Integrating Remote Sensing and Temporal Data for the Predictive Modeling of Total Suspended Solids (TSS)

by

1. Ananta Das Titu (Reg.: 2019333541)
2. Md. Aminul Islam (Reg.: 2019333571)

A thesis submitted to the

Department of Civil Engineering

in partial fulfillment for the degree of

BACHELOR OF SCIENCE IN CIVIL ENGINEERING

Department of Civil Engineering

Sylhet Engineering College

Sylhet, Bangladesh

July 2025

CERTIFICATE OF APPROVAL

The thesis/project titled "Integrating Remote Sensing and Temporal Data for the Predictive Modeling of Total Suspended Solids (TSS)" submitted by Ananta Das Titu (Reg.: 2019333541) and Md. Aminul Islam (Reg.: 2019333571), has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Civil Engineering on 22nd July, 2025.

____

1. Chairman of The Exam Committee

____

2. Supervisor

____

3. Co-Supervisor

DECLARATION

I hereby declare that this thesis/project is our own work and to the best of our knowledge it contains no materials previously published or written by another person, or have been accepted for the award of any other degree or diploma at Sylhet Engineering College (SEC) or any other educational institution. We also declare that the intellectual content of this thesis is the product of our own work and any contribution made to the research by others, with whom I have worked at SEC or elsewhere, is explicitly acknowledged.

Ananta Das Titu

Md. Aminul Islam

Dedicated

to

To our parents, who never stopped believing in us, and to everyone who stood beside us during challenges—this work is for you.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to our supervisor, Md. Abdur Rashid Sarker, for his continuous guidance, encouragement, and insightful feedback throughout the course of our thesis. We are also thankful to our co-supervisor, Sibbir Ahmed, for his valuable suggestions and kind support.

We are deeply grateful to our families and friends for their unwavering encouragement and motivation. Their support, regardless of background or belief, has been a great source of strength for us.

Finally, we appreciate the help and cooperation of all who contributed to the successful completion of our thesis.

ABSTRACT

Total Suspended Solids (TSS) serves as a critical indicator of river water quality, impacting aquatic ecosystems, sediment transport, and flood risk management. Conventional TSS monitoring methods, limited by high costs, logistical challenges, and sparse spatial coverage, are insufficient for large river systems like the Rhine. This study proposes a novel, scalable framework for predicting TSS concentrations by integrating remote sensing and machine learning, focusing on two contrasting Rhine River locations in Germany: Duisburg, a downstream industrial region, and Bonn, an upstream agricultural area. Utilizing Landsat 5 and 7 satellite imagery (1994–2009) from Google Earth Engine (GEE), river discharge data from the Global Runoff Data Centre(GRDC), and TSS measurements from the Global River Water Quality Archive(GRQA), the research employs Random Forest, XGBoost, Linear Regression, Bagging Regressor, Gradient Boosting, CatBoost, and an Artificial Neural Network (ANN) to model TSS dynamics. Input features, including satellite-derived indices (NDVI, MNDWI, EVI) and discharge data, capture the influences of industrial runoff in Duisburg and agricultural runoff in Bonn, alongside geographical variations (floodplains vs. steep slopes). Test set evaluations identified hyper-tuned XGBoost as the top performer ($R^2 = 0.9871$, MSE = 0.0016), closely followed by Gradient Boosting ($R^2 = 0.9845$, MSE = 0.0019) and ANN ($R^2 = 0.9863$, MSE = 0.0017), demonstrating high predictive accuracy. Random Forest and Bagging Regressor also performed robustly ($R^2 = 0.9823$ and 0.9820, respectively), while Linear Regression lagged significantly ($R^2 = 0.9357$, MSE = 0.0080), struggling with non-linear relationships. These findings highlight the superiority of ensemble models and the transformative potential of remote sensing for real-time, spatially comprehensive TSS monitoring. The study provides actionable insights for managing industrial and agricultural impacts on river health, advocating for the integration of satellite data and machine learning into environmental policy and sediment control strategies to enhance water quality management in large river systems

TABLE OF CONTENT

CHAPTER 1
INTRODUCTION

1.1 Background and Motivations

Total Suspended Solids (TSS) are particulate materials suspended in water, ranging from sediments and organic matter to pollutants and microorganisms. TSS is considered one of the most significant indicators of water quality in rivers and lakes, as it plays a critical role in regulating aquatic ecosystem health [1]. Elevated TSS concentrations reduce water clarity, obstructing the penetration of sunlight, which is essential for aquatic plant photosynthesis. This lack of light can disrupt aquatic food webs, leading to oxygen depletion and hypoxic or anoxic conditions, harming fish and other aquatic organisms. High TSS concentrations also carry heavy metals, pesticides, and nutrients, further deteriorating water quality and causing eutrophication in the affected water bodies [2].

