlation (Ca, Mg, HCO_3 and Mg). The season has a strong correlation with the parameters EC, pH, TDS, T, Ca, HCO_3, Mg, NO_2+NO_3, SiO_2, and DO ($p < 0.05$). Cl, CO_3, F, K, Na, NH_3-N, P-Tot, SO_4, BOD and COD had a non-significant correlation with *r-value* between moderate to low range. The aforementioned suggests changing natural and anthropogenic sources in the catchment. However, including other prediction variables like LULC, flow and rainfall could be more beneficial to conclude this precisely.

5.2.2 Spatiotemporal Clustering

The CA classified 20 monitoring stations in this study into four distinct clusters based on the measured variables: Cluster 1, Cluster 2, Cluster 3, and Cluster 4 (called C1, C2, C3, and C4) and was confirmed by investigating the cluster quality by the silhouette of cohesion (Chang et al., 2012; Ay & Kisi, 2014; Shamitha & Ilango, 2019). A silhouette score of < 0.6 is observed when k=4 for the non-monsoon and monsoon seasons (Figure 5.2 & Figure 5.3).

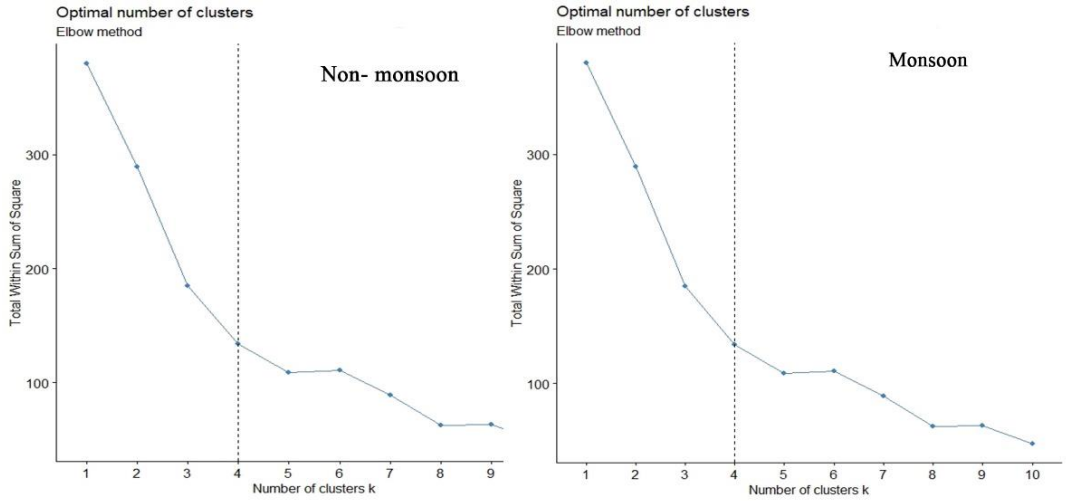


Figure 5.2 Seasonal identification of optimum number of clusters Elbow method

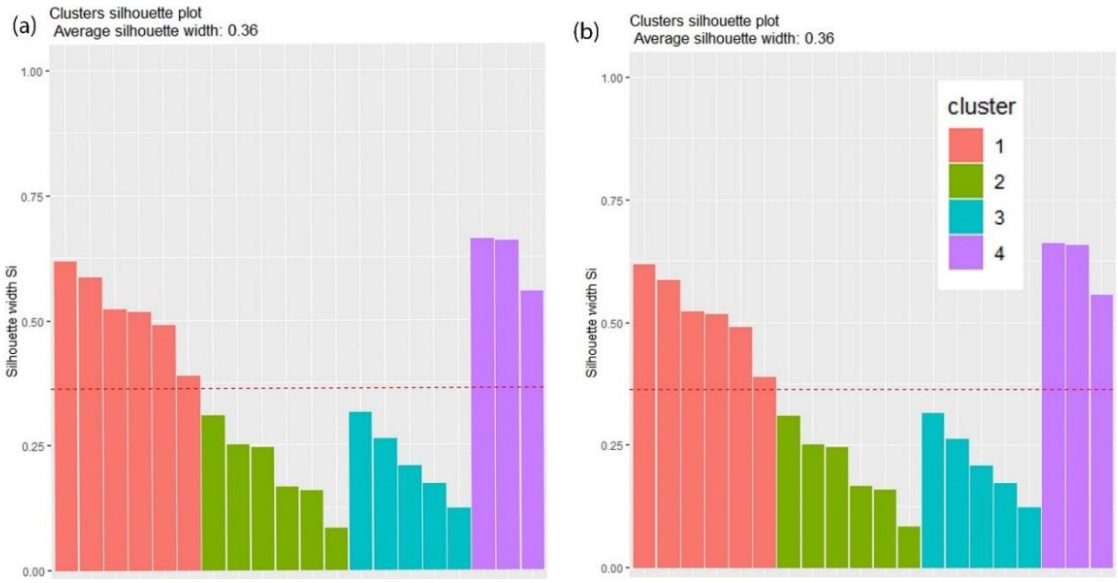


Figure 5.3 Cluster silhouette plot for non-monsoon and monsoon season

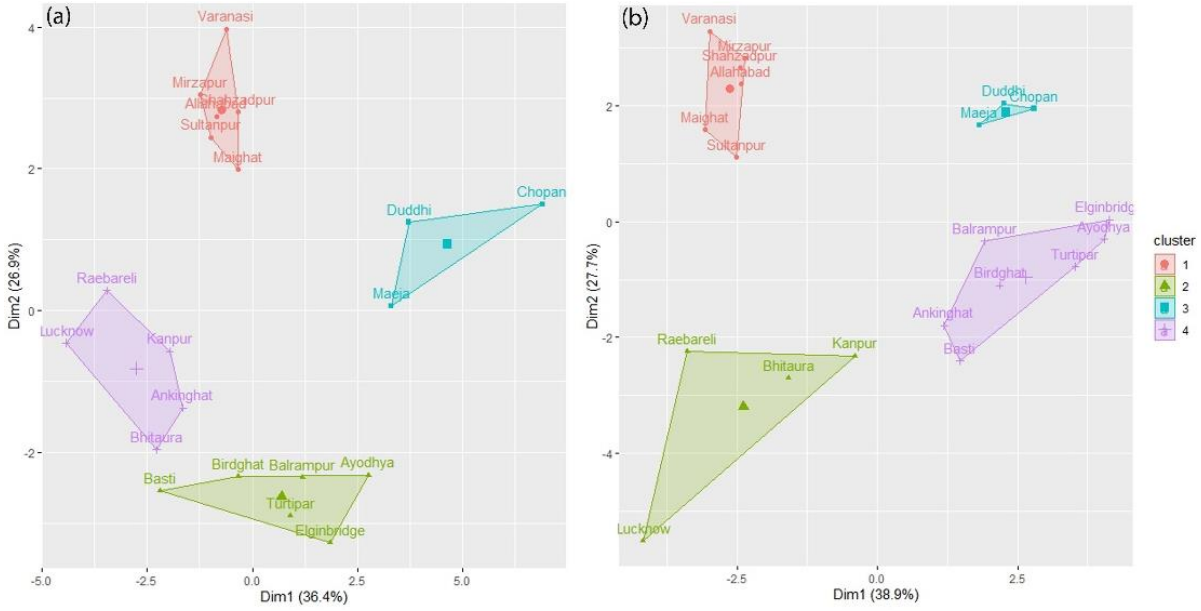


Figure 5.4 Clusters identified along the study area for non-monsoon (a) and monsoon (b) seasons

As presented in Figure 5.4, Allahabad, Maighat, Mirzapur, Shahzadpur, Sultanpur, and Varanasi are the stations located along the downstream side of the study area that falls in C1. The agricultural, barren land and built-up are identified as primary land use along these clusters. Built-up areas were discovered in a few pockets away from the monitoring stations along this cluster. Allahabad is one of the most populous districts along C1 and is located at the confluence of three rivers: The Ganga, the Yamuna, and the Saraswati. Heavy metal pollution and sediments above acceptable concentrations

have been identified along these rivers. Sultanpur is another urban settlement along the river bank which receives a considerable quantity of untreated domestic, urban sewage and agricultural runoff from the industries like sugar mills, distilleries, paper/pulp and agrochemicals etc (Singh et al. 2004; Iqbal et al. 2019). As a result, a thorough run-off analysis would be required to determine the precise pollution contributions from these land use.

Stations along the C2 and C4 clusters exhibit spatial variations during the non-monsoon and monsoon seasons. During the non-monsoon season, Basti, Birdghat, Balrampur, Ayodhya, Turtipar, and Eliginbridge, located on the study area's extreme eastern side, are falls in C2 and Lucknow, Raebareli, Bhitaura and Kanpur for the monsoon season. This spatial shift during the monsoon season along the clusters indicates the higher contribution of NPS pollutants along these stretches. C1 and C3 are the clusters that displayed no spatial variations along the study area. Located on the extreme eastern side of the study area, Maejja, Dudhi, and Chopan fall in C3 during the non-monsoon and monsoon seasons. Besides agricultural land, thick vegetation is another important land use identified along these clusters. C4 is situated along the upstream side and spreads from the west to the eastern side of the study area. Lucknow, Raebareli, Kanpur, Ankinghat and Bhitaura in non-monsoon and Ankinghat, Basti, Birdghat, Balrampur, Turtipar, Ayodhya and Elginbridge in the monsoon season are the stations comprised along this cluster. A robust spatial distribution of agricultural land use and the barren area is identified as a primary land use along all these clusters. The list of different clusters and spatial distribution during the non-monsoon and monsoon seasons are discussed in (Table 5.3 & Figure 5.5). For better understanding of WQPs, the spatiotemporal patterns that exist in the individual clusters are analysed using Box and Whisker plots for different seasons.

Table 5.3 Spatiotemporal cluster identified for non-monsoon and monsoon season

Station name	Lat	Long	Cluster non-monsoon	Cluster monsoon
Ankinghat	26.9066	80.0727	C4	C4
Ayodhya	26.8133	82.2069	C2	C4
Balampur	27.4369	82.2286	C2	C4
Basti	26.7827	82.7147	C2	C4

Bhitaura	26.028	80.8477	C4	C2
Birdghat	26.7213	83.3502	C2	C4
Allahabad	25.3983	81.9122	C1	C1
Chopan	24.5258	83.0461	C3	C3
Duddhi	24.7263	83.2719	C3	C3
Elginbridge	27.0933	81.4844	C2	C4
Kanpur	26.4647	80.3794	C4	C2
Lucknow	26.8602	80.9425	C4	C2
Maighat	25.6383	82.8625	C1	C1
Maeja Road	25.233	82.038	C3	C3
Mirzapur	25.158	82.5461	C1	C1
Raubareli	26.2411	81.2055	C4	C2
Shahzadpur	25.6613	81.435	C1	C1
Sultanpur	26.2872	82.1255	C1	C1
Turtipar	26.1419	83.8376	C2	C4
Varanasi	25.3247	83.0363	C1	C1

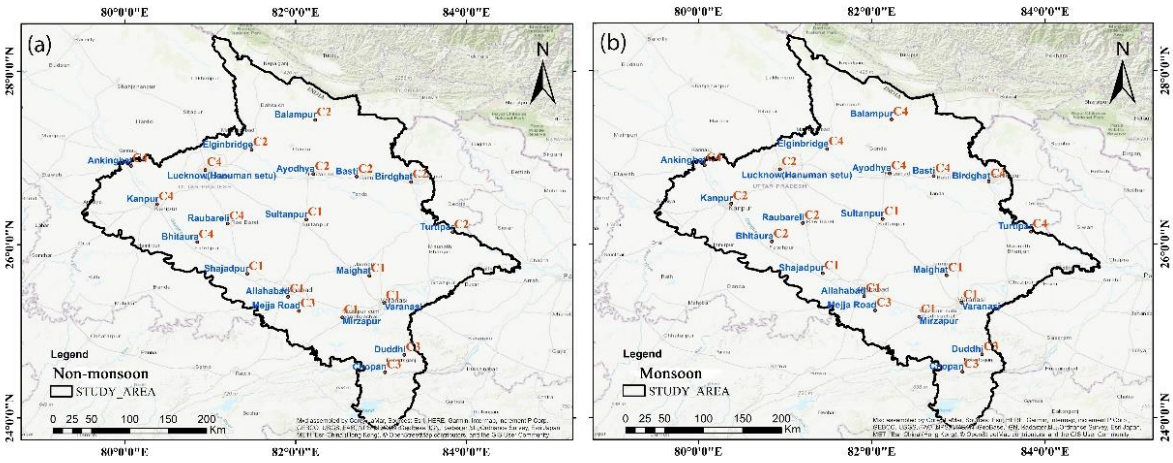


Figure 5.5 Geographical location of clusters for non-monsoon and monsoon seasons

Many sections of the Ganga basin have high EC values surpassing the allowable limit of 3000 µS/cm (CWC and NRSC 2014). The reported value for this study ranges between 68-4460 µmhos/cm, thus, 75% of observations are not achieving the desired limit of EC in both seasons. Significantly lower values are observed in the monsoon

season for C1 and C3, which could be due to dilution (Tibebe et al. 2019). Whereas the observed desired limit of TDS is satisfied for Class A, C and E along the study area. From the graph (Figure 5.6), we can also see a similar trend during the monsoon season.

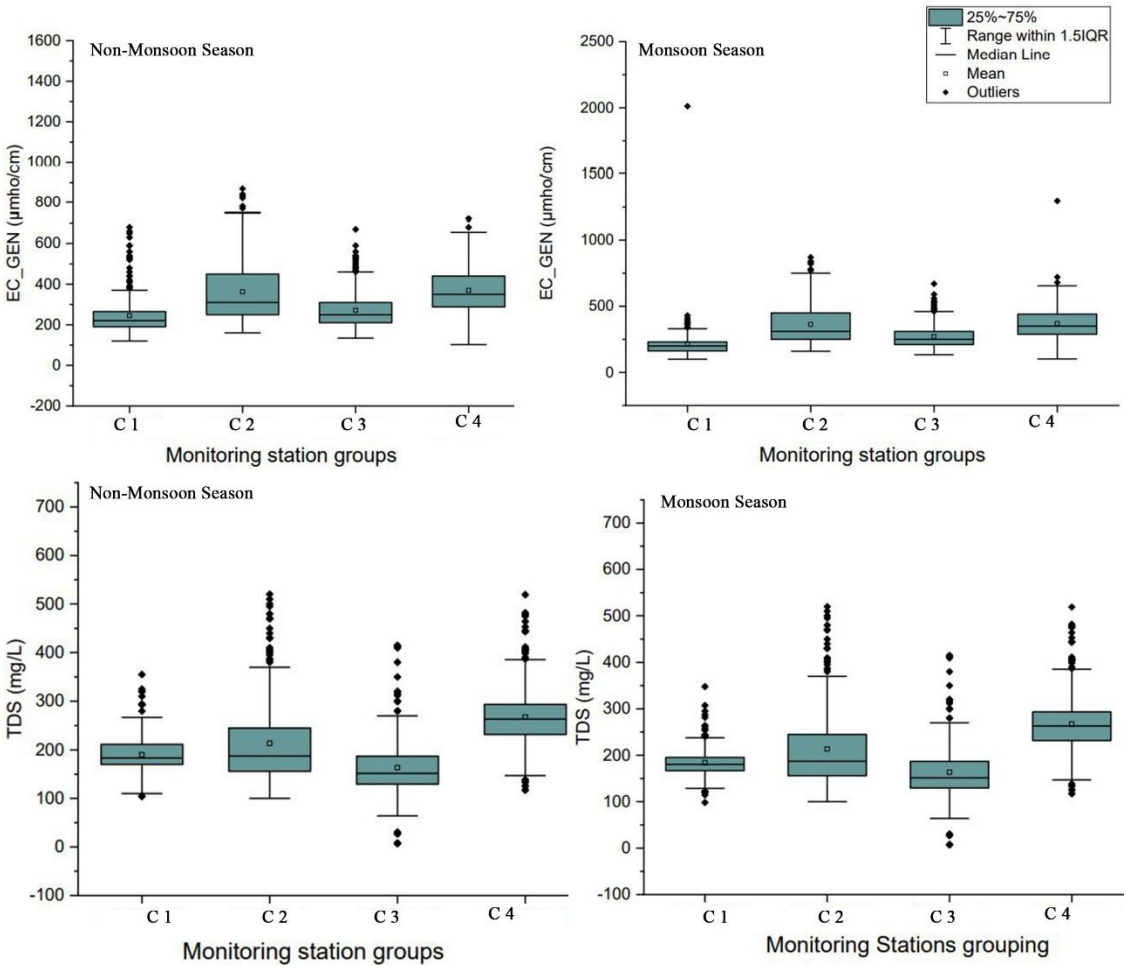


Figure 5.6 Spatiotemporal pattern for EC and TDS along different clusters

The average minimum and maximum temperature of the whole basin are published as 18.44°C and 32.05°C, respectively (CWC and NRSC 2014). Tropical and subtropical temperature zones predominate across the Ganga basin. The tropical temperature zone in the basin has a mean annual temperature of more than 24°C and a mean January temperature of more than 18°C. In contrast, the subtropical temperature zone has a mean annual temperature of 17°C - 24°C and a mean January temperature of more than 10°C - 18°C. As per observation, about 25-75% of data are falling in the range of 18-32°C for the non-monsoon season (Figure 5.7). Of the four clusters, C1 displays a different trend than other clusters. C1 has a minimum and mean value of 18 and 23°C,

respectively, with no outliers. Clusters 2, 3 and 4 have a similar trend with minimum and maximum values of 26-32°C, respectively. During the monsoon season, all the clusters show a similar trend with a mean value of 25-27°C. The non-monsoon season has a higher mean value than the monsoon season. Temp fluctuations are usually attributed to waste discharge from thermal industries and organic waste discharges. Many researchers have identified the correlation between the quantity of water and rainfall with temperature (Ay & Kisi, 2014; Sandoval et al., 2014; Álvarez-Cabria et al., 2016). The studies conducted by the CWC using IMD gridded data have identified, for the past the 35 years (1969-2004), the average annual mean temperature of the Ganga basin was 24.82°C, with a high of 32.05°C in 1987 and an average yearly lowest temperature of 18.44°C, but the annual minimum temperature in 1971 was 17.68°C (CWC and NRSC 2014). Temperature is an essential factor that affects physical, chemical and biological parameters (Chang et al., 2015). We have identified a strong correlation of temperature with EC, pH, DO, SiO_2 and Ca along the study area.

As per IS specification, the desirable limit for pH is 6.5-8.5 for non-monsoon and monsoon seasons. In our study, a similar trend is experienced in both seasons, as shown in Figure 5.7 for non-monsoon and monsoon seasons ranged from 7.0-8.7, and around 25-75% of data falls within the desirable range with a mean value of 8-8.1, maximum values in all clusters exceeding 8.5, which is undesirable. pH plays a critical role in water chemistry and is an essential measurement concerning water quality. pH more than 7 represents the presence of more free hydroxyl ions in water. Moreover, the higher pH level can be attributed to the increase in temperature.