The importance of monitoring TSS lies not only in its effects on aquatic life but also in its role in sediment transport and flood risk management [3]. For instance, TSS impacts soil erosion and sediment deposition, particularly in river systems during floods. This sediment movement can lead to reservoir sedimentation, affecting water storage capacities and infrastructure such as dams and levees. The sediment dynamics also play a vital role in maintaining morphological features of river ecosystems, such as wetlands and river deltas, which provide crucial ecosystem services, including storm surge protection and carbon storage [4].

Given these significant impacts of TSS, it is crucial to develop methods for monitoring and predicting its concentrations with high accuracy [5]. However, traditional methods for measuring TSS, such as field sampling and laboratory analysis, face major limitations. These techniques are expensive, time-consuming, and often limited by spatial coverage—sampling can only be done at discrete locations along a river, making it challenging to assess larger-scale water quality. Furthermore, traditional methods lack real-time data, which is especially important during events like storms, floods, or periods of high discharge, when TSS concentrations can change rapidly [6].

1.2 The Role of Remote Sensing in TSS Monitoring

Given the limitations of traditional methods, there has been increasing interest in using remote sensing technologies to monitor TSS across large spatial scales. Satellite-based systems such as Landsat, Sentinel-2, and MODIS provide global, high-resolution imagery that can be used to detect and estimate TSS concentrations over vast geographical areas [7]. Remote sensing enables the collection of spatially distributed data that provides comprehensive coverage of water bodies, offering continuous monitoring without the need for costly field visits [8].

In this research, satellite imagery serves as a primary source of data, with specific bands and spectral indices used to assess TSS levels. For example, the Near Infrared (NIR) band and Red band are particularly sensitive to suspended particles, allowing for the estimation of TSS concentrations. Various spectral indices, such as the Normalized Difference Water Index (NDWI) and Modified NDWI (MNDWI), are also used to enhance the accuracy of TSS predictions by distinguishing water bodies from land and identifying areas with high turbidity [9].

Remote sensing alone, however, cannot provide the level of accuracy needed for TSS prediction. This is where machine learning (ML) comes in. By combining satellite-derived data (such as reflectance values from various bands) with discharge data (which indicates sediment transport dynamics), machine learning models can uncover complex, non-linear relationships between environmental variables and TSS concentrations.

1.3 Machine Learning techniques for TSS Prediction

In recent years, machine learning algorithms have become essential tools for predicting water quality parameters like TSS [10]. These models can process large datasets, identify patterns, and make predictions based on input features, including satellite-derived bands, vegetation indices, and river discharge data [11]. In this study, we employ several advanced machine learning algorithms to predict TSS concentrations in the Rhine River:

- Random Forest (RF): An ensemble learning method that constructs multiple decision trees and combines their predictions. Random Forest is particularly suited for handling non-linear relationships and can effectively capture complex interactions between various input variables, making it highly reliable for predicting TSS concentrations across different environmental conditions.

- XGBoost (Extreme Gradient Boosting): A powerful gradient-boosting algorithm that builds models sequentially, correcting the errors made by previous models. XGBoost is known for its ability to handle large datasets efficiently, making it an ideal choice for TSS prediction where high-dimensional data is involved. Its ability to model complex relationships makes it a strong performer in predicting TSS concentrations in the presence of varying water quality parameters.

- Linear Regression (LR): A simpler model that assumes a linear relationship between the input variables (satellite bands, discharge data) and the target variable (TSS). While it serves as a baseline model for comparison, it has limitations in handling the complex, non-linear nature of TSS data.

The use of these models allows for the comparison of performance metrics, such as R^2, MSE, RMSE, and MAE, to assess how well each model can predict TSS concentrations. In particular, the study aims to evaluate the effectiveness of XGBoost and Random Forest, which have demonstrated superior performance in previous studies of similar environmental data.

1.4 Research Objectives

This study aims to develop machine learning models that predict TSS concentrations in the Rhine River, using satellite imagery, discharge data, and machine learning techniques. The specific research objectives are:

1. Develop predictive models for TSS using satellite-derived features and discharge data.

2. Compare the impact of industrial vs. agricultural runoff and floodplains vs. steep slopes on TSS concentrations at Duisburg and Bonn.

3. Evaluate the performance of machine learning models (Random Forest, XGBoost, Linear Regression) in predicting TSS concentrations based on standard evaluation metrics (R^2, MSE, RMSE, MAE).

4. Assess the potential of remote sensing as a tool for real-time, non-invasive monitoring of water quality and TSS concentrations in large river systems.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Total Suspended Solids (TSS) is a critical water quality parameter that plays a major role in maintaining the health of aquatic ecosystems. It influences key water quality aspects such as light penetration, oxygen levels, sediment transport, and nutrient cycling. Accurate prediction and monitoring of TSS concentrations are essential for managing water bodies, especially large river systems, where varying land uses and anthropogenic activities often complicate the situation. Traditional methods of measuring TSS, such as field sampling and laboratory analysis, are time-consuming and expensive, making them less suitable for large-scale or long-term monitoring projects. Consequently, there has been increasing interest in leveraging remote sensing data in combination with machine learning techniques for efficient, cost-effective, and scalable TSS prediction. This literature review examines several studies that explore the use of machine learning and remote sensing for TSS prediction, discussing their methodologies, findings, and limitations, and identifying key gaps that still need to be addressed.