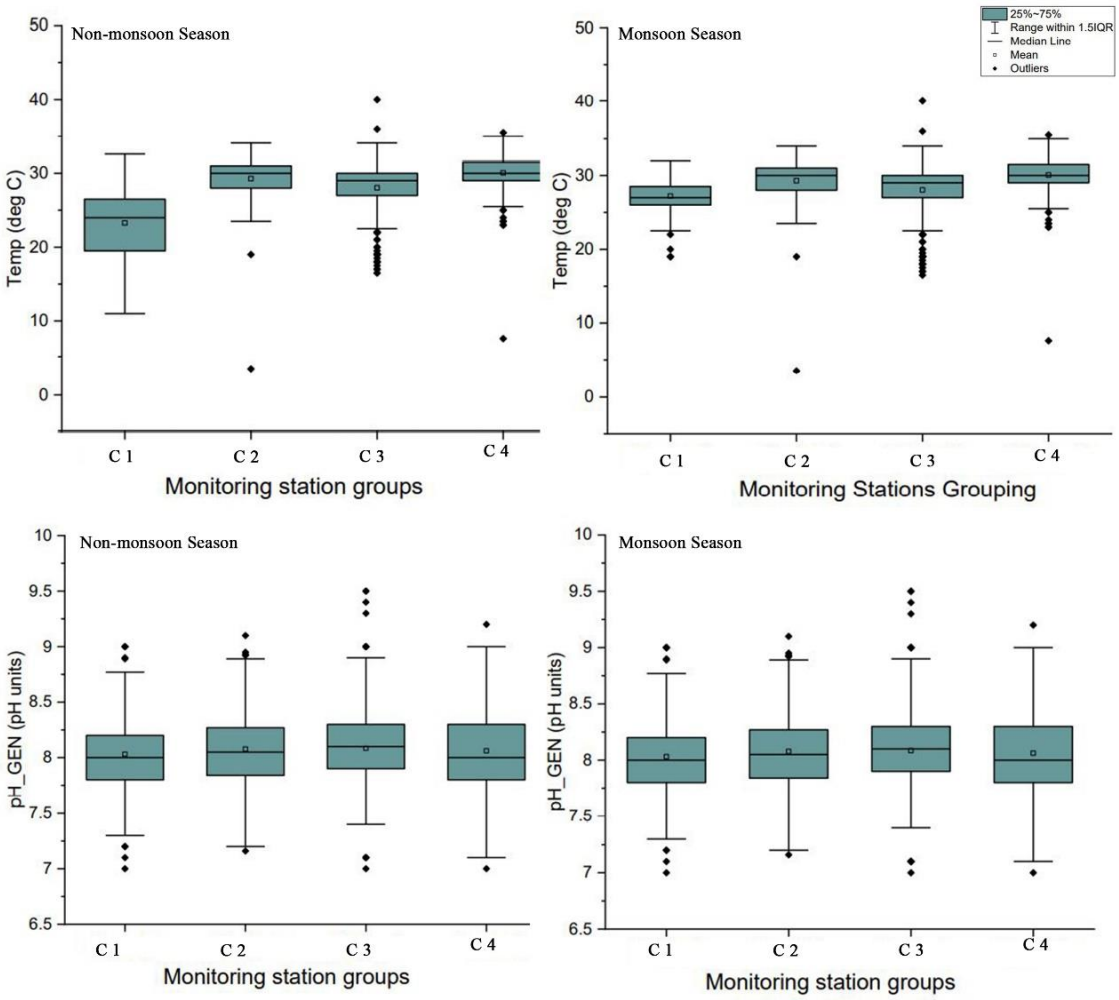


Figure 5.7 Spatiotemporal pattern for Temp and pH along different clusters

As per IS 2296:1992, the maximum BOD value for other criteria should be less than 3 mg/l, while the measured value is 0.2-16.0 mg/l (CWC and NRSC 2014). The DO level in water is greatly affected by the content of BOD, and higher BOD values can be related to the lower DO range of the water. About 40-75% of data crosses BOD's desired limits. A high BOD value indicates the faecal contamination from urban land use discharges into the river (Shukla et al. 2014). BOD is a crucial indicator in estimating the amount of organic matter in river water quality. BOD in monsoon season shows lower values than a non-monsoon season, as this could be due to the dilution of effluents in monsoon season (Sundaray et al. 2006). The COD along the study area showed a higher mean value at C2 in both seasons. The higher COD values, commonly observed in the basin with top agricultural practices, could be attributed to the use of chemical fertilizers (Shukla et al. 2014) and indicates the pollution strength due to

industrial and sewage waste through industrial, agricultural and urban run-off. Spatiotemporal patterns of BOD, COD and DO are presented in Figure 5.8.

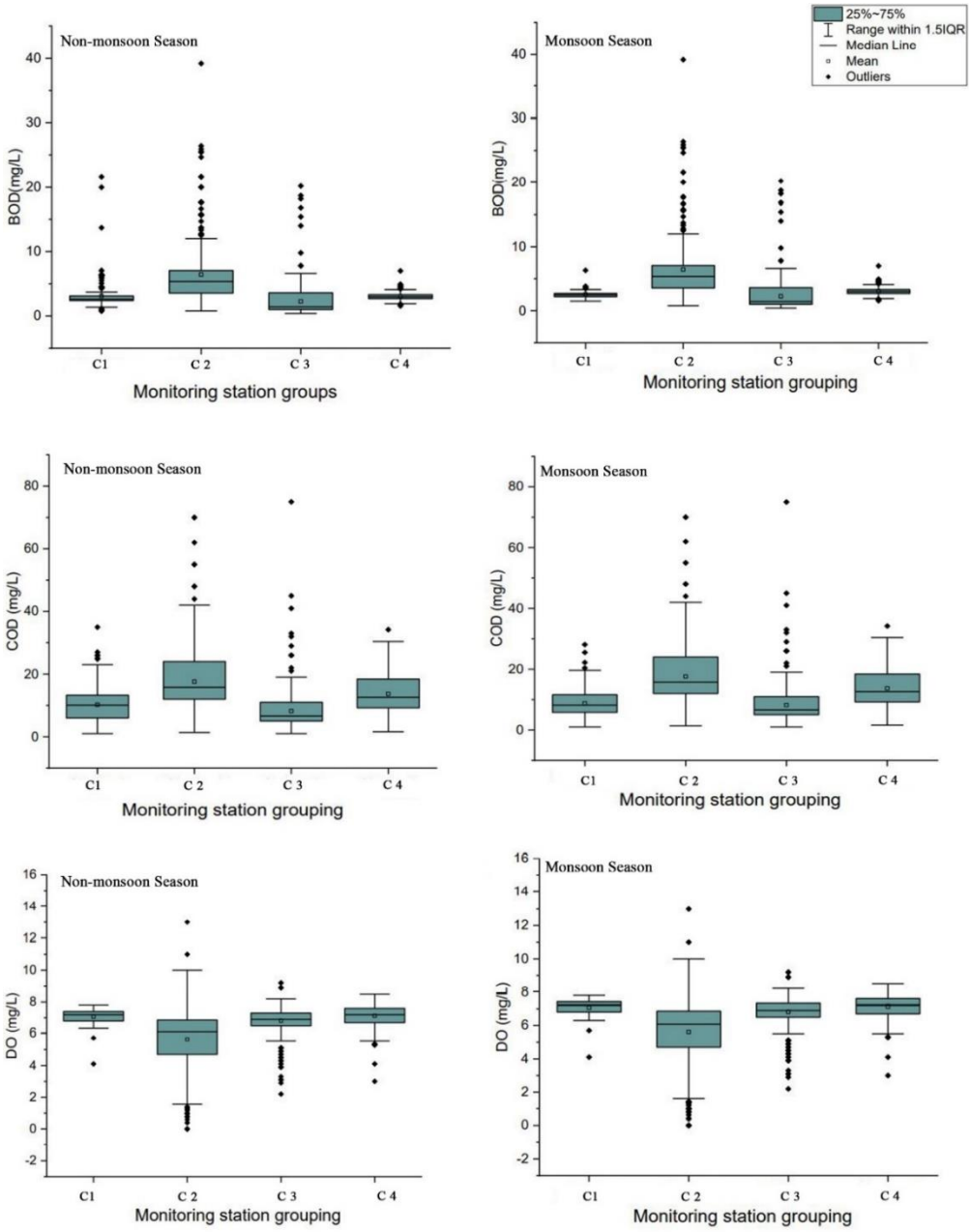


Figure 5.8 Spatiotemporal pattern for BOD, COD and DO along different clusters

The DO on the other hand is ranged between 1.50-10.0 mg/l in both seasons. A lower value of DO is observed along all the clusters during the non-monsoon and monsoon seasons, specifically >2 mg/l along C2 in the non-monsoon season. Typically, DO values are lower in the non-monsoon season when water temperatures are high compared to the monsoon season. Additionally, lower DO values indicate high organic

point sources and NPS, such as sewage from urban areas, industrial effluents, and agricultural run-off. Moreover, a high photosynthesis rate can also directly correlate with low DO. Concentrations below 4.0 mg/l will also adversely affect aquatic life (Tibebe et al. 2019).

5.2.3 Data reduction and Feature selection

The study examined and eliminated features with low variance and high correlation in this section. Specifically, the eigenvectors with the lowest eigenvalues in the dataset contained the least information about the data's variance and were thus eliminated. Eventually, three PCs with eigenvalue >1 are loaded (Singh et al., 2004; Razmkhah et al., 2010) (listed in Table 5.4) and are validated separately for different seasons using the explained variance ratio, which is presented in Figure 5.9.

Table 5.4 Loading of 20 WQPs on three significant PCs for non-monsoon and monsoon season

Variables	Non-monsoon Season			Monsoon Season		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
EC	0.402	**0.858***	-0.033	**-0.977***	-0.463	0.020
pH	**0.786***	-0.686	0.385	**-0.782***	0.403	0.226
TDS	-0.324	-0.354	-0.085	**0.989***	-0.180	-0.316
Temp	**0.789***	-0.321	-0.385	0.633	0.056	-0.316
Ca	-0.196	-0.785	-0.628	0.624	0.491	-0.605
Cl	0.639	-0.451	0.393	-0.184	0.456	0.464
CO_3	-0.485	-0.623	-0.279	**0.766***	0.184	-0.017
F	-0.104	0.492	0.611	0.554	-0.624	0.597
HCO_3	-0.048	0.458	0.573	0.392	-0.605	0.486
K	-0.035	-0.312	-0.355	**0.922***	-0.267	-0.155
Mg	0.361	-0.544	0.229	0.115	0.616	0.057
Na	0.240	**-0.965***	0.532	-0.561	**0.754***	0.453
NH_3-N	0.531	0.623	-0.346	-0.338	-0.118	-0.482
NO_2+NO_3	**-0.927***	0.316	0.654	**0.893***	-0.652	0.650
P-Tot	0.398	**0.978***	-0.191	**0.900***	-0.592	-0.111

SiO2	0.293	-0.463	0.066	-0.234	**0.937***	0.020
SO4	0.264	**0.903***	0.054	-0.453	-0.524	0.068
BOD	**0.752***	0.492	-0.295	**-0.803***	-0.086	-0.382
COD	**0.821***	-0.517	-0.645	**0.902***	0.001	-0.425
DO	**-0.727***	**-0.965***	-0.253	**-0.921***	-0.639	-0.231
Eigenvalue	5.641	4.777	1.1711	6.999	4.322	1.425
Variance%	36.14	28.85	10.7	44.220	27.330	9.020
Cumulative %	36.14	64.99	75.69	44.220	71.550	80.570

*Above 0.70 PC scores are marked as bold**

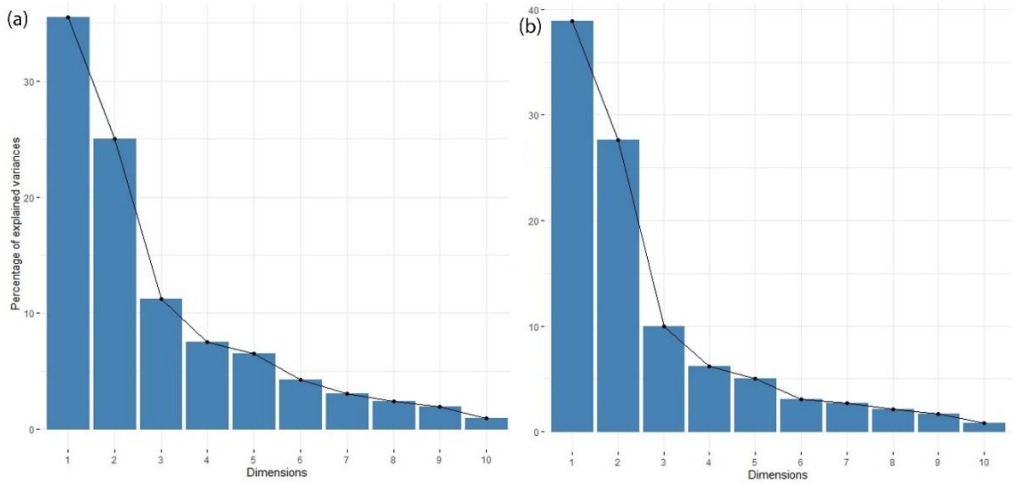


Figure 5.9 Explained variance ratio for non-monsoon (a) and monsoon (b) seasons

PC1 explained 36.14% of the data variance during the non-monsoon season. A positive loading (>0.70) was observed for pH, Temp, BOD and COD and (-ve >0.70) for NO_2+NO_3 and DO. The high -ve loading of NO_2+NO_3 and P-Tot indicates less contribution of agriculture and urban runoff in the non-monsoon season. Temp and DO show a moderate inverse proportion to each other. The moderate to high BOD and COD indicates the discharge of industrial and domestic point source of pollution. Aside from PC1, PC2 also resulted in a variance of 28.85%, which is significant. The higher positive contribution of EC, P-Tot, and SO_4 (0.858, 0.978, and 0.903, respectively) indicate a robust anthropogenic activity in the form of point source and NPS pollution along the basin. In comparison, PC3 explained 10.7% of the variance with no significant associations with any parameters. Nevertheless, the loading of PC1 and PC2 shows a moderate to high negative loading of DO during the non-monsoon season.

Further investigation may be necessary to determine the origin and thresholds of these pollutants to implement the most effective management strategy. The moderate to high levels of pH, BOD, COD and NO_2+NO_3, EC, P-Tot, and SO_4 along PC1 and PC2 also indicate anthropogenic factors interventions. However, a similar trend was observed during the monsoon season, with total variance exceeding that of the non-monsoon season (80.57%). PC1, PC2, and PC3 explained 44.220%, 27.33%, and 9.020% of the total variance. TDS, CO_3, K, P-Tot, NO_2+NO_3 and COD have higher significant contributions (> 0.70) for PC1. Specifically, we have observed a TDS score of 0.989, which is higher during the monsoon season and has a lower value during the summer (-0.32). The higher TDS levels in the river can harm agricultural, industrial, and domestic water users. During the monsoon season, however, many WQPs were discovered to have positive scores. Across PC2, a moderate to high connection with Na and SiO_2, scoring 0.754 and 0.937, respectively, is observed. The primary source of this could be the run-off from various LULCs. When the WQPs were compared seasonally, the monsoon season had a more significant impact. Furthermore, the scores obtained indicate a substantial influence of natural and anthropogenic interventions.

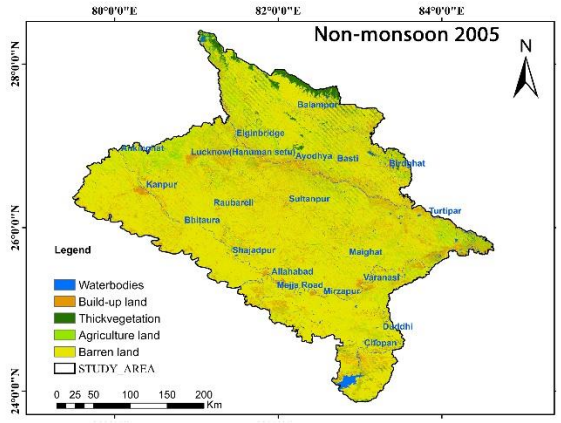
As a result, this study successfully applied PCA to withdraw redundant variables from the data without compromising much information. Furthermore, the complex 20 stations X 20 WQPs dataset is reduced to a lesser dimension of 20 stations X 3 PCs. Thus, dominant WQPs for 2005-2018 along MGB are identified using PCA. Based on the results and data availability, the WQPs EC, pH, Temp, TDS, NO_2+NO_3, P-Tot, BOD, COD and DO are chosen for further examination.

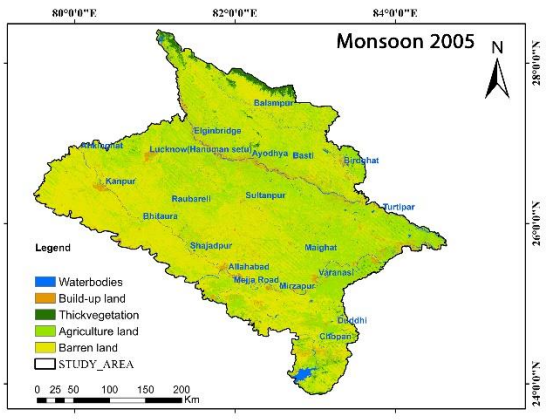
5.3 EVALUATION OF WATER POLLUTION DUE TO ANTHROPOGENIC CHANGES

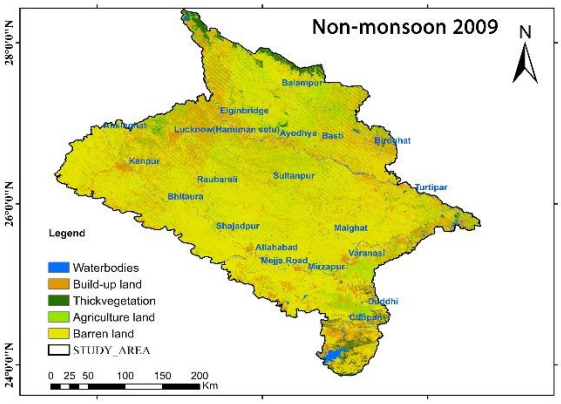
5.3.1 Land use land cover and change analysis

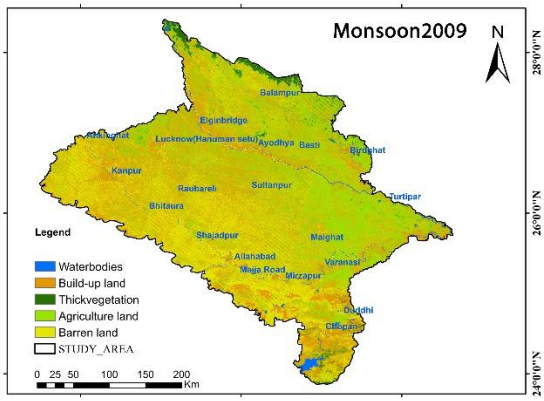
The final results of LULC are chosen based on the highest accuracies obtained from three algorithms, RF, SVM, and CART, for seasons from 2005 to 2018 (Figure 5.10). The selected image is classified by defining the region of interest (ROI) with points and polygons. Each class had approximately 80-90 ROIs considered for calibration and 60-70 ROIs for validation. Waterbody, built-up, barren land, agriculture and thick vegetation are the five classes selected in this study. The class water includes open

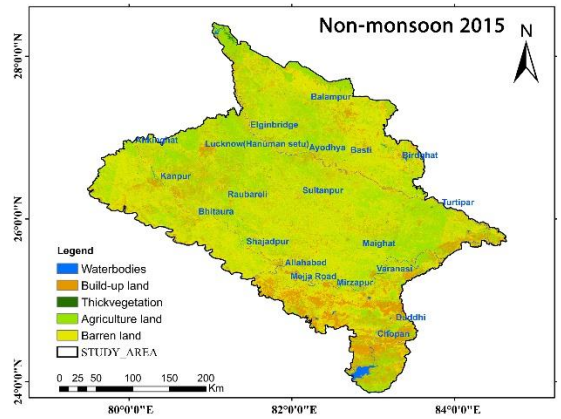
bodies of water, rivers, and ponds, while built-up land includes industrial, residential, and commercial developments, as well as roads, railways, and pavements. The bare land, open land, and quarries classes are all included in the barren land class. The class agriculture primarily considers cropland and plantations, whereas the class thick vegetation primarily considers reserve forests. Overall classification accuracy of 85.6-91% was attained with Kappa statistics of 0.89, suggesting an acceptable relationship between LULC and reference GCPs is observed (Kulithalai Shiyam Sundar and Deka 2021). Besides, the change analysis is also done for the year 2005-2009, 2009-2015, and 2015-2018 to identify the pattern. During the non-monsoon season, the built-up class increased by 7.5% from 2005 to 2018 (Table 5.5).

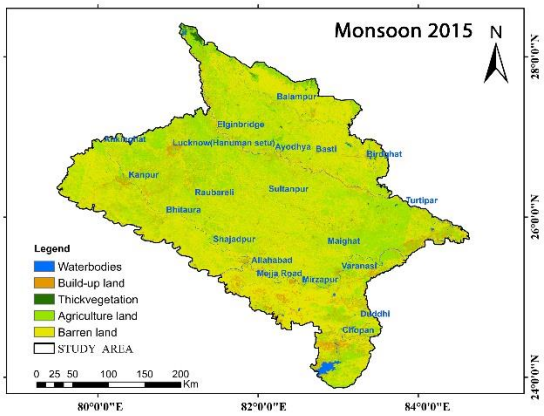

Non-monsoon 2005
80°0'0"E
82°0'0"E
84°0'0"E
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km


Monsoon 2005
80°0'0"E
82°0'0"E
84°0'0"E
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km


Non-monsoon 2009
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km


Monsoon2009
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km


Non-monsoon 2015
80°0'0"E
82°0'0"E
84°0'0"E
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km


Monsoon 2015
80°0'0"E
82°0'0"E
84°0'0"E
28°0'0"N
26°0'0"N
24°0'0"N
N
Balampur
Elginbridge
Lucknow(Hanuman setu)
Ayodhya
Basti
Birdghat
Kanpur
Raubareli
Sultanpur
Turtipar
Bhitaura
Shajadpur
Maighat
Allahabad
Mirzapur
Varanasi
Mejia Road
Daddhi
Citopani
Legend
Waterbodies
Build-up land
Thickvegetation
Agriculture land
Barren land
STUDY_AREA
0 25 50 100 150 200
Km

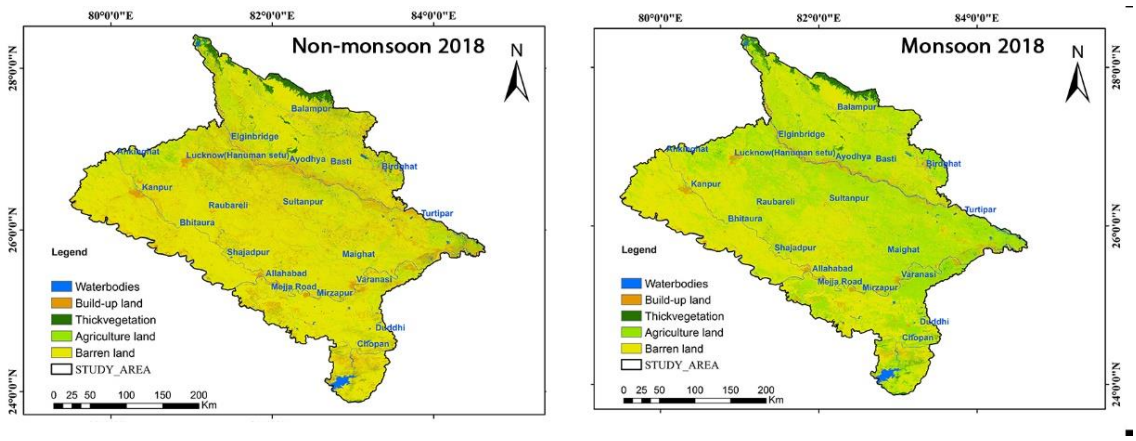


Figure 5.10 LULC classification non-monsoon and monsoon season

The thick vegetation has a coverage of 2.1% in the non-monsoon season in 2005, which was reduced to 1.9% in 2009, then to 1.4 and 1.3% in 2015 and 2018, respectively. In contrast, the agricultural land has increased from 54.6% to 58.1% from 2005-2018. On the other hand, water bodies decreased from 1.8% in 2005 to 1.3% in 2018. The findings suggest that except for agricultural lands and built-up areas, the other 3 LULCs chosen in this study have shown a decreasing trend.

Table 5.5 Non-monsoon Season % change in LULC from 2005-2018

LULC Types	**2005%**	**2009%**	**2015%**	**2018%**	**% Change 2005-2009**	**% Change 2009-2015**	**% Change 2015-2018**
Water bodies	1.8	1.7	1.1	1.3	-0.1	-0.6	-0.2
Build-up	14.3	19.9	21.2	21.8	5.6	1.3	0.6
Thick vegetation	2.1	1.9	1.4	1.3	-0.2	-0.5	-0.1
Agriculture Land	54.6	56.0	57.9	58.1	1.4	1.9	0.2
Barren Land	27.2	20.6	18.3	17.6	-6.6	-2.3	-0.7

Table 5.6 Monsoon Season % change in LULC from 2005-2018

LULC Types	2005%	2009%	2015%	2018%	% Change 2005-2009	% Change 2009-2015	% Change 2015-2018
Water bodies	1.9	1.7	1.7	1.9	-0.2	0.0	0.2
Build-up	14.2	19.2	20.9	21.3	5.0	1.7	0.4
Thick vegetation	2.7	2.2	1.7	1.4	-0.5	-0.5	-0.3
Agriculture Land	55.0	56.7	57.0	58.0	1.7	0.3	1.0
Barren Land	26.2	20.3	19.5	17.4	-5.9	-0.8	-2.1

During the monsoon season, there is a slightly higher percentage of dense vegetation and waterbodies. In 2005-2009, 2009-2015, and 2015-2018, build-up land use changed by 5%, 1.8%, and 0.4%, respectively (Table 5.6). Similarly, higher percentage changes in barren land are observed, with the highest being around -5.9% from 2005 to 2009 (Figure 5.11).

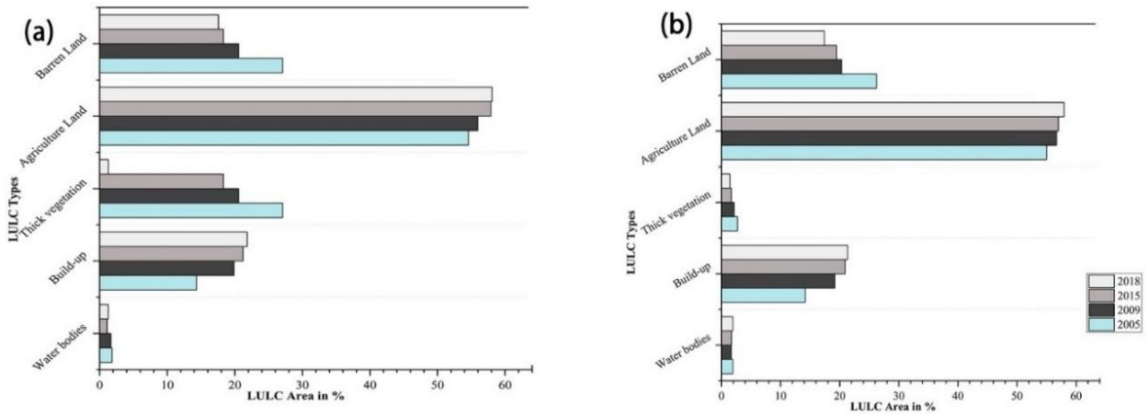


Figure 5.11 LULC area in percentage from 2005-2018 for monsoon (a) and non-monsoon (b) season

5.3.2 Effects of land use land cover pattern on different scales among clusters

The connections between LULC metrics and WQPs are investigated using stacked ensemble modelling (SEM), Redundancy Analysis (RDA) and Correlation analysis. The LULC and WQPs are modelled along each cluster to identify the feature of

importance, and finally, SEM was created based on the features that had been identified. Each cluster is studied seasonally among different scales, as discussed in the following section.

5.3.2.1 Effect of LULC on WQPs along Cluster1 Non-monsoon and monsoon season at different scale

Along C1, predictors explained more than 53% of the alterations in WQPs during the monsoon seasons and 29% during the non-monsoon season for the watershed scale. Contributors of LULC showed varying patterns on the catchment to reach and riparian scale. The predictors described 58% and 54% variation at the reach scale throughout the monsoon and non-monsoon seasons (Table 5.7). However, when the scale was reduced to riparian, this increased to 65% and 60% in monsoon and non-monsoon seasons, respectively (Table 5.7). Besides that, the explanatory ability decreased as the scale was increased to the catchment. A similar trend is observed in Pearson's correlation analysed at different scales ($P < 0.05$) (Figure 5.12 for C1). This suggests that predictors have a more significant impact along the river's banks.

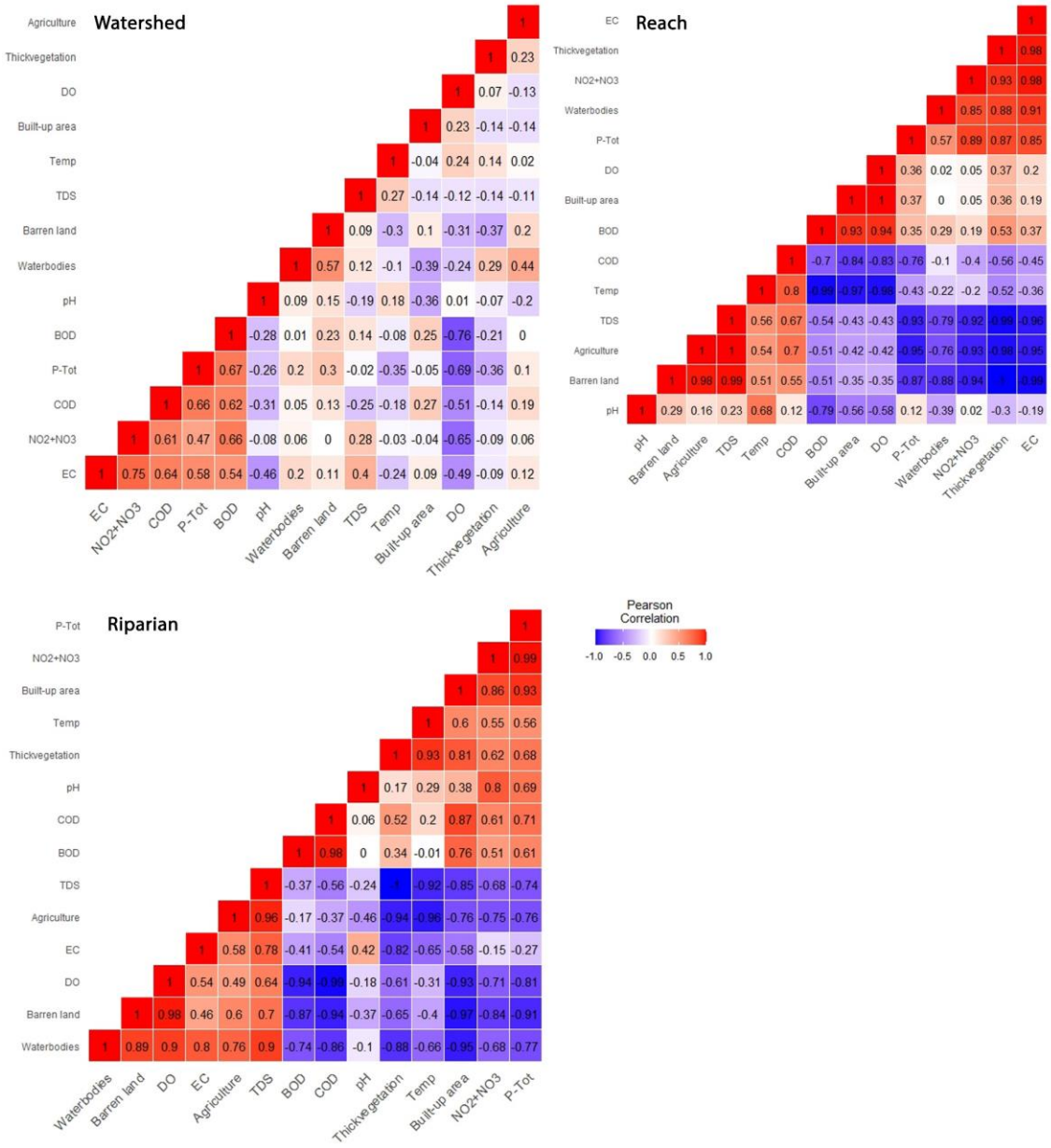


Figure 5.12 The correlation coefficient between WQPs and LULC along C1 at different scale for the monsoon season

Referring to RDA results in Figure 5.13, a strong relationship between build-up area and P-Tot on the riparian scale, agriculture with temperature, and BOD with COD can be seen. We can also observe that the relationship between predictor variables and response variables varies with scale. However, thick vegetation at all scales has not been associated with WQPs at any scale.

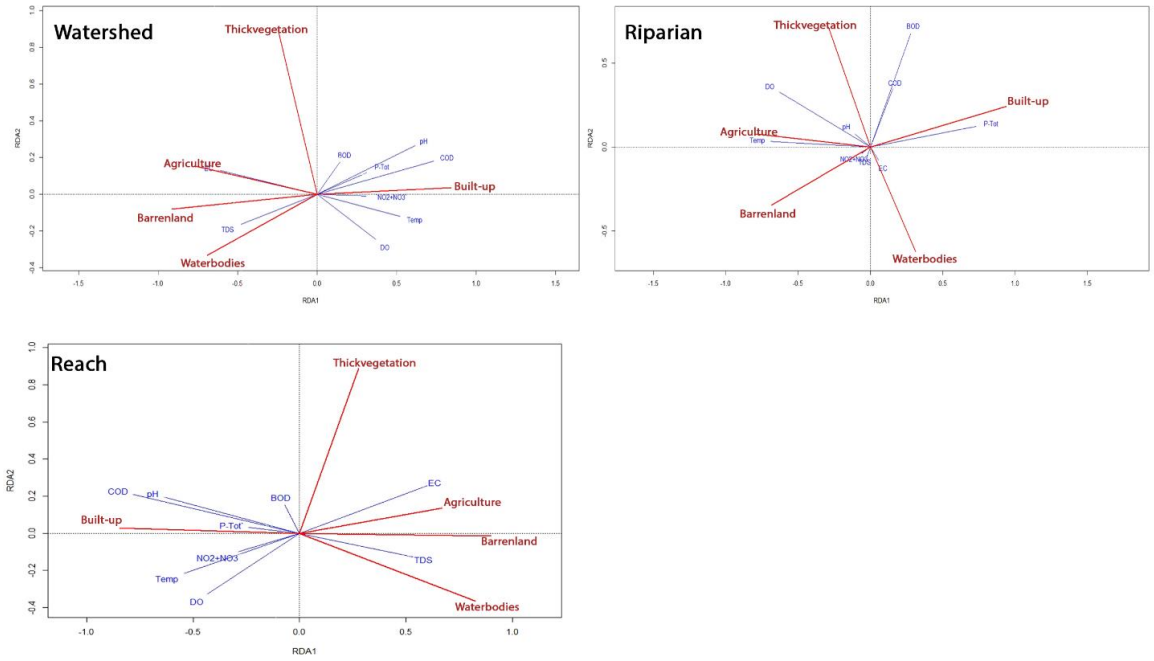


Figure 5.13 Association between WQPs and LULC at different scales as per RDA for C1 in monsoon season

Agriculture and barren LU were more likely to produce NPS pollution along this cluster than other classes. Though the contribution of built-up land ranks second (Table 5.7), which is consistent with the results of LULC classification and correlation analysis, as well as the results published in the Water Resource Information System (WRIS) (WRIS 2022). C1 includes Allahabad, Maighat, Mirzapur, Shahzadpur, Sultanpur and Varanasi. Due to urban pockets receiving a considerable quantity of treated and untreated wastewater directly into the river course, major cities such as Allahabad and Varanasi contributed to high levels of BOD. They were designated as unfit for outdoor bathing (Dutta et al., 2020), which can be true when referring to the riparian scale RDA. Due to the river's proximity to the highly urbanised city of Allahabad, significant deterioration in water quality was observed at several locations (WRIS 2022). In addition, the monitoring stations Ankinghat (Kanpur), Chhatnag (Allahabad), and Varanasi are located in the Gangetic plains, which are characterised by high anthropogenic activity (Kumar Shukla et al. 2018). Besides that, the major agriculture and barren land dominant watersheds are Maighat, Mirzapur, Shahzadpur, and Sultanpur may also be associated with high NPS loading. Moreover, it indicates the role of significant point source pollutants in the river stretch along C1.

During the non-monsoon season (Figure 5.14), a higher contribution of P-Tot is observed along different scales, and thick vegetation, on the other hand, has no association with any WQPs. This result demonstrate direct sewage disposal from the build-up area along the basin (Álvarez-cabria et al. 2016). Because of the impervious surfaces of metropolitan environments, even moderate rain events can create a wide range of pollutants. Moreover, the contributions of different LULCs along different scales can also be seen based on their length.

DO and agricultural land observed a strong association along the riparian scale. Compared to different scales along the reach scale, a high to moderate association between P-Tot-built-up and DO-barren land is observed. The association of different LULCs on WQPs shows seasonal variation in C1. It is clear from the RDA that thick vegetation plays a vital role in improving the water quality along C1 in both seasons (Mello et al. 2018). Moreover, some of LULC and its association with WQPs point out that the variation in WQPs is irrespective of LULC present there. This could be attributed to studies indicating that sewage accounts for 70% of GRB pollution, industrial waste accounts for 20%, and NPS such as garbage in the river, open defecation, and agricultural run-off account for 10% (Namami Gange 2020).

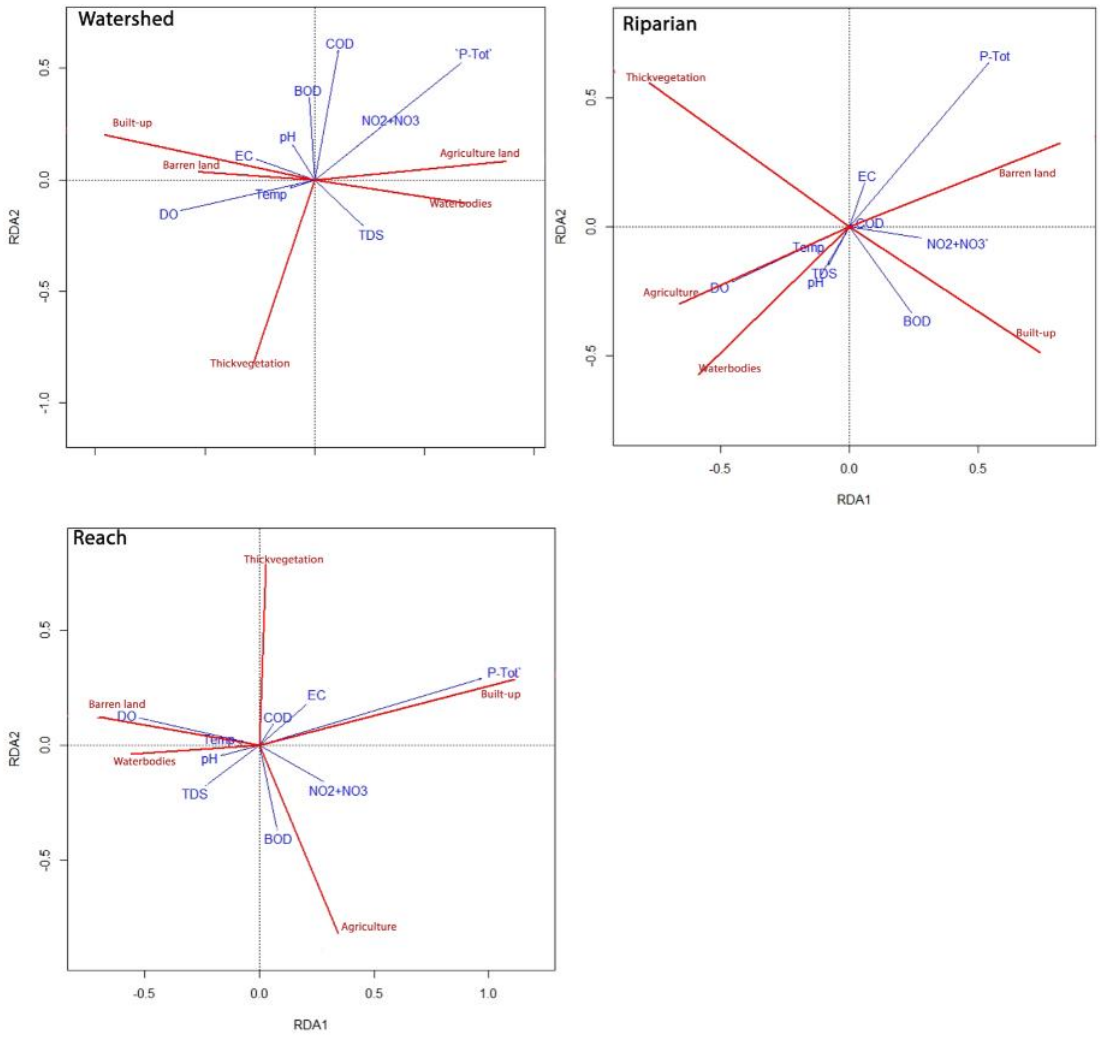


Figure 5.14 Association between WQPs and LULC at different scales as per RDA for C1 in non-monsoon season

5.3.2.2 Effect of LULC on WQPs along Cluster 2 Non-monsoon and Monsoon season at different scale

Ayodhya, Balrampur, Basti, Birdghat, Elginbridge and Turtipar are the monitoring stations present in C2, and agricultural land, barren land, and thick vegetation are the significant land use present along this cluster. The RDA results along the catchment, reach, and riparian scale indicated the association of EC, COD, Temp, TDS, P-Tot and DO with agriculture land use. At the same time, thick vegetation and waterbodies are not associated with any WQPs. The predictors during the monsoon season RDA explained 40% of the variance in WQPs, whereas it decreased by 5% during the non-monsoon season for watershed scale. At the reach scale, the predictors explained 69%

and 54% during the monsoon and non-monsoon seasons, respectively, but on an increased scale, i.e., riparian, this became 68% and 52% in the respective seasons (Table 5.7). This could be attributed to the higher agricultural practices along the banks of the river. Besides that, traditional farming tends to increase the risk of soil erosion, and high nutrient usage could indeed result in NPS pollution through stream water due to surface runoff (Shi et al. 2017). As observed in C1, dense vegetation does not affect WQPs and can be seen as cleaner along this cluster. From Table 5.7, we could see higher explained variance at the reach scale, with agriculture and barren land use being the most associated with WQPs.

5.3.2.3 Effect of LULC on WQPs along Cluster 3 Non-monsoon and Monsoon season at different scale

The stations along cluster C3 are Dudhi, Chopan and Maeja, no difference in scale effect during the non-monsoon and monsoon seasons is observed here. EC, pH, TDS and Temp had high loading along RDA1 and slightly less in RDA2. During the non-monsoon season, 34.56% variance is explained by RDA1 and 20% more in the monsoon season (Table 5.7). This indicates the impact of NPS of pollution on water quality degradation is higher than the point source in the non-monsoon season. The relationship between LULC and WQPs has shown a similar trend as in other clusters, though much less variation is explained here. This could be attributed to the presence of denser vegetation than in other clusters. Although, during the monsoon and non-monsoon seasons thick vegetation, barren, and agriculture land have explained better variations at reach and riparian scales. This cluster includes the stations Dudhi, Chopan, and Majea, which have a high percentage of deciduous forest, degraded/scrubs, which are classified as thick vegetation here, and barren land. Vegetation contributes substantially to NPS nutrient trapping, and vegetation cover is inversely linked with most WQPs, presumably due to reduced soil erosion (Shi et al. 2017). As a result, the WQP loading at RDA1 and 2 could be traced to a point source pollution.

5.3.2.4 Effect of LULC on WQPs along Cluster 4 Non-monsoon and Monsoon season at different scale

During the non-monsoon season, the catchment scale explains approximately 67% of the variance, while the reach and riparian scales explain 68%. Agricultural and barren land is the most contributing land use along this cluster (Table 5.7). The total explained variance remains constant at the reach and riparian scales during both seasons but is greater during the non-monsoon season. The influence of Lucknow and Kanpur along this cluster during the non-monsoon season (Figure 5.4) could be linked to higher pollution loading during the non-monsoon season. Domestic and municipal sewage run-off from agricultural lands and industrial effluents from distilleries, agrochemicals, sugar mills, paper/pulp, etc., are left untreated and discharged directly into the river from Lucknow (Singh et al. 2004).

On the other hand, Kanpur is a prominent city well-known for its textile and leather industries. As these industries are on the river's bank, their effluent reaches GRB. The RDA results (Figure 5.15 & Figure 5.16) show that land uses like agriculture, built-up areas, and barren land significantly impact water quality degradation. In contrast, as discussed in other clusters, thick vegetation with less loading controls water quality degradation.

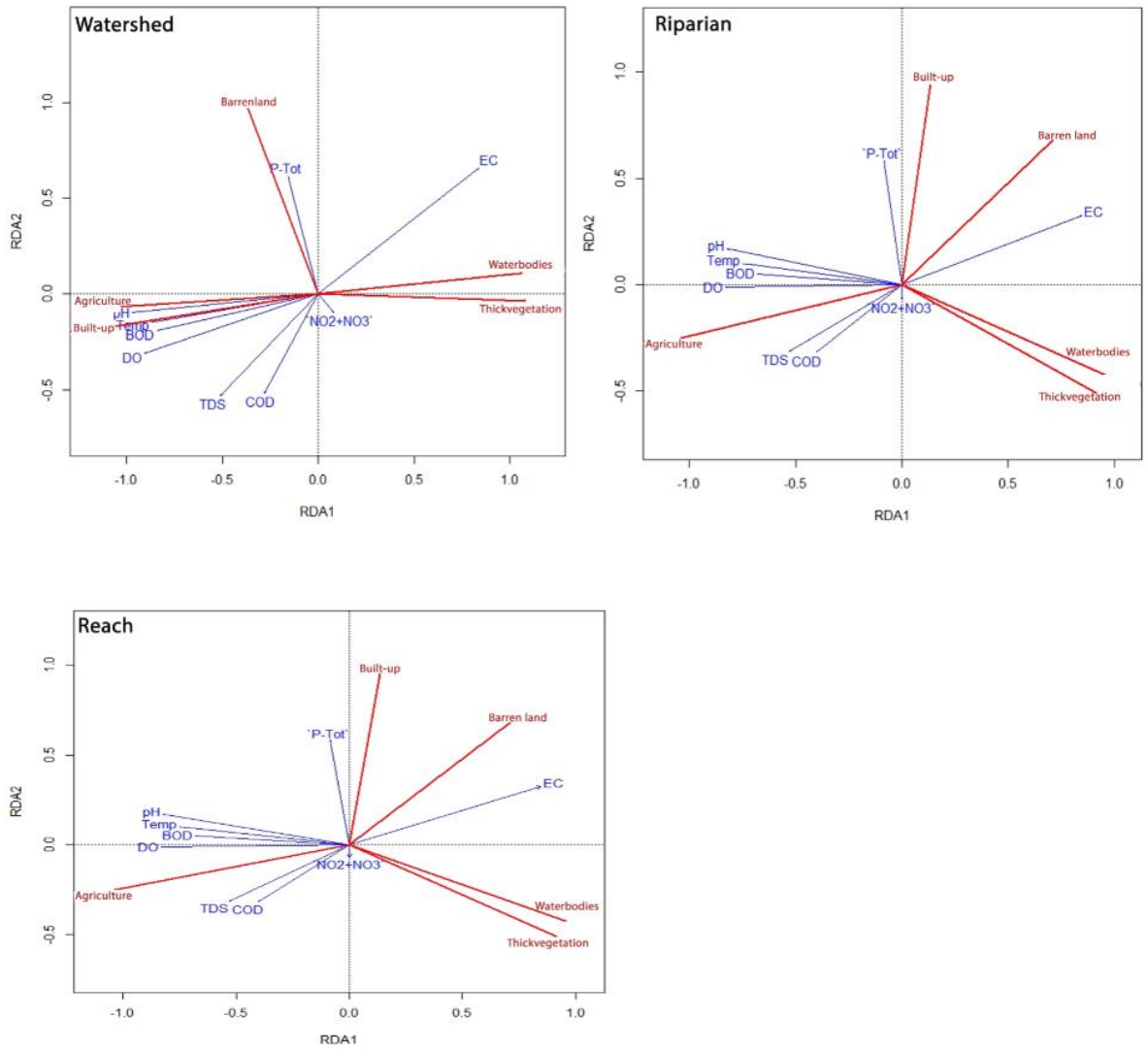


Figure 5.15 Association between WQPs and LULC at different scales as per RDA for C4 in Non-monsoon season

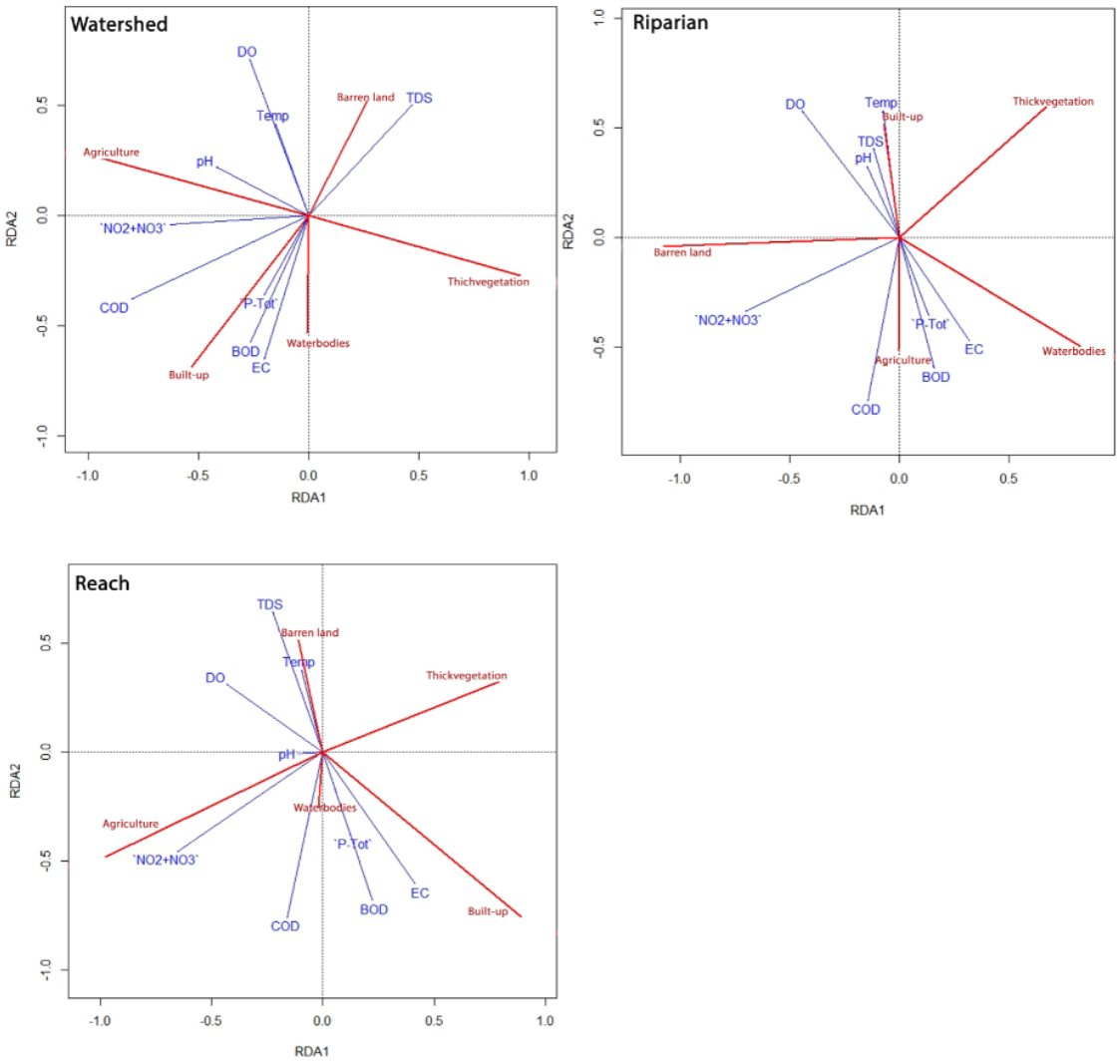


Figure 5.16 Association between WQPs and LULC at different scales as per RDA for C4 in Monsoon season

The study then uses an SEM regression algorithm to assess the predictive capability of various LULCs with WQPs.

Table 5.7 Multi-scale explanations (in %) for different LULC classes on WQPs

Clusters	Season	Scale/LULC Classes	Waterbody	Build-Up Land	Thick vegetation	Agriculture Land	Barren Land	Total
C1	Non-Monsoon	Watershed	0.96	1.89	2.2	13.5	10.46	29.0
		Reach	2.03	10.05	3.9	19.5	18.52	54.0
		Riparian	**1.12**	**10.53**	**4.23**	**20.73**	**23.39**	**60.0**
	Monsoon	Watershed	1.05	10.45	2.43	21.93	17.14	53.0
		Reach	2.4	10.02	2.93	24.37	18.28	58.0
		Riparian	**2.56**	**10.05**	**6.03**	**26.32**	**20.04**	**65.0**
C2	Non-Monsoon	Watershed	1.57	5.73	3.5	13.15	11.05	35.0
		Reach	1.82	5.13	4.2	24.82	18.03	54.0
		Riparian	1.52	4.28	4.05	24.79	17.36	52.0
	Monsoon	Watershed	2.57	6.73	4.5	14.15	12.05	40.0
		Reach	**2.82**	**7.26**	**5.28**	**32.61**	**21.03**	**69.0**
		Riparian	**1.52**	**4.28**	**4.05**	**24.79**	**17.36**	**68.0**
C3	Non-Monsoon	Watershed	2.45	1.19	11.86	9.51	9.59	34.6
		Reach	**2.45**	**0.93**	**30.59**	**16.7**	**19.33**	**70.0**

		Riparian	**2.45**	**0.93**	**33.59**	**16.7**	**19.33**	**73.0**
	Monsoon	Watershed	2.45	1.19	21.86	10.51	18.59	54.6
		Reach	**2.45**	**0.93**	**31.78**	**17.84**	**19**	**72.0**
		Riparian	**2.45**	**0.93**	**33.59**	**16.7**	**19.33**	**75.0**
C4	Non-Monsoon	**Watershed**	**2.78**	**10.58**	**1.56**	**30.3**	**21.78**	**67.0**
		Reach	**1.23**	**7.58**	**1.69**	**34.5**	**23**	**68.0**
		Riparian	**1.56**	**8.93**	**1.72**	**36.78**	**19.01**	**68.0**
	Monsoon	**Watershed**	**2.78**	**7.64**	**3.63**	**26.99**	**19.96**	**61.0**
		Reach	1.26	6.8	1.6	23.34	21	54.0
		Riparian	1.58	7.53	1.1	23.78	20.01	54.0

60% and above values are marked in bold

5.3.2.5 Stacked Regression Models to Predict water quality of LULC along different clusters

Although the SEM's predictive powers (Table 5.8) vary significantly (R^2 = 0.12 to 0.96), the model validation performance was good, with a slight difference between predicted and measured values. Except for C1, the best models for EC were identified at the riparian scale, with a predictive score of 0.6 to 0.95 at other clusters, with C4 being the highest. EC is critical for determining the water quality of a specific area. As conductivity measures dissolved ionic concentration, it is a baseline for measuring other WQPs (Ahmad et al. 2021). A substantial distinction in the conductivity value indicates a change in water quality, the presence of contaminants, or some pollution source impacts the area. Agriculture and build-up land are essential predictors of EC, and the abundance of these land use in these clusters indicates a surge (Mello et al. 2020). pH has an excellent predictive score along the watershed scale for the monsoon season, along the riparian scale for the non-monsoon season, and also the highest (R^2 = 0.83) for C1 along the riparian scale for the monsoon season. Point source pollution is a prevalent cause that can increase or decrease pH based on the chemicals present.

Furthermore, the contaminants emitted by agricultural runoff, domestic sewage, or industrial runoff could cause a significant change in pH levels due to the land use present within the catchment. The TDS model exhibited varied scale effects during both seasons, with the highest being explained at riparian for C4 (R^2= 0.85), a dominant agriculture region during the monsoon season. The presence of severe anthropogenic sources along the river's course and runoff with elevated suspended matter could be attributed to the high TDS concentration in the rivers. Temp, on the other hand, mostly displayed a uniform scale effect, with the riparian scale being most predominant (R^2= 0.36-0.96). The highest prediction capacity is observed for C4 (R^2= 0.96) during the non-monsoon season. NO_2+NO_3 is a significant criterion of eutrophic waters, especially those contaminated by fertiliser run-off, domestic sewage and animal waste (Álvarez-cabria et al. 2016; El-Zeiny and El-Kafrawy 2017). It should be noted that along all clusters, the riparian scale had a better performance for the the NO_2+NO_3 model. Seasonally speaking, monsoon season had the highest value across all the clusters (C4 R^2= 0.86), which explains the high level of runoff from agricultural land use. The

variation at the scale level could be associated with highly increased bacterial activity coupled with increased nitrogen content and drainage water with higher levels of nutrients (Siqueira et al. 2015; El-Zeiny and El-Kafrawy 2017). Concurrently, P-Tot followed a similar pattern, with higher performance (R^2 = 0.9 in monsoon and R^2 = 0.84 in non-monsoon) all over the riparian and seasonally higher values in the monsoon season. P-Tot, the primary source of urban runoff, had higher R^2 estimates along C1 (R^2 = 0.8) and C4 (R^2= 0.9) for the riparian scale, which is clustered with significant cities as discussed earlier. Unplanned urban growth, intensive farming exercises, and deforestation are all associated with higher water quality levels of phosphorus and nitrogen (Afed Ullah et al. 2018). BOD is a direct indicator of the level of pollution in waterbodies. Seasonally, all the clusters has exbhitted a similar trend for BOD with higher R^2 at the riparian scale. These levels indicate a significant load of organic matter release in segments that receive a large quantity of agricultural as well as sewage waste from multiple sources, which were believed to be due to the abundant supply of microbes and microbes activity present in the region (El-Zeiny and El-Kafrawy 2017). The clusters C1, C2 and C4 had values ranging from R^2 of 0.8 to 0.85 for different seasons. The majority of COD clusters and all DO clusters had good riparian scale performance, with C4 having the highest (R^2 = 0.95 for COD and 0.87 for DO) for the monsoon season. Moreover, in both cases monsoon season performed better than the non-monsoon. Lower values of DO (<2 mg/L) are reported along Kanpur and Lucknow, which fall in C2 and C4. The value of BOD continues to remain reasonably stable throughout both seasons within most locations, and the DO values are lower during the monsoon season.

Table 5.8 Coefficient of determination R^2 for different scales and seasons

WQPs	Scale	Monsoon Season				Non-Monsoon			
		C1	C2	C3	C4	C1	C2	C3	C4
EC	Watershed	0.8	0.45	0.61	**0.85**	0.53	0.53	0.55	**0.82**
	Reach	0.76	0.55	0.64	**0.87**	0.55	0.55	0.57	**0.81**
	Riparian	0.55	0.63	0.72	**0.95**	0.6	0.6	0.61	**0.83**
pH	Watershed	0.7	0.6	0.76	0.75	0.72	0.72	0.65	0.76
	Reach	0.72	0.55	0.71	0.72	0.72	0.72	0.75	0.78

	Riparian	**0.83**	0.55	0.71	0.64	0.73	0.73	0.77	0.73
TDS	Watershed	**0.82**	0.53	0.73	**0.81**	0.5	0.5	0.46	**0.8**
	Reach	**0.84**	0.5	0.74	**0.83**	0.46	0.46	0.5	0.73
	Riparian	0.68	0.48	0.76	**0.85**	0.43	0.43	0.55	0.74
Temp	Watershed	**0.81**	0.48	0.36	**0.81**	0.48	0.48	0.36	**0.86**
	Reach	**0.85**	0.55	0.38	**0.84**	0.55	0.55	0.38	**0.87**
	Riparian	**0.84**	0.6	0.45	**0.84**	0.6	0.6	0.45	**0.96**
NO2+NO3	Watershed	0.46	0.27	0.14	**0.81**	0.7	0.44	0.27	0.52
	Reach	0.56	0.32	0.12	**0.83**	0.66	0.46	0.32	0.57
	Riparian	0.75	0.5	0.12	**0.86**	0.72	0.48	0.43	0.6
P-Tot	Watershed	0.75	0.47	0.35	**0.85**	**0.82**	0.57	0.32	0.76
	Reach	0.75	0.56	0.37	**0.85**	**0.84**	0.58	0.34	0.77
	Riparian	**0.8**	0.62	0.42	**0.9**	**0.83**	0.59	0.35	**0.81**
BOD	Watershed	0.75	0.73	0.55	**0.8**	0.72	0.72	0.52	0.75
	Reach	0.78	0.74	0.56	**0.85**	0.73	0.73	0.54	0.76
	Riparian	**0.88**	**0.85**	0.58	**0.88**	**0.8**	**0.8**	0.55	0.77
COD	Watershed	0.72	0.54	0.45	**0.94**	**0.81**	0.51	0.32	0.63
	Reach	0.6	0.63	0.46	**0.95**	0.62	0.62	0.32	0.74
	Riparian	0.69	0.7	0.47	**0.95**	0.65	0.65	0.33	0.74
DO	Watershed	0.85	0.78	0.37	**0.87**	0.75	0.75	0.45	**0.81**
	Reach	0.85	0.79	0.56	**0.8**	0.77	0.77	0.56	**0.81**
	Riparian	**0.87**	**0.82**	0.57	**0.87**	**0.82**	**0.82**	0.57	**0.84**

Above 0.80 R^2 scores are marked as bold

Overall, the RDA and SEM results indicate that agricultural land, barren land, and human settlement close to the river bank are going to have a serious influence on WQPs. The majority of forest tracts in the basin have been seriously hampered due to over-exploitation. As an outcome, the GRB forest ecosystem is under severe stress (Consortium of 7 IITs 2012). The root cause of riparian biodiversity degradation is both natural and anthropogenic. Construction works, the agricultural land expansion for food, and grazing pressure is some of the major concerns. Most WQPs are not connected with the thick vegetation class, presumably due to less soil erosion (Singh

and Mishra 2014). Riparian vegetation restoration is critical for maintaining and improving stream water quality because the vegetation cover along this stretch of the river reduces pollution load. Agricultural land use significantly impacts the sediment and nutrient levels in the waterbody. Rising inflows of organic manure, inorganic fertiliser, and pesticides are the key variables affecting water quality in agricultural areas (Ding et al. 2016). The results show that the impact of LULC patterns on WQPs varies with season and spatial scale. This implies that water quality management is primarily a regional problem. As a result, water quality management and land use planning must take a multi-scale approach.

5.4 MAPPING THE CONCENTRATION OF WQPS USING LANDSAT-8 AND MACHINE LEARNING ALGORITHMS

Remote sensing has long been recognised as having the potential to supplement traditional lake monitoring methods after taking into account the failure of conventional *insitu* data and the analysis to explain the water quality problem on a finer spatiotemporal scale. In this part of the study, a machine learning approach for assessing spatiotemporal water quality is introduced.

5.4.1 Feature selection criteria

Statistical tests were performed on extracted R_{rs} data to check the inconsistency, and outliers were removed or corrected by applying a z-score (Sudheer et al. 2007). To identify the best features for modelling, the Pearson correlation matrix (r) between Landsat-8 *Rrs* values on different bands and band ratios with WQPs is investigated at various stations. In this study, multi-spectral bands and their combinations with correlation (i.e. $r \geq 0.50$) were selected to form the input layer (Sharaf El Din et al., 2017; Hafeez et al., 2019). A maximum of 50% significance level or $p<0.05$ is considered to finalize the input parameters (Nas et al., 2010; Swain & Sahoo, 2017a; Abdelmalik, 2018). As presented in Table 5.9, a significant ($p<0.05$) correlation of WQPs with bands B1-B4 except for EC is observed. A similar trend has been experienced with other stations as well. However, the rest of the Landsat-8 bands, such as Cirrus, thermal infrared 1 (TIR1), and thermal infrared 2 (TIR2), were less correlated (i.e. $r < 0.50$) within the WQPs. The lower r values achieved between TIR1 and TIR2

bands and WQPs are due to the fact that these bands are primarily designed to detect surface temperatures. Simultaneously, Cirrus is commonly used for cloud detection. The study also created many band combinations with significant correlations with WQPs to improve the relationship between input and output variables for the machine learning algorithm. The criteria for selecting features for different clusters are explained below.

The Pearson correlation technique and the ExtraTreesRegressor based on Gini importance have identified the best correlated (above 0.50) with significance ($p<0.05$) bands and band combination for the model input. Around 166 input parameters (not presented here), including bands and their different combinations, are identified for various stations with WQPs (Temp, EC, pH, SiO_2, and DO). A correlation of 0.567-0.923 is observed on different combinations with a significance of $p<0.05$. Feature importance scores are also identified through ExtraTreeRegressor and are plotted for features based on the Gini importance of various combinations.

Table 5.9 Pearson correlation between Rrs and WQPs

	B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**EC**	**pH**	**TDS**	**Temp**	**SiO_2**	**DO**
B1	1												
B2	**0.99**	1											
B3	**0.95**	**0.96**	1										
B4	**0.86**	**0.89**	**0.94**	1									
B5	0.29	0.29	0.36	0.49	1								
B6	0.16	0.16	0.16	0.33	**0.85**	1							
B7	0.16	0.16	0.14	0.32	**0.76**	**0.98**	1						
EC	**-0.55**	**-0.54**	**-0.62**	**-0.53**	-0.10	0.03	0.06	1					
pH	**0.53**	**0.51**	**0.63**	**0.51**	0.18	0.04	0.01	**-0.74**	1				
TDS	**0.54**	**0.53**	**0.60**	**0.57**	0.09	-0.03	-0.05	-1.00	**0.73**	1			
Temp	**0.57**	**0.56**	**0.62**	**0.53**	0.28	0.15	0.13	**-0.65**	**0.68**	**0.63**	1		
SiO_2	**0.53**	**0.52**	**0.62**	**0.51**	0.19	0.05	0.01	**-0.79**	**0.87**	**0.78**	**0.63**	1	
DO	**0.57**	**0.60**	**0.68**	**0.74**	**0.65**	0.04	0.01	**-0.61**	**0.89**	**0.61**	**0.56**	**0.77**	1

5.4.2 Hyperparameter optimization for XGBoost

The XGBoost models are developed using scikit-learn compatible API. The database for the model is first converted into an optimized data structure called Dmatrix, as this is the specific format that XGBoost can handle. Hyperparameters of XGBoost are then optimized by applying Optuna. Optuna employs a historical record of trial details to determine the promising search area in order to optimise the hyperparameters in less time. Learning rate, max_depth, l1_reg (L1 regularization term on weights), l2_reg (L2 regularization term on weights) and n_estimators are the hyperparameters are applied in this study. The pruning feature automatically stops the unpromising trails in the early stages of training and is also accounted for in the modelling process. One WQPs at a time as output is consider here because the best features identified for these were different. Although, the same hyperparameters are applied throughout all the stations for different WQPs. The optimized hyperparameters identified are displayed in Table 5.10.

Table 5.10 Optimized hyperparameters for different WQPs along different Clusters in XGBoost

Clusters	WQPs	Learning Rate	Max_depth	l1_reg	l2_reg
C1	pH	0.237338	4	0.2236841	0.00059588
	Temp	0.205234461	8	0.0115554	2.8569528
	SiO_2	0.11117548	7	0.0038749	0.92009494
	DO	0.1492115	8	0.0001574	0.18985823
	TDS	1.05713312	5	1.144E-05	1.65E-05
C2	EC	0.0356512	6	0.0001455	1.1245E-05
	pH	0.178377066	7	0.014976	0.00103462
	Temp	0.114854821	7	0.0001023	0.01760738
	SiO_2	0.558027582	7	3.582E-05	9.11E-05
	DO	0.267273993	6	0.0034017	0.23174842
	TDS	1.055113332	5	1.914E-05	1.12E-05
C3	EC	0.03212539	8	0.0001335	1.3344E-05
	pH	0.54051175	7	0.228148	0.00035278

	Temp	0.138631144	4	6.56E-05	0.01605706
	TDS	0.030726507	5	4.51E-05	1.51E-04
	SiO_2	0.084894633	4	0.0025398	9.96E-01
	DO	0.158258279	4	0.0054742	9.86053923
C4	EC	0.831185087	6	0.0001335	1.51E-04
	pH	0.097761003	5	0.005306	0.09855875
	Temp	0.114854821	7	0.0001023	0.01760738
	TDS	1.055113332	5	1.91E-05	1.12E-05
	SiO_2	0.558027582	7	3.58E-05	9.11E-05
	DO	0.480724584	5	0.3392282	0.00656407

5.4.3 Hyperparameter optimization for MLP

MLPRegressor is a multi-layer perceptron algorithm for regression tasks in scikit-learn's neural network module. It can train a neural network on input data and predict continuous target variables. It can also handle multiple hidden layers and various activation functions, allowing for a wide range of modelling capabilities. To optimize its hyperparameters, GridSearchCv is applied here. Grid search is a method of hyperparameter tuning that generates and assesses a model systematically for each combination of algorithm parameters supplied in a grid. The hyperparameters (Table 5.11) have been used by applying a 3-7 fold cross-validation to optimise the best estimator for this investigation. The ratio of train:test was changed from 70-80 until the accuracy for both training and testing became the same, or the difference was negligible. This procedure for the hyperparameters search is carried out for all the clusters by taking one WQPs at a time.

Table 5.11 Optimized hyperparameters for different WQPs along different Clusters in MLP regressor

Clusters	WQPs	Activation	Hidden Layers	Learning Rate	Solver
C1	EC	logistic	(50,150)	Constant	L-BFGS
	pH	identity	(150,100,50)	Constant	L-BFGS
	Temp	relu	(100,50)	Constant	L-BFGS
	SiO_2	relu	(150,100,50)	Constant	L-BFGS

	DO	relu	(100,50)	Constant	L-BFGS
C2	EC	relu	(150,50,100)	Constant	Adam
	pH	relu	(100,150,50)	Constant	L-BFGS
	Temp	tanh	(100,)	Constant	L-BFGS
	SiO_2	relu	(150,50,100)	Constant	L-BFGS
	DO	tanh	(150,100,50)	Constant	L-BFGS
C3	EC	relu	(50,100,150)	Constant	Adam
	pH	relu	(100,150,50)	Constant	L-BFGS
	Temp	tanh	(100,)	Constant	L-BFGS
	SiO_2	relu	(50,100,150)	Constant	L-BFGS
	DO	relu	(100,50,150)	Constant	L-BFGS
C4	EC	relu	(50,150,100)	Constant	L-BFGS
	pH	relu	(100,150,50)	Constant	L-BFGS
	Temp	logistic	(50,)	Constant	L-BFGS
	SiO_2	relu	(100,50,150)	Constant	L-BFGS
	DO	relu	(50,150,100)	Constant	L-BFGS

5.4.4 Evaluation and Comparisons of Results

The entire dataset consists of a ground truth dataset, and the pixel value dataset was split randomly into a 70% training set and a 30% test set to develop XGboost and MLP, regression models. R^2, *RMSE* and *adjusted R* for predicted EC, pH, Temp, SiO_2 and DO were calculated as model evaluation for each cluster separately (Table 5.12 & Table 5.13).

Table 5.12 Regression statistics of XGBoost regressor along different cluster

Clusters	WQPs	Train R^2	Test R^2	RMSE	RRMSE
C1	EC	0.32	0.27	1.23	0.005
	pH	0.94	0.78	0.08	0.010
	Temp	0.88	0.73	0.15	0.007
	SiO_2	0.98	0.98	0.01	0.001
	DO	0.98	0.97	0.01	0.001
C2	EC	0.35	0.33	2.57	0.011

	pH	0.74	0.74	0.19	0.025
	Temp	0.87	0.89	0.10	0.004
	SiO_2	0.96	0.96	0.01	0.001
	DO	0.98	0.927	0.01	0.001
C3	EC	0.23	0.21	2.85	0.013
	pH	0.74	0.74	0.26	0.033
	Temp	0.85	0.89	0.08	0.004
	SiO_2	0.97	0.97	0.00	0.000
	DO	0.97	0.9	0.01	0.001
C4	EC	0.34	0.32	2.58	0.011
	pH	0.81	0.76	0.09	0.012
	Temp	0.87	0.9	0.01	0.000
	SiO_2	0.98	0.97	0.00	0.001
	DO	0.97	0.96	0.01	0.001

Table 5.13 Regression statistics of MLP regressor along different clusters

Clusters	**WQPs**	R^2	**RMSE**	**RRMSE**
C1	EC	0.37	0.1786	0.00079
	pH	0.89	0.06198	0.00796
	Temp	0.82	0.0812	0.00374
	SiO_2	0.93	0.00426	0.00053
	DO	0.93	0.00595	0.00081
C2	EC	0.27	0.29654	0.00131
	pH	0.87	0.00535	0.00069
	Temp	0.93	0.00065	0.00003
	SiO_2	0.91	0.00472	0.00058
	DO	0.87	0.02724	0.00373
C3	EC	0.23	0.13671	0.00060
	pH	0.84	0.0183	0.00230
	Temp	0.95	0.0023	0.00011

	SiO_2	0.97	0.0063	0.00077
	DO	0.81	0.02233	0.00300
C4	EC	0.35	0.00001	0.00000
	pH	0.87	0.00509	0.00065
	Temp	0.92	0.00065	0.00003
	SiO_2	0.92	0.00472	0.00058
	DO	0.82	0.02724	0.00373

Except for EC, the R^2 values for all WQPs were high and close to 1, showing a significant relationship between satellite reflectance data and *insitu* observations. The feature of importance from each WQPs modelling was studied to identify the optimum band and combinations. Therefore, it is evident that the developed Landsat-8 based water quality modelling could be a highly recommended, cost-effective and time-saving methods for monitoring optically active and non-active WQPs. A high coefficient of determination are observed for WQPs like pH, Temp, SiO_2 and DO (R^2 0.74-0.98) with XGBoost and MLP with a *p-value <0.005*. However, EC performed poorly in all the clusters, with R^2 ranging from 0.23-0.37 for XGBoost. However, MLP produces comparably superior results, with R^2 values ranging from 0.81 to 0.97 for C2 (except for EC), 0.81 of DO in C3 and 0.97 of SiO_2 in C3, with the exception of C1 which has R^2 values of 0.32 and 0.27 in the training and testing phases, respectively. The performance evaluation measures and scatter diagrams in the testing phase for XGBoost, and MLP for pH, Temp, SiO2 and DO are presented in Figure 5.17, Figure 5.18, Figure 5.19 and Figure 5.20.

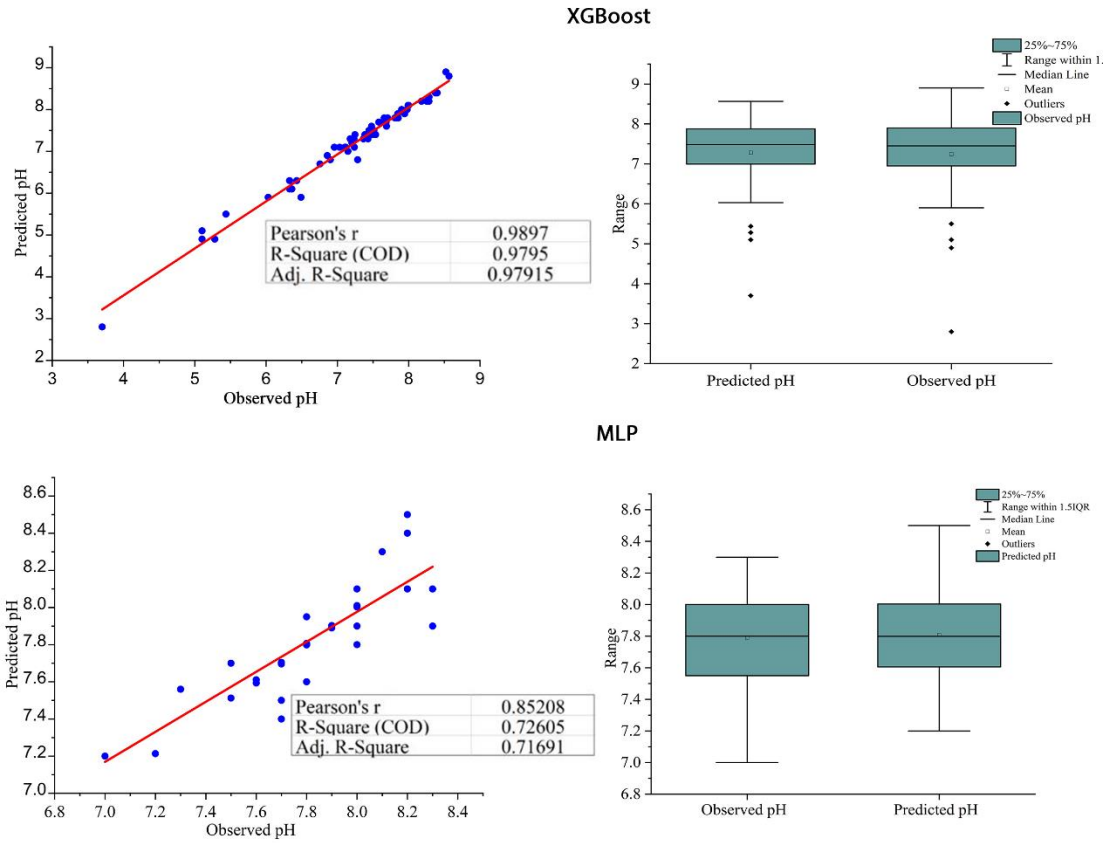


Figure 5.17 Scatter plot, Box and whisker plot for pH

Better performance for XGBoost (R^2 = 0.88- 0.98) can be observed from the displayed scatter plot for all the parameters compared to MLP (R^2 = 0.72-0.97).

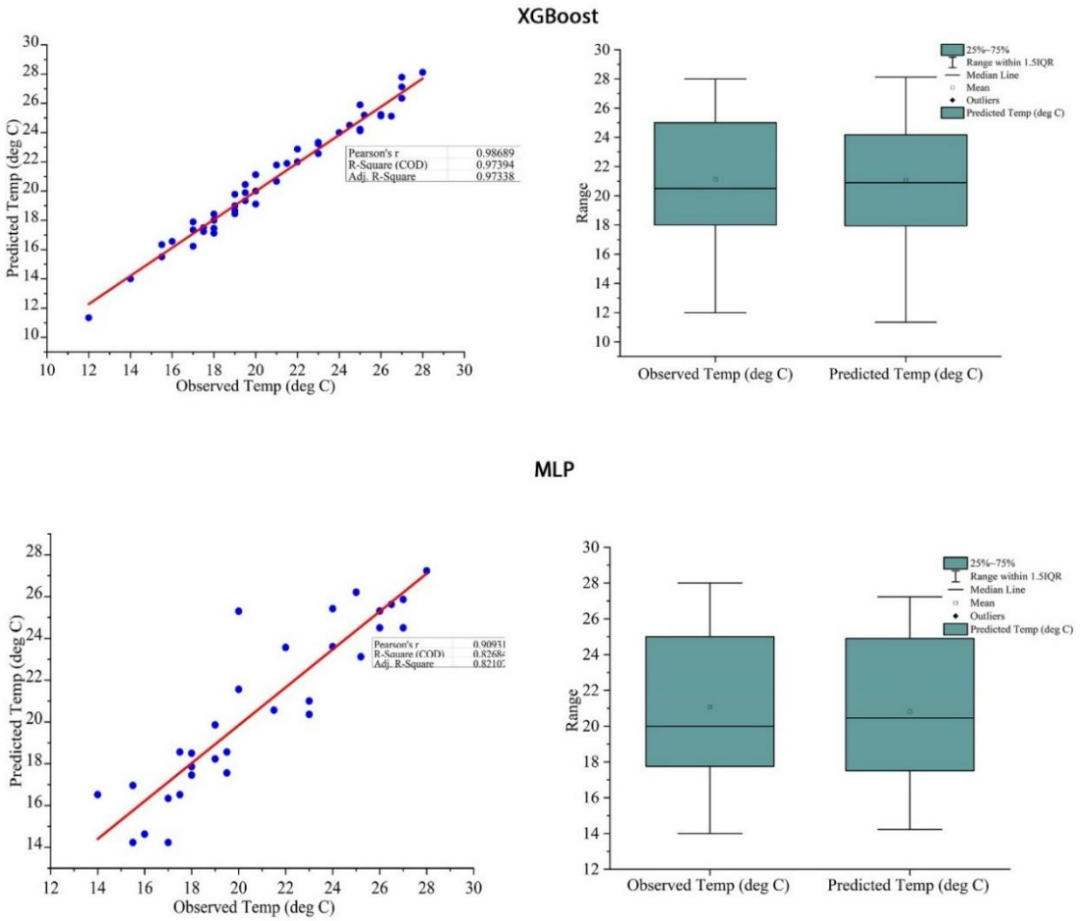


Figure 5.18 Scatter plot, Box and whisker plot for Temp

Box plots were created for all WQPs to compare the observed minimum, maximum, and mean values with the predicted. A minimal difference in mean value is observed across all the clusters for both models.

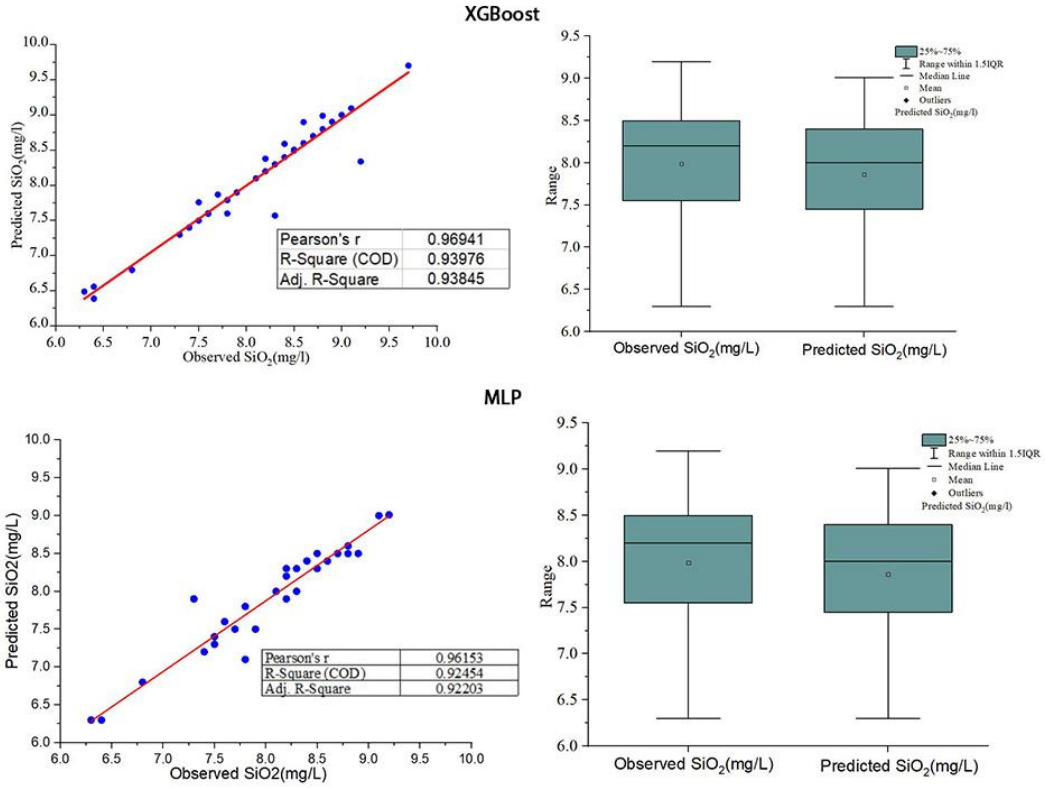


Figure 5.19 Scatter plot, Box and whisker plot for SiO2

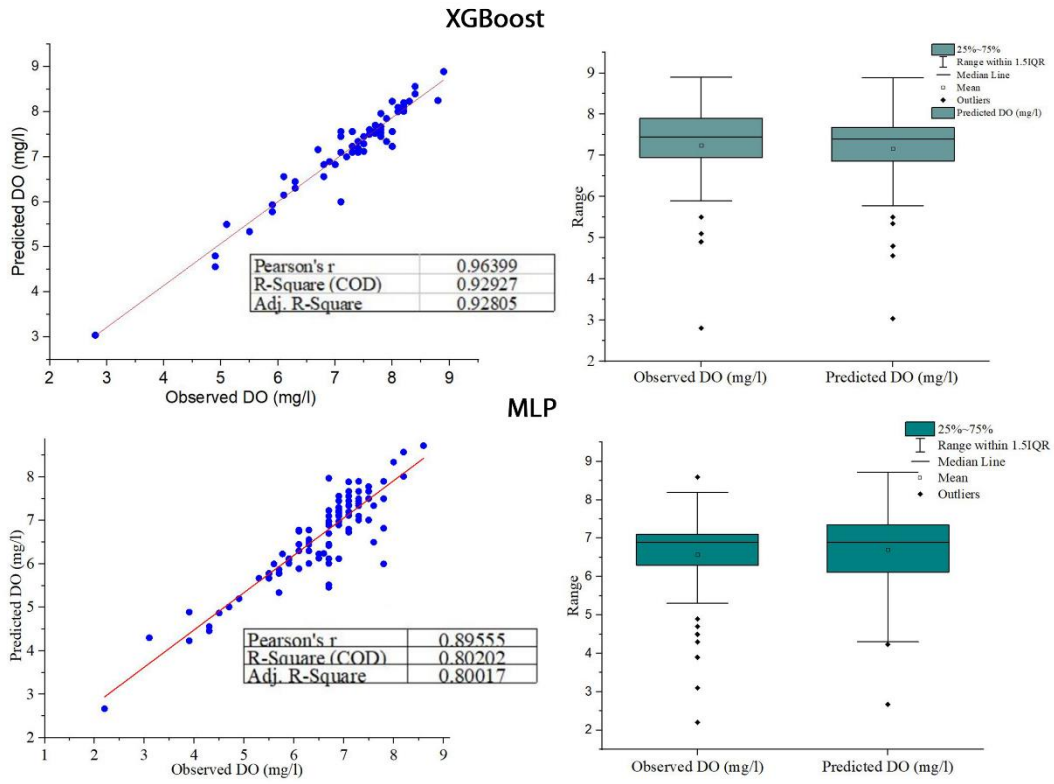


Figure 5.20 Scatter plot, Box and whisker plot for DO

5.4.5 Spatial Distribution of WQPs

From the developed models, spatiotemporal maps were plotted using the developed regression model to understand the spatial distribution of different WQPs in all the clusters. The 2017 non-monsoon and monsoon season map along C1 and C4 is presented here. The spatial pattern proves that land-use modifications and seasonal variation primarily regulated the water quality conditions. EC addresses the total concentration of water-ionized constituents; a higher conductivity concentration reflects higher water pollution. GRB shows high EC concentrations, surpassing the allowable limit of 3000 µS/cm across many parts (CWC and NRSC 2014). The typical electrical conductivity value is 300 µS/cm (Bhuyan et al. 2018). The concentration of EC along different clusters has not shown any seasonal shift for 2017 (Figure 5.21). The values are 370-630 µS/cm in non-monsoon and monsoon seasons 230-620 µS/cm. Thus, no seasonal trend is observed but showing an increase level in the monsoon

season. A lower value of EC is observed in C1 and C3 for the non-monsoon season. The permissible limits for drinking water purposes should be 1500 μmhos/cm. Except for different ranges of value, no spatial change in the patterns were observed in the non-monsoon and monsoon seasons. The high concentrations are marked in Lucknow, Allahabad Varanasi, and nearby stations fall in C1 and C4. The higher concentrations of EC could be accredited to the high degree of anthropogenic activities such as agricultural runoff and waste disposal. Inorganic compounds are better conductors than organic compounds due to the input of industrial effluents, making conductivity a good indicator of inorganic pollution. Therefore, measuring conductivity will provide a good indication of the state of inland water.

As per IS specification, the desirable limit for pH is 6.5– 8.5 mg/l. Generally, the pH concentration increases because of the photosynthetic algae activities that consume carbon dioxide dissolved in it. Broadly, pH ranges from 6.5 to 9, which is relevant for aquatic life. A slight variation in values is observed in both seasons (Figure 5.21). Keeping the aquatic ecosystem within this range is important because high and low pH can be destructive in nature (Al-Badaii et al. 2013). Downstream of the study area has shown high values of TDS for non-monsoon and monsoon seasons, 216.96-285 mg/l and 231.281 mg/l, respectively (Figure 5.21). The desired limits of TDS as per IS:2296 are 500, 1500 and 2100 mg/l for classes A, C and E, respectively. In a few stations, along with clusters C1 and C2 the predicted values for 2017 in non-monsoon and monsoon seasons are well within limits. The high TDS concentration in the rivers could be attributed to extreme anthropogenic activities along the river course and runoff with high suspended matter.

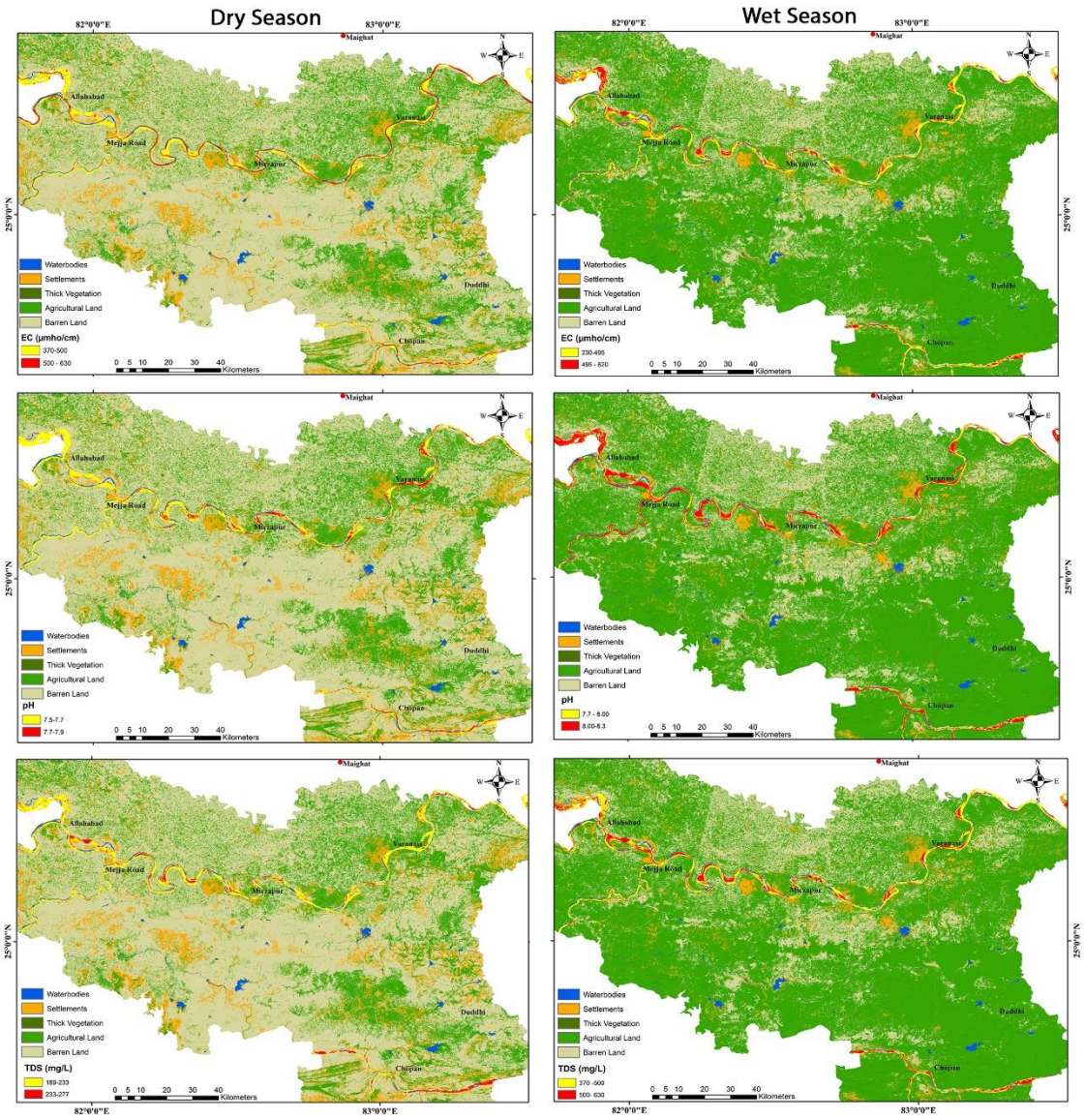


Figure 5.21 Spatial variation of EC, pH and TDS along the parts of study area

The temperature of the study area, mainly of the state of Uttar Pradesh (UP), varied from 0 to 46°C. In 2014, the average temperature during pre-monsoon and monsoon varied from 20°C–25°C upstream, 25°C–27°C in the middle stream and 27°C–30°C in downstream and the average temperature in post-monsoon varied from 13°C–15°C in upstream, 15°C–18°C in the middle stream and 18°C–20°C in downstream (Khan et al. 2017). Water temperature is directly linked with toxic absorption, salinity, and DO; temperature also influences the rate of photosynthesis by algae and aquatic plants. Increasing temperature reduces the DO, bringing harmful effects to aquatic life. The higher level of water temperature could be attributed to human activities such as the

discharge of industrial effluents, agriculture, and forest harvesting. Nevertheless, the favourable loading of temperature is associated with seasonal variation. The inverse relationship between temperature and DO is a natural process in water because warm water quickly becomes saturated with oxygen and thus holds less DO (Bhat et al. 2014). The desired value for DO should be more than 4 mg/l. The observed value for DO in the basin lies in the range of 4.3–9.2 mg/l. The DO concentrations above 5 mg/L in most locations were ideal for bathing purposes. The increasing value of DO may also be attributed to reducing wastes from various NPS. However, the decrease could be linked to an increase in water temperature and the biological activity of aquatic organisms along the river basin. Most of the stations in C2 and C3 show a clear seasonal shift, with the minimum value being 5.50-5.8 and 5.70-5.93 mg/l in non-monsoon and monsoon seasons, respectively. The stations Kanpur, Lucknow, and Ankinghat followed the same pattern during non-monsoon and monsoon seasons with a range of 6.12-7.05 mg/l. Although, higher values of DO were observed along some stretches of river 7.67-8.30 and 7.30-7.80 in non-monsoon and monsoon seasons, respectively.

CHAPTER 6

CONCLUSIONS

6.1 GENERAL

In this study, a 14-year (2005-2018, monthly data), 20 water quality variables covering 20 monitoring stations (67,200 entries) from Ankhinghat to Chopan under the Central Water Commission in the Middle Ganga Basin and LULC relationship between them are studied to understand the water quality problem using various multivariate techniques. The study also tried to illustrate and prove a significant empirical relationship between Landsat-8 OLI surface reflectance of 30 m spatial and 16-day temporal resolution data with *insitu* WQPs.

6.2 CONCLUSIONS

- To examine the temporal fluctuations of river water quality, the Spearman non-parametric correlation coefficient test (Spearman *r*) is used. Temperature with the season has the greatest Spearman *r* (-0.866) with a highly significant *p-level* of (0.0000). The season exhibited a substantial correlation with the parameters EC, pH, TDS, T, Ca, HCO_3, Mg, NO_2+NO_3, SiO_2, and DO ($p < 0.05$). Cl, CO_3, F, K, Na, NH_3-N, P.Tot, SO_4, BOD and COD had a non-significant correlation with *r*-value.
- During both seasons, the K-means cluster analysis classified the monitoring stations into four groups based on similar water quality criteria (C1, C2, C3, C4). The clusters C2 and C4 showed a seasonal shift when analyzed separately. The stations such as Kanpur, Lucknow, Raebareli, Bhitaura, Balrampur, Birdghat, Turtipar, Basti, Ayodhya, and Elginbridge fall into these clusters. Information like this can reduce the number of river sampling sites while retaining as much information as possible. Moreover, breaking down 20 monitoring stations into 4 clusters further reduced the modelling complexity in this study

- Most significant WQPs from spatial and seasonal variations from an extensive data set are identified using PCA. It is a data reduction procedure and a more traditional method of increasing the speed of machine learning algorithms. A reduced number of 3 PCs are identified for 20 WQPs in 20 stations with a variance of explanation 75.84% and 80.57% in the non-monsoon and monsoon seasons. DO, EC, P-Tot, and SO_4 are the most dominating parameters with a PC score of more than 0.8 in the non-monsoon season; similarly, TDS, K, COD, Cl, Na, SiO_2 in the monsoon season. A satisfactory result is drawn in PCA, which reduces the complexity of the model from 20 stations X 20 WQPs to 20 stations X 3 PCs. As a result, the monitoring programme can be limited to the identified dominant WQPs.
- The RDA results showed that along most of the clusters, the contributors of LULC varied spatially on the catchment scale, although they remained the same at the reach and riparian scale. The seasonal comparison indicates that the monsoon season has a more significant explanation. It has also been discovered that some LULC classes and their associations with WQPs are not associated with LULC. The redundancy analysis also revealed that thick vegetation along most clusters is essential in keeping water clean, whereas agriculture and urban areas degrade water quality.
- The model's predictive power across different clusters and scales is then evaluated using the SEM between LULC classes and WQPs. Overall, the riparian scale surpasses the watershed and reach scales regarding prediction scores. Since urban and agricultural land use is concentrated in riparian areas, future research will focus on the impact of riparian land use on water quality. Furthermore, a multi-scale methodology is suggested for better land use planning in water quality management along the Middle Ganga Basin.

Given the laborious, time-consuming, and costly nature of the *insitu* monitoring network. Besides that, the analyses are restricted to a single point in space and time, which makes it difficult when these values are critical for watershed assessments and management practices. Remote sensing technology has proven to be an excellent tool for bridging the gap between accuracy and large-scale analyses.

Therefore, the next phase of the study assessed six years of satellite data (2013-2018) to characterize the trends of dominant physicochemical WQPs such as EC, TDS, SiO_2 and DO across the four clusters identified in the preceding sections. The study also tried to illustrate and prove the presence of a significant empirical relationship between Landsat-8 OLI surface reflectance of 30 m spatial and 16-day temporal resolution data with *insitu* WQPs as explained below:

- The demonstrated regression techniques, namely XGBoost and MLPRegressor, have established a substantial spatial and temporal distribution of significant WQPs in the water bodies. Thus, it can be proposed as a rapid, inexpensive, and convenient method to obtain helpful water quality information from satellite data.
- Pearson's correlation coefficient values were used to assess the strength of the relationship between model inputs and outputs. In this context, the multi-spectral bands correlated (i.e. $r \geq 0.50$) with selected WQPs were selected to develop a regression equation for water quality retrieval.
- The study results indicated that the band ratio is more effective than the single bands due to the reflectance ratio reducing and eventually eliminating the effect of the changes in illumination conditions and the sediment type. Moreover, the binary combination factor weakens particle size's impact on reflectance.
- The applied hyperparameter optimization techniques used on these models have helped achieve optimal hyperparameters. Further, it drastically improved the model performance XGBoost (R^2 = 0.88- 0.98) and MLP (R^2 = 0.81-0.97), than the model without hyperparameter optimization.
- According to IS:2296 drinking water standards, the predicted values for different WQPs are compared. Along some stretches of GRB, the high critical values of WQPs conclude a pressing need to address the river basin to alleviate the pollution issues for a sustainable riverine ecosystem.
- However, the non-availability and presence of cloud cover of Landsat-8, especially during the monsoon season, has failed the model to discuss the spatiotemporal trend in some stations present in C1. Given this, future work concerns might include combining the Landsat-8 with other relevant sensor data

or with extremely high spatial, spectral, and temporal resolution datasets. However, it will be constructive to develop generalized models for estimating different WQPs in the GRB without being entirely dependent on water sampling.

6.3 LIMITATIONS AND FUTURE PERSPECTIVES

The non-availability and presence of cloud cover of Landsat-8, especially during the monsoon season, has failed the model to discuss the spatiotemporal trend in some stations present in C1. Future work issues could focus on fusing the Landsat-8 with other suitable sensor data with high spatial, spectral, and temporal resolution datasets. However, it will be beneficial to develop generalized models for estimating different WQPs in the GRB that are not entirely dependent on water sampling. Nevertheless, including more predictors to model the relationship between WQPs and LULC classes could have given a clear understanding of essential predictors in defining spatiotemporal patterns in water quality. Furthermore, the study did not consider the influx from the upstream catchment or the Upper Ganga Region due to computational constraints and data availability, assuming it was insignificant.

REFERENCES

Abdelmalik, K. W. (2018). “Role of statistical remote sensing for Inland water quality parameters prediction.” *Egyptian Journal of Remote Sensing and Space Science*, 21(2), 193–200.

Abdul-Aziz, O. I., and Al-Amin, S. (2016). “Climate, land use and hydrologic sensitivities of stormwater quantity and quality in a complex coastal-urban watershed.” *Urban Water Journal*, 13(3), 302–320.

Abdulkareem, J. H., Sulaiman, W. N. A., Pradhan, B., and Jamil, N. R. (2018). “Long-Term Hydrologic Impact Assessment of Non-point Source Pollution Measured Through Land Use/Land Cover (LULC) Changes in a Tropical Complex Catchment.” *Earth Systems and Environment*, 2, 67–84.

Afed Ullah, K., Jiang, J., and Wang, P. (2018). “Land use impacts on surface water quality by statistical approaches.” *Global Journal of Environmental Science and Management*, 4(2), 231–250.

Ahmad, W., Iqbal, J., Nasir, M. J., Ahmad, B., Khan, M. T., Khan, S. N., and Adnan, S. (2021). “Impact of land use/land cover changes on water quality and human health in district Peshawar Pakistan.” *Scientific Reports*, 11(1), 1–14.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., and García-Nieto, J. (2019). “Efficient water quality prediction using supervised machine learning.” *Water (Switzerland)*, 11(11), 1–14.

Al-Badaii, F., Shuhaimi-Othman, M., and Gasim, M. B. (2013). “Water quality assessment of the Semenyih river, Selangor, Malaysia.” *Journal of Chemistry*.

Allee, R. J., and Johnson, J. E. (1999). “Use of satellite imagery to estimate surface chlorophyll a and Secchi disc depth of Bull Shoals Reservoir, Arkansas, USA.” *International Journal of Remote Sensing*, 20(6), 1057–1072.

Álvarez-cabria, M., Barquín, J., and Peñas, F. J. (2016). “Modelling the spatial and seasonal variability of water quality for entire river networks : Relationships with natural and anthropogenic factors.” *Science of the Total Environment*, 545–546, 152–

162.

Amato, F., Tonini, M., Murgante, B., and Kanevski, M. (2018). “Fuzzy definition of Rural Urban Interface: An application based on land use change scenarios in Portugal.” *Environmental Modelling and Software*, 104, 171–187.

Andrzej Urbanski, J., Wochna, A., Bubak, I., Grzybowski, W., Lukawska-Matuszewska, K., Łącka, M., Śliwińska, S., Wojtasiewicz, B., and Zajączkowski, M. (2016). “Application of Landsat 8 imagery to regional-scale assessment of lake water quality.” *International Journal of Applied Earth Observation and Geoinformation*, 51, 28–36.

Antonini, K., Langer, M., Farid, A., and Walter, U. (2017). “SWEET CubeSat – Water detection and water quality monitoring for the 21st century.” *Acta Astronautica*, 140, 10–17.

Antonopoulos, V. Z., Papamichail, D. M., and Mitsiou, K. A. (2001). “Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece.” *Hydrology and Earth System Sciences*, 5(4), 679–692.

Arora, M., Casas-mulet, R., Costelloe, J. F., Peterson, T. J., Mccluskey, A. H., and Stewardson, M. J. (2017). *Chapter 6. Impacts of Hydrological Alterations on Water Quality*. *Water for the Environment*, Elsevier Inc.

Ay, M., and Kisi, O. (2014). “Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques.” *Journal of Hydrology*, 511, 279–289.

Azhar, S. C., Aris, A. Z., Yusoff, M. K., Ramli, M. F., and Juahir, H. (2015). “Classification of River Water Quality Using Multivariate Analysis.” *Procedia Environmental Sciences*, 30, 79–84.

Baban, S. M. J. (1993). “Detecting water quality parameters in the norfolk broads, U.K., using landsat imagery.” *International Journal of Remote Sensing*, 14(7), 1247–1267.

Barrett, D. C., and Amy E. Frazier. (2016). “Automated Method for Monitoring Water Quality Using Landsat Imagery.” *Water*, 8, 257.

Batur, E., and Maktav, D. (2019). “Assessment of Surface Water Quality by Using Satellite Images Fusion Based on PCA Method in the Lake Gala, Turkey.” *IEEE Transactions on Geoscience and Remote Sensing*, 57(5), 2983–2989.

Bertels, L., Vanderstraete, T., Coillie, S. Van, Knaeps, E., Sterckx, S., Goossens, R., and Deronde, B. (2008). “Mapping of coral reefs using hyperspectral CASI data; a case study: Fordata, Tanimbar, Indonesia.” *International Journal of Remote Sensing*, 29(8), 2359–2391.

Bhat, S. A., Meraj, G., Yaseen, S., and Pandit, A. K. (2014). “Statistical Assessment of Water Quality Parameters for Pollution Source Identification in Sukhnag Stream: An Inflow Stream of Lake Wular (Ramsar Site), Kashmir Himalaya.” *Journal of Ecosystems*, 2014, 1–18.

Bhuyan, M. S., Bakar, M. A., Sharif, A. S. M., Hasan, M., and Islam, M. S. (2018). “Water Quality Assessment Using Water Quality Indicators and Multivariate Analyses of the Old Brahmaputra River.” *Pollution*, 4(3), 481–493.

Bonansea, M., Ledesma, M., Rodriguez, C., and Pinotti, L. (2018). “Using new remote sensing satellites for assessing water quality in a reservoir.” *Hydrological Sciences Journal*, 64(1), 34–44.

Bonansea, M., Rodriguez, M. C., Pinotti, L., and Ferrero, S. (2015). “Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina).” *Remote Sensing of Environment*, 158, 28–41.

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R*. *Numerical Ecology with R*.

Brando, V. E., and Dekker, A. G. (2003). “Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality.” *IEEE Transactions on Geoscience and Remote Sensing*, 41(6), 1378–1387.

Breiman, L. (2001). “Random forests.” *Machine Learning*, 45(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification*

And Regression Trees. Classification and Regression Trees, Routledge.

Carstens, D., and Amer, R. (2019). "Spatio-temporal analysis of urban changes and surface water quality." *Journal of Hydrology*, 569(August 2018), 720–734.

Chan, T. K. H., Cheung, C. M. K., and Lee, Z. W. Y. (2017). "The state of online impulse-buying research: A literature analysis." *Information and Management*, 54(2), 204–217.

Chang, N. Bin, Imen, S., and Vannah, B. (2015a). "Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: A 40-year perspective." *Critical Reviews in Environmental Science and Technology*, 45(2), 101–166.

Chang, N. Bin, Imen, S., and Vannah, B. (2015b). "Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: A 40-year perspective." *Critical Reviews in Environmental Science and Technology*, 45(2), 101–166.

Chang, N. Bin, Wimberly, B., and Xuan, Z. (2012a). "Identification of spatiotemporal nutrient patterns in a coastal bay via an integrated k-means clustering and gravity model." *Journal of Environmental Monitoring*, 14(3), 992–1005.

Chang, N. Bin, Yang, Y. J., Daranpob, A., Jin, K. R., and James, T. (2012b). "Spatiotemporal pattern validation of chlorophyll-a concentrations in Lake Okeechobee, Florida, using a comparative MODIS image mining approach." *International Journal of Remote Sensing*, 33(7), 2233–2260.

Chang, N., Caselles, V., Juan, M. S., Camacho, A., Delegido, J., and Vannah, B. W. (2015c). "Integrated satellite data fusion and mining for monitoring lake water quality status of the Albufera de Valencia in Spain." *Journal of Environmental Management*, 151, 416–426.

Chen, F., Xiao, D., and Li, Z. (2016a). "Developing water quality retrieval models with in situ hyperspectral data in Poyang Lake, China." *Geo-Spatial Information Science*, 19(4), 255–266.

Chen, Q., Mei, K., Dahlgren, R. A., Wang, T., Gong, J., and Zhang, M. (2016b). "Impacts of land use and population density on seasonal surface water quality using a modified geographically weighted regression." *Science of the Total Environment*, 572, 450–466.

Cheng, X., Chen, L., Sun, R., and Kong, P. (2018). "Land use changes and socio-economic development strongly deteriorate river ecosystem health in one of the largest basins in China." *Science of the Total Environment*, 616–617, 376–385.

Consortium of 7 IITs. (2012). *Riparian Floral Diversity of Ganga River GRBMP : Ganga River Basin Management Plan*.

Consortium of 7 IITs. (2013a). *Ganga River Basin environment management plan: interim report*.

Consortium of 7 IITs. (2013b). *Demographic and Analysis in Middle Ganga Basin*.

Consortium of 7 IITs. (2013c). *Status of Urbanization and Industrialization in Middle Ganga Basin*.

Consortium of 7 IITs. (2014a). *Surface and Groundwater Modelling of the Ganga River Basin GRBMP : Ganga River Basin*.

Consortium of 7 IITs. (2014b). "Assessment of Domestic Pollution Load from Urban Agglomeration."

CWC and NRSC. (2014). *Ganga Basin Report*.

Dalu, T., Dube, T., Froneman, P. W., Sachikonye, M. T. B., Clegg, B. W., and Nhiwatiwa, T. (2015). "An assessment of chlorophyll-a concentration spatio-temporal variation using Landsat satellite data, in a small tropical reservoir." *Geocarto International*, 30(10), 1130–1143.

Dash, S., Borah, S. S., and Kalamdhad, A. (2018). "Monitoring and assessment of deepor beel water quality using multivariate statistical tools." *Water Practice and Technology*, 13(4), 893–908.

Dimri, D., Daverey, A., Kumar, A., and Sharma, A. (2021). "Monitoring water quality

of River Ganga using multivariate techniques and WQI (Water Quality Index) in Western Himalayan region of Uttarakhand, India." *Environmental Nanotechnology, Monitoring and Management*, 15, 100375.

Ding, J., Jiang, Y., Liu, Q., Hou, Z., Liao, J., Fu, L., and Peng, Q. (2016). "Influences of the land use pattern on water quality in low-order streams of the Dongjiang River basin, China : A multi-scale analysis." *Science of the Total Environment*, 551–552(19), 205–216.

Dutta, V., Dubey, D., and Kumar, S. (2020). "Cleaning the River Ganga : Impact of lockdown on water quality and future implications on river rejuvenation strategies." *Science of the Total Environment*, 743, 140756.

El-Magd, I. A., and El-Zeiny, A. (2014). "Quantitative hyperspectral analysis for characterization of the coastal water from Damietta to Port Said, Egypt." *Egyptian Journal of Remote Sensing and Space Science*, 17(1), 61–76.

El-Zeiny, A., and El-Kafrawy, S. (2017). "Assessment of water pollution induced by human activities in Burullus Lake using Landsat 8 operational land imager and GIS." *Egyptian Journal of Remote Sensing and Space Science*, 20, S49–S56.

Espinoza-Villar, R., Martinez, J. M., Armijos, E., Espinoza, J. C., Filizola, N., Santos, A. Dos, Willems, B., Fraizy, P., Santini, W., and Vauchel, P. (2018). "Spatio-temporal monitoring of suspended sediments in the Solimões River (2000–2014)." *Comptes Rendus - Geoscience*, 350(1–2), 4–12.

Fu, L., and Gan Wang, Y. (2012). "Statistical Tools for Analyzing Water Quality Data." *Water Quality Monitoring and Assessment*.

Garg, V., Aggarwal, S. P., and Chauhan, P. (2020). "Changes in turbidity along Ganga River using Sentinel-2 satellite data during lockdown associated with COVID-19." *Geomatics, Natural Hazards and Risk*, 11(1), 1175–1195.

Garg, V., Senthil Kumar, A., Aggarwal, S. P., Kumar, V., Dhote, P. R., Thakur, P. K., Nikam, B. R., Sambare, R. S., Siddiqui, A., Muduli, P. R., and Rastogi, G. (2017). "Spectral similarity approach for mapping turbidity of an inland waterbody." *Journal*

of Hydrology, 550, 527–537.

Gholizadeh, M. H., Melesse, A. M., and Reddi, L. (2016). "A comprehensive review on water quality parameters estimation using remote sensing techniques." *Sensors (Switzerland)*, 16(8).

Giri, S., and Qiu, Z. (2016). "Understanding the relationship of land uses and water quality in Twenty First Century : A review." *Journal of Environmental Management*, 173, 41–48.

Glasgow, H. B., Burkholder, J. A. M., Reed, R. E., Lewitus, A. J., and Kleinman, J. E. (2004). "Real-time remote monitoring of water quality: A review of current applications, and advancements in sensor, telemetry, and computing technologies." *Journal of Experimental Marine Biology and Ecology*, 300(1–2), 409–448.

González-Márquez, L. C., Torres-Bejarano, F. M., Torregroza-Espinosa, A. C., Hansen-Rodríguez, I. R., and Rodríguez-Gallegos, H. B. (2018). "Use of LANDSAT 8 images for depth and water quality assessment of El Guájaro reservoir, Colombia." *Journal of South American Earth Sciences*, 82, 231–238.

González, S., García, S., Ser, J. Del, Rokach, L., and Herrera, F. (2020). "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities." *Information Fusion*, 64, 205–237.

Gu, Q., Hu, H., Ma, L., Sheng, L., Yang, S., Zhang, X., Zhang, M., Zheng, K., and Chen, L. (2019). "Characterizing the spatial variations of the relationship between land use and surface water quality using self-organizing map approach." *Ecological Indicators*, 102(March), 633–643.

Günen, M. A. (2022). "Performance comparison of deep learning and machine learning methods in determining wetland water areas using EuroSAT dataset." *Environmental Science and Pollution Research*, 29(14), 21092–21106.

Günen, M. A., Atasever, U. H., and Beşdok, E. (2020). "Analyzing the contribution of training algorithms on deep neural networks for hyperspectral image classification."

Photogrammetric Engineering and Remote Sensing, 86(9), 581–588.

Hafeez, S., Wong, M., Ho, H., Nazeer, M., Nichol, J., Abbas, S., Tang, D., Lee, K., and Pun, L. (2019). "Comparison of Machine Learning Algorithms for Retrieval of Water Quality Indicators in Case-II Waters: A Case Study of Hong Kong." *Remote Sensing*, 11(6), 617.

Haji Gholizadeh, M., Melesse, A. M., and Reddi, L. (2016). "Spaceborne and airborne sensors in water quality assessment." *International Journal of Remote Sensing*, 37(14), 3143–3180.

Hajigholizadeh, M., and Melesse, A. M. (2017a). "Study on Spatiotemporal Variability of Water Quality Parameters in Florida Bay Using Remote Sensing." *Journal of Remote Sensing & GIS*, 6.

Hajigholizadeh, M., and Melesse, A. M. (2017b). "Assortment and spatiotemporal analysis of surface water quality using cluster and discriminant analyses." *Catena*, 151, 247–258.

Hajigholizadeh, M., Moncada, A., Kent, S., and Melesse, A. M. (2021). "Land – Lake Linkage and Remote Sensing Application in Water." *Land*, 10, 147.

Handcock, R. N., Torgersen, C. E., Cherkauer, K. A., Gillespie, A. R., Tockner, K., Faux, R. N., and Tan, J. (2012). "Thermal Infrared Remote Sensing of Water Temperature in Riverine Landscapes." *Fluvial Remote Sensing for Science and Management*, 85–113.

Hsu, C., Chang, C., and Lin, C.-J. (2016). *A Practical Guide to Support Vector Classification*.

Ibrahem, A., Osman, A., Najah, A., Fai, M., Feng, Y., and El-shafie, A. (2021). "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia." *Ain Shams Engineering Journal*, 12(2), 1545–1556.

Imen, S., Chang, N. Bin, and Yang, Y. J. (2015). "Developing the remote sensing-based early warning system for monitoring TSS concentrations in Lake Mead." *Journal of Environmental Management*, 160, 73–89.

Iqbal, K., Ahmad, S., and Dutta, V. (2019). “Pollution mapping in the urban segment of a tropical river: is water quality index (WQI) enough for a nutrient-polluted river?” *Applied Water Science*, 9(8), 1–16.

Julian, J. P., Davies-Colley, R. J., Gallegos, C. L., and Tran, T. V. (2013). “Optical Water Quality of Inland Waters: A Landscape Perspective.” *Annals of the Association of American Geographers*, 103(2), 309–318.

Kamble, S. R., and Vijay, R. (2011). “Assessment of water quality using cluster analysis in coastal region of Mumbai, India.” *Environmental Monitoring and Assessment*, 178(1–4), 321–332.

Kay, J. E., Kampf, S. K., Handcock, R. N., Cherkauer, K. A., Gillespie, A. R., and Burges, S. J. (2005). “Accuracy of lake and stream temperatures estimated from thermal infrared images.” *Journal of the American Water Resources Association*, 41(5), 1161–1175.

Keiner, L. E., and Yan, X. H. (1998). “A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery.” *Remote Sensing of Environment*, 66(2), 153–165.

Keith, D. J., Schaeffer, B. A., Lunetta, R. S., Gould, R. W., Rocha, K., and Cobb, D. J. (2014). “Remote sensing of selected water-quality indicators with the hyperspectral imager for the coastal ocean (HICO) sensor.” *International Journal of Remote Sensing*, 35(9), 2927–2962.

Khan, M. Y. A., Gani, K. M., and Chakrapani, G. J. (2017). “Spatial and temporal variations of physicochemical and heavy metal pollution in Ramganga River—a tributary of River Ganges, India.” *Environmental Earth Sciences*, 76(5).

Kiangala, S. K., and Wang, Z. (2021). “An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment.” *Machine Learning with Applications*, 4, 100024.

Kondratyev, K. Y., Pozdnyakov, D. V., and Pettersson, L. H. (1998). “Water quality

remote sensing in the visible spectrum." *International Journal of Remote Sensing*, 19(5), 957–979.

Kong, J., Sun, X., Wang, W., Du, D., Chen, Y., and Yang, J. (2015). "An optimal model for estimating suspended sediment concentration from Landsat TM images in the Caofeidian coastal waters." *International Journal of Remote Sensing*, 36(19–20), 5257–5272.

Koponen, S., Pulliainen, J., Kallio, K., and Hallikainen, M. (2002). "Lake water quality classification with airborne hyperspectral spectrometer and simulated MERIS data." *Remote Sensing of Environment*, 79, 51–59.

Koskinen, J., Leinonen, U., Vollrath, A., Ortmann, A., Lindquist, E., D'Annunzio, R., Pekkarinen, A., and Käyhkö, N. (2019). "Participatory mapping of forest plantations with Open Foris and Google Earth Engine." *ISPRS Journal of Photogrammetry and Remote Sensing*, 148, 63–74.

Kotekani, S. S., and Ilango, V. (2022). "HEMClust : An Improved Fraud Detection Model for Health Insurance using Heterogeneous Ensemble and K-prototype Clustering." *International Journal of Advanced Computer Science and Applications*, 13(3), 127–139.

Kulithalai Shiyam Sundar, P., and Deka, P. C. (2021). "Spatio-temporal classification and prediction of land use and land cover change for the Vembanad Lake system, Kerala: a machine learning approach." *Environmental Science and Pollution Research*, 29, 86220–86236.

Kulluk, S., Gülmez, B., Oztürk, G., and Ozer, S. (2023). "FC-Kmeans : Fixed-centered K-means algorithm." *Expert Systems With Applications*, 211, 118656.

Kumar Shukla, A., Shekhar Prasad Ojha, C., Mijic, A., Buytaert, W., Pathak, S., Dev Garg, R., and Shukla, S. (2018). "Population growth, land use and land cover transformations, and water quality nexus in the Upper Ganga River basin." *Hydrology and Earth System Sciences*, 22(9), 4745–4770.

Lamaro, A. A., Mariñelarena, A., Torrusio, S. E., and Sala, S. E. (2013). "Water

surface temperature estimation from Landsat 7 ETM+ thermal infrared data using the generalized single-channel method: Case study of Embalse del Río Tercero (Córdoba, Argentina).” *Advances in Space Research*, 51(3), 492–500.

Lepistö, A., Huttula, T., Koponen, S., Kallio, K., Lindfors, A., Tarvainen, M., and Sarvala, J. (2010). “Monitoring of spatial water quality in lakes by remote sensing and transect measurements.” *Aquatic Ecosystem Health and Management*, 13(2), 176–184.

Li, D., and Liu, S. (2018). *Water Quality Evaluation*. *Water Quality Monitoring and Management*.

Li, L., Zhang, B., and Li, J. (2016). “Statistically modelling and mining remotely sensed data in urban areas based on topic models - A conceptual analysis.” *Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing*, 0, 1–5.

Li, Y., He, L., Peng, B., Fan, K., and Tong, L. (2018). “Remote sensing inversion of water quality parameters in longquan lake based on PSO-SVR algorithm.” *International Geoscience and Remote Sensing Symposium (IGARSS)*, 9268–9271.

Liu, D., Chen, N., Zhang, X., Wang, C., and Du, W. (2020). “Annual large-scale urban land mapping based on Landsat time series in Google Earth Engine and OpenStreetMap data: A case study in the middle Yangtze River basin.” *ISPRS Journal of Photogrammetry and Remote Sensing*, 159(November 2019), 337–351.

Liu, J., Zhang, Y., Yuan, D., and Song, X. (2015). “Empirical Estimation of Total Nitrogen and Total Phosphorus Concentration of Urban Water Bodies in China Using High Resolution IKONOS Multispectral Imagery.” *Water*, 7(12), 6551–6573.

Liu, W., Wang, S., Yang, R., Ma, Y., Shen, M., You, Y., Hai, K., and Baqa, M. F. (2019). *Remote sensing retrieval of turbidity in alpine rivers based on high spatial resolution satellites*. *Remote Sensing*.

Lounis, B., Aissa, A. B., Rabia, S., and Ramoul, A. (2013). “Hybridisation of fuzzy systems and genetic algorithms for water quality characterisation using remote sensing data.” *International Journal of Image and Data Fusion*, 4(2), 171–196.

Maeda, E. E., Lisboa, F., Kaikkonen, L., Kallio, K., Koponen, S., Brotas, V., and Kuikka, S. (2019). “Temporal patterns of phytoplankton phenology across high latitude lakes unveiled by long-term time series of satellite data.” *Remote Sensing of Environment*, 221(January), 609–620.

Maillard, P., and Pinheiro Santos, N. A. (2008). “A spatial-statistical approach for modeling the effect of non-point source pollution on different water quality parameters in the Velhas river watershed - Brazil.” *Journal of Environmental Management*, 86(1), 158–170.

Mainali, J., and Chang, H. (2018). “Landscape and anthropogenic factors affecting spatial patterns of water quality trends in a large river basin, South Korea.” *Journal of Hydrology*, 564, 26–40.

Martín, J., Sáez, J. A., and Corchado, E. (2021). “On the suitability of stacking-based ensembles in smart agriculture for evapotranspiration prediction.” *Applied Soft Computing*, 108, 107509.

Matthews, M. W. (2011). “A current review of empirical procedures of remote sensing in Inland and near-coastal transitional waters.” *International Journal of Remote Sensing*, 32(21), 6855–6899.

Mello, K. De, Averna, R., Randhir, T. O., Cordeiro, A., and Alberto, C. (2018). “Effects of land use and land cover on water quality of low-order streams in Southeastern Brazil : Watershed versus riparian zone.” *Catena*, 167, 130–138.

Mello, K. de, Taniwaki, R. H., Paula, F. R. de, Valente, R. A., Randhir, T. O., Macedo, D. R., Leal, C. G., Rodrigues, C. B., and Hughes, R. M. (2020). “Multiscale land use impacts on water quality: Assessment, planning, and future perspectives in Brazil.” *Journal of Environmental Management*, 270, 110879.

Miller, J. D., and Hutchins, M. (2017). “Journal of Hydrology : Regional Studies The impacts of urbanisation and climate change on urban fl ooding and urban water quality : A review of the evidence concerning the United Kingdom.” *Journal of Hydrology: Regional Studies*, 12(June), 345–362.

Misra, R. (2015). *Uttar Pradesh. State Politics in India*.

Moradkhani, K., and Fathi, A. (2022). "Segmentation of waterbodies in remote sensing images using deep stacked ensemble model." *Applied Soft Computing*, 124, 109038.

Morel, A., and Prieur, L. (1977). "Analysis of variations in ocean color." *Limnology and Oceanography*, 22(4), 709–722.

Naganna, S. R., and Deka, P. C. (2019). "Artificial intelligence approaches for spatial modeling of streambed hydraulic conductivity." *Acta Geophysica*, 67(3), 891–903.

Namami Gange. (2020). "National Mission for Clean Ganga | NMCG." *Department of Water Resources, River Development & Ganga Rejuvenation*, <https://nmcg.nic.in/index.aspx> (Dec. 30, 2022).

Nas, B., Ekercin, S., Karabörk, H., Berktay, A., and Mulla, D. J. (2010). "An application of landsat-5TM image data for water quality mapping in Lake Beysehir, Turkey." *Water, Air, and Soil Pollution*, 212(1–4), 183–197.

Nazeer, M., and Nichol, J. E. (2015). "Combining Landsat TM / ETM + and HJ-1 A / B CCD Sensors for Monitoring Coastal Water Quality in Hong Kong." *IEEE Geoscience and Remote Sensing Letters*, 12(9), 1898–1902.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). "The application of data mining techniques in fi nancial fraud detection : A classi fi cation framework and an academic review of literature." *Decision Support Systems*, 50(3), 559–569.

Novoa, S., Chust, G., Valencia, V., Froidefond, J. M., and Morichon, D. (2011). "Estimation of chlorophyll-a concentration in waters over the continental shelf of the bay of biscay: A comparison of remote sensing algorithms." *International Journal of Remote Sensing*, 32(23), 8349–8371.

Olmanson, L. G., Brezonik, P. L., and Bauer, M. E. (2013). "Remote Sensing of Environment Airborne hyperspectral remote sensing to assess spatial distribution of water quality characteristics in large rivers : The Mississippi River and its tributaries in Minnesota." *Remote Sensing of Environment*, 130, 254–265.

Panda, S. S., Garg, V., and Chaubey, I. (2004). “Artificial Neural Networks Application in Lake Water Quality Estimation Using Satellite Imagery.” *Journal of Environmental Informatics*, 4(2), 65–74.

Pathak, D., Whitehead, P. G., Futter, M. N., and Sinha, R. (2018). “Water quality assessment and catchment-scale nutrient flux modeling in the Ramganga River Basin in north India: An application of INCA model.” *Science of the Total Environment*, 631–632, 201–215.

Patra, S., Sahoo, S., Mishra, P., and Chandra, S. (2018). “Impacts of urbanization on land use / cover changes and its probable implications on local climate and groundwater level.” *Journal of Urban Management*, (April), 1–15.

Peneva, E., Griffith, J. A., and Carter, G. A. (2008). “Seagrass Mapping in the Northern Gulf of Mexico using Airborne Hyperspectral Imagery: A Comparison of Classification Methods.” *Journal of Coastal Research*, 244(244), 850–856.

Peterson, K. T., Sagan, V., and Sloan, J. J. (2020). “Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing.” *GIScience and Remote Sensing*, 57(4), 510–525.

Phinn, S., Roelfsema, C., Dekker, A., Brando, V., and Anstee, J. (2008). “Mapping seagrass species, cover and biomass in shallow waters: An assessment of satellite multi-spectral and airborne hyper-spectral imaging systems in Moreton Bay (Australia).” *Remote Sensing of Environment*, 112(8), 3413–3425.

Pourhabibi, T., Ong, K., Kam, B. H., and Ling, Y. (2020). “Fraud detection : A systematic literature review of graph-based anomaly detection approaches.” *Decision Support Systems*, 133(April), 113303.

Ramsey, E. W., Jensen, J. R., Mackey, H., and Gladden, J. (1992). “Remote sensing of water quality in active to inactive cooling water reservoirs.” *International Journal of Remote Sensing*, 13(18), 3465–3488.

Razmkhah, H., Abrishamchi, A., and Torkian, A. (2010). “Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: A case study on

Jajrood River (Tehran, Iran).” *Journal of Environmental Management*, 91(4), 852–860.

Ritchie, J. C., Zimba, P. V, and Everitt, J. H. (2003). “Remote Sensing Techniques to Assess Water Quality.” *Photogrammetric Engineering and Remote Sensing*, 69(February), 695–704.

Rostom, N. G., Shalaby, A. A., Issa, Y. M., and Afifi, A. A. (2017). “Evaluation of Mariut Lake water quality using Hyperspectral Remote Sensing and laboratory works.” *Egyptian Journal of Remote Sensing and Space Science*, 20, S39–S48.

Rubin, H. J., Lutz, D. A., Steele, B. G., Cottingham, K. L., Weathers, K. C., Ducey, M. J., Palace, M., Johnson, K. M., and Chipman, J. W. (2021). “Remote Sensing of Lake Water Clarity : Performance and Transferability of Both Historical Algorithms and Machine Learning.” *Remote Sensing*, 13, 1434.

Saadi, A. M. El, Yousry, M. M., and Jahin, H. S. (2014). “Statistical estimation of Rosetta branch water quality using multi-spectral data.” *Water Science*, 28(1), 18–30.

Said, S., and Khan, S. A. (2021). “Remote sensing-based water quality index estimation using data-driven approaches: a case study of the Kali River in Uttar Pradesh, India.” *Environment, Development and Sustainability*, 23(12), 18252–18277.

Sandoval, S., Torres, A., Duarte, M., and Velasco, A. (2014). “Assessment of rainfall influence over water quality effluent of an urban catchment: A data driven approach.” *Urban Water Journal*, Taylor & Francis.

Santos, A. L. M. R. dos, Martinez, J. M., Filizola, N. P., Armijos, E., and Alves, L. G. S. (2018). “Purus River suspended sediment variability and contributions to the Amazon River from satellite data (2000–2015).” *Comptes Rendus - Geoscience*, 350(1–2), 13–19.

Shamitha, S. K., and Ilango, V. (2019). “A roadmap for intelligent data analysis using clustering algorithms and implementation on health insurance data.” *International Journal of Scientific and Technology Research*, 8(10), 2008–2018.

Shapiro, A. S. S., and Wilk, M. B. (1965). “An Analysis of Variance Test for

Normality (Complete Samples) Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : https://www.jstor.org/stable/2333709 An analysis of variance test for normality (complete samples) t.” *Oxford University Press*, 52(3), 591–611.

Sharaf El Din, E., Zhang, Y., and Suliman, A. (2017a). “Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework.” *International Journal of Remote Sensing*, 38(4), 1023–1042.

Sharaf El Din, E., Zhang, Y., and Suliman, A. (2017b). “Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework.” *International Journal of Remote Sensing*, 38(4), 1023–1042.

Sharma, B. (2018). “Appraisal of river water quality using open-access earth observation data set : Appraisal of river water quality using open-access earth observation data set : a study of river Ganga at Allahabad (India).” *Sustainable Water Resources Management*, 5, 755–765.

Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., and Skakun, S. (2017). “Exploring Google earth engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping.” *Frontiers in Earth Science*, 5(February), 1–10.

Shi, P., Zhang, Y., Li, Z., Li, P., and Xu, G. (2017). “Influence of land use and land cover patterns on seasonal water quality at multi-spatial scales.” *Catena*, 151, 182–190.

Shukla, A. K., Shekhar, C., Ojha, P., Mijic, A., Buytaert, W., Pathak, S., and Dev, R. (2017). “Population Growth – Land Use / Land Cover Transformations-Water Quality Nexus in Upper Ganga River Basin.” *Hydrology and Earth System Sciences*, (October).

Shukla, S., Khire, M. V, and Gedam, S. S. (2014). “Effects of land use/land cover changes on water quality of a sub-tropical river basin.” *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 3188–3191.

Singh, K. P., Malik, A., Mohan, D., and Sinha, S. (2004). “Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - A case study.” *Water Research*, 38(18), 3980–3992.

Singh, S., and Mishra, A. (2014). “Spatiotemporal analysis of the effects of forest covers on stream water quality in Western Ghats of peninsular India.” *Journal of Hydrology*, 519, 214–224.

Siqueira, J. De, Luma, C., Costa, S., Lu, M., Alves, E. M., Peixoto, P., and Jose, A. (2015). “Impact of land use / land cover changes on water quality and hydrological behavior of an agricultural subwatershed.” *Environ Earth Sci*, 74, 5373–5382.

Snyder, H. (2019). “Literature review as a research methodology: An overview and guidelines.” *Journal of Business Research*, 104(August), 333–339.

Song, K., Liu, G., Wang, Q., Wen, Z., Lyu, L., Du, Y., Sha, L., and Fang, C. (2020). “Quantification of lake clarity in China using Landsat OLI imagery data.” *Remote Sensing of Environment*, 243(March), 111800.

Steissberg, T. E., Schladow, S. G., and Hook, S. J. (2010). “Monitoring Past, Present, and Future Water Quality Using Remote Sensing.” *Southern Nevada Public Lands Management Act Lake Tahoe Environmental Improvement Program*.

Sudheer, K. P., Chaubey, I., and Garg, V. (2007). “Lake water quality assessment from Landsat Thematic Mapper data using neural network : an approach to optimal band combination selection.” *Journal of the American Water Resources Association(JAWRA)*, 42(6), 1683–1695.

Sundaray, S. K., Panda, U. C., Nayak, B. B., and Bhatta, D. (2006). “Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of the Mahanadi river-estuarine system (India) - A case study.” *Environmental Geochemistry and Health*, 28(4), 317–330.

Swain, R., and Sahoo, B. (2017a). “Mapping of heavy metal pollution in river water at daily time-scale using spatio-temporal fusion of MODIS-aqua and Landsat satellite imageries.” *Journal of Environmental Management*, 192, 1–14.

Swain, R., and Sahoo, B. (2017b). “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.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., and Brisco, B. (2020). “Google Earth Engine for geo-big data applications: A meta-analysis and systematic review.” *ISPRS Journal of Photogrammetry and Remote Sensing*, 164(January), 152–170.

Tanaka, M. O., Lúcia, A., Souza, T. De, Moschini, L. E., and Oliveira, A. K. De. (2016). “Agriculture , Ecosystems and Environment In fl uence of watershed land use and riparian characteristics on biological indicators of stream water quality in southeastern Brazil.” *“Agriculture, Ecosystems and Environment,”* 216, 333–339.

Teodoro, A. C., Veloso-Gomes, F., and Gonçalves, H. (2007). “Retrieving TSM concentration from multispectral satellite data by multiple regression and artificial neural networks.” *IEEE Transactions on Geoscience and Remote Sensing*, 45(5), 1342–1350.

Tian, Z., Xiao, J., Feng, H., and Wei, Y. (2020). “Credit Risk Assessment based on Gradient Boosting Decision Tree.” *Procedia Computer Science*, 174, 150–160.

Tibebe, D., Kassa, Y., Melaku, A., and Lakew, S. (2019). “Investigation of spatio-temporal variations of selected water quality parameters and trophic status of Lake Tana for sustainable management, Ethiopia.” *Microchemical Journal*, 148(April), 374–384.

Umwali, E. D., Kurban, A., Isabwe, A., Mind’je, R., Azadi, H., Guo, Z., Udahogora, M., Nyirarwasa, A., Umuhoza, J., Nzabarinda, V., Gasirabo, A., and Sabirhazi, G. (2021). “Spatio-seasonal variation of water quality influenced by land use and land cover in Lake Muhazi.” *Scientific Reports*, 11(1), 1–16.

Vanhellemont, Q. (2020). “Automated water surface temperature retrieval from Landsat 8/TIRS.” *Remote Sensing of Environment*, 237(November 2019), 111518.

Vanhellemont, Q., and Ruddick, K. (2015). “Advantages of high quality SWIR bands

for ocean colour processing: Examples from Landsat-8." *Remote Sensing of Environment*, 161, 89–106.

Wang, L., Diao, C., Xian, G., Yin, D., Lu, Y., Zou, S., and Erickson, T. A. (2020). "A summary of the special issue on remote sensing of land change science with Google earth engine." *Remote Sensing of Environment*, 248(April 2018).

Wang, X., Cai, Q., Ye, L., and Qu, X. (2012). "Evaluation of spatial and temporal variation in stream water quality by multivariate statistical techniques: A case study of the Xiangxi River basin, China." *Quaternary International*, 282, 137–144.

Wang, X., Ma, L., and Wang, X. (2010). "Apply semi-supervised support vector regression for remote sensing water quality retrieving." *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2757–2760.

Wang, X., and Yang, W. (2019). "Water quality monitoring and evaluation using remote-sensing techniques in China: A systematic review." *Ecosystem Health and Sustainability*, 5(1), 47–56.

Wang, Y., Liu, X., Wang, T., Zhang, X., Feng, Y., Yang, G., and Zhen, W. (2021). "Relating land-use/land-cover patterns to water quality in watersheds based on the structural equation modeling." *Catena*, 206(August 2020), 105566.

Wen, X., and Yang, X. (2009). "Monitoring of Water Quality Using Remote Sensing Data Mining." *Knowledge-Oriented Applications in Data Mining*.

Woerd, H. Vander, and Pasterkamp, R. (2004). "Mapping of the North Sea turbid coastal waters using SeaWiFS data." *Canadian Journal of Remote Sensing*, 30(1), 44–53.

WRIS. (2022). "India-WRIS." <https://indiawris.gov.in/wris/#/lulc> (Dec. 27, 2022).

Wu, T., Zhang, W., Jiao, X., Guo, W., and Alhaj Hamoud, Y. (2021). "Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration." *Computers and Electronics in Agriculture*, 184(March 2020), 106039.

Wunderlin, D. A., María Del Pilar, D., María Valeria, A., Fabiana, P. S., Cecilia, H. A., and María De Los Ángeles, B. (2001). “Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A Case Study: Suquía River basin (Córdoba-Argentina).” *Water Research*, 35(12), 2881–2894.

Yepez, S., Laraque, A., Martinez, J. M., Sa, J. De, Carrera, J. M., Castellanos, B., Gallay, M., and Lopez, J. L. (2018). “Retrieval of suspended sediment concentrations using Landsat-8 OLI satellite images in the Orinoco River (Venezuela).” *Comptes Rendus - Geoscience*, 350(1–2), 20–30.

Yu, X., Yi, H., Liu, X., Wang, Y., Liu, X., and Zhang, H. (2016). “Remote-sensing estimation of dissolved inorganic nitrogen concentration in the Bohai Sea using band combinations derived from MODIS data.” *International Journal of Remote Sensing*, 37(2), 327–340.

Yuan, H. L., and Chen, H. Q. (2011). “Analysis of water quality variation of rainfall in Xi’an.” *ISWREP 2011 - Proceedings of 2011 International Symposium on Water Resource and Environmental Protection*, 1, 482–485.

Zhan, H., Shi, P., and Chen, C. (2003). “Retrieval of oceanic chlorophyll concentration using support vector machines.” *IEEE Transactions on Geoscience and Remote Sensing*, 41(12 PART II), 2947–2951.

Zhang, H., Qi, Z. fang, Ye, X. yue, Cai, Y. bin, Ma, W. chun, and Chen, M. nan. (2013). “Analysis of land use/land cover change, population shift, and their effects on spatiotemporal patterns of urban heat islands in metropolitan Shanghai, China.” *Applied Geography*, 44, 121–133.

Zhang, Y., Wu, L., Ren, H., Deng, L., and Zhang, P. (2020). “Retrieval of water quality parameters from hyperspectral images using hybrid Bayesian probabilistic neural network.” *Remote Sensing*, 12(10), 1–31.

Zhou, C., Zhang, C., Tian, D., Wang, K., Huang, M., and Liu, Y. (2017). “A software sensor model based on hybrid fuzzy neural network for rapid estimation water quality in Guangzhou section of Pearl River, China.” *Journal of Environmental Science and*

Health - Part A Toxic/Hazardous Substances and Environmental Engineering, 53(1), 91–98.

Zhou, P., Huang, J., Gilmore, R., Jr, P., and Hong, H. (2016). "New insight into the correlations between land use andwater quality in a coastal watershed of China: Does point source pollution weaken it?" *Science of the Total Environment*, 543, 591–600.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R. (2021). "Ensemble machine learning paradigms in hydrology: A review." *Journal of Hydrology*, 598(April), 126266.

PUBLICATIONS

International Journals

Krishnaraj, A., and Deka, P. C. (2020). "Spatial and temporal variations in river water quality of the Middle Ganga Basin using unsupervised machine learning techniques." Environ Monit Assess, 192:744. doi: 10.1007/s10661-020-08624-4.

Krishnaraj, A., and Honnasiddaiah, R. (2022). "Remote sensing and machine learning based framework for the assessment of spatio - temporal water quality in the Middle Ganga Basin." Environmental Science and Pollution Research, 29, 64939–64958. doi: 10.1007/s11356-022-20386-9.

International Conferences

Krishnaraj, A and Honnasiddaiah, R. "A remote sensing and LSTM based water quality prediction along Middle Ganga Basin", AGU fall meeting 2021 (Poster presentation).

Ashwitha, S. K, Deka, P. C and Parthasarathy, K. S. S. "Monitoring the changes in river water quality due to the natural and anthropogenic factors with machine learning models using Landsat imagery." 24th HYDRO 2019, International Conference, Osmania university, Hydrabad, India.

BIO-DATA

Name : Ashwitha S K

Date of Birth : 14-08-1987

Address : #B-2405, NorthernSky City
Ujjodi, Pumpwell
Mangaluru, Karnataka-575002

Mobile : +919995089795

E-mail : ashwitha.ashi@gmail.com

Qualification : M.Tech (Remote sensing & GIS)
B.Tech (Civil Engineering)
Diploma (Civil Engineering)

Publications : Journals: 2

International Conferences: 2