2.2 Machine learning for TSS Prediction in Specific Water Bodies

One of the earliest studies that highlighted the application of machine learning for TSS prediction was conducted by Balahaha Hadi Ziyad Sami et al. [12], who investigated the use of artificial neural networks (ANN) for predicting TSS concentrations in Taiwan's Fei Tsui Reservoir. This study is significant because it demonstrated the effectiveness of ANN in handling multiple environmental parameters for TSS prediction. The authors used 26 parameters, including nitrate concentrations, total phosphorus, turbidity, and the trophic state index, collected over a ten-year period. By employing a three-layer ANN model, the researchers achieved an R^2 of 0.9589, a Willmott index (WI) of 0.9933, and an RMSE of 0.4753, suggesting excellent model performance. These results demonstrated the potential of machine learning in water quality monitoring and highlighted the ability of ANN models to process complex, high-dimensional datasets.

However, while this study offered promising results, its application is limited to a specific, localized water body. The Fei Tsui Reservoir, being a small and isolated water body, is subject to limited environmental influences, mainly urban pollution. The study's findings are not easily generalized to large and more complex river systems, especially those with multiple anthropogenic impacts. Large rivers like the Rhine River, which experience industrial, agricultural, and urban runoff, present different challenges in TSS prediction. The study's narrow geographic scope means it does not account for variations in TSS levels due to diverse land uses, such as the industrial runoff seen in Duisburg or agricultural runoff in Bonn. This highlights the need for studies that apply machine learning models to more complex environments, where multiple types of anthropogenic impacts are present.

In a similar vein, Chongyang Wang et al. [13] explored the use of Landsat imagery to develop a retrieval model for TSS concentrations in estuarine and coastal waters across five regions in China. Their study used a quadratic model based on the logarithmic transformation of the red and near-infrared bands of Landsat imagery to estimate TSS levels. The model was able to explain approximately 72% of the variation in TSS concentration ($R^2 = 0.72$) and achieved a satisfactory RMSE of 25 mg/L. The study's focus on coastal and estuarine regions makes it relevant to understanding TSS dynamics in coastal waters, where TSS concentrations can fluctuate due to natural factors such as tides, waves, and sedimentation processes.

Despite the promising results, Wang et al.'s [13] model has limitations in its applicability to more complex river systems. First, the model's performance was constrained by a narrow range of TSS concentrations (4.3–577.2 mg/L), which may not adequately represent the extreme TSS levels observed in industrialized or agricultural regions. Moreover, the study was geographically limited to five regions in China, which means the model's applicability to other regions with different land uses or anthropogenic influences, such as industrial discharge or agricultural runoff, remains uncertain. The study also did not account for the influence of human activities, such as pollution from urban, industrial, or agricultural sources, which can cause substantial variations in TSS concentrations, particularly in larger rivers. The lack of variability in land use and pollution sources in the study suggests that it may not be directly applicable to rivers like the Rhine, which experience a diverse set of anthropogenic influences.

2.3 Remote Sensing and Machine Learning for TSS Prediction in Large Water Bodies

Recent studies have increasingly focused on using remote sensing data combined with machine learning techniques for large-scale TSS prediction. Lucas Silveira Kupssinskü et al. [14] employed Sentinel-2 spectral data and machine learning algorithms to predict both TSS and chlorophyll-a concentrations across various water bodies. Their models demonstrated an R^2 value above 0.8 for both parameters, indicating the potential of remote sensing for estimating water quality parameters. The study utilized high-resolution satellite data, which provided more precise estimates compared to lower-resolution sensors.

However, the study's reliance on high-resolution satellite imagery presents challenges for large-scale or long-term monitoring projects. High-resolution data, while accurate, is often expensive and not always accessible for large-scale applications, particularly in regions where frequent monitoring is required. Furthermore, Kupssinskü et al. [14] did not fully address the impact of human activities on TSS variability, which is a significant concern in industrial or agricultural regions. In areas like Duisburg, where industrial runoff significantly impacts water quality, or Bonn, where agricultural runoff contributes to TSS levels, the ability to account for these anthropogenic influences is crucial. The study's narrow focus on spectral data without considering the full range of potential pollutants makes it less applicable to regions with complex land-use dynamics.

In the same year, Nicole M. DeLuca et al. [15] expanded on the use of multispectral data for TSS estimation in Chesapeake Bay. By using multispectral information from the MODIS sensor, they compared the performance of a Random Forest regression model with a widely used single-band algorithm for TSS estimation. The multispectral model outperformed the single-band approach, particularly under high TSS conditions, suggesting that multispectral data can improve the accuracy of remote sensing models. While this study provided valuable insights, it also highlighted the increased complexity and data requirements associated with multispectral models. The computational cost and data demands may limit their use in real-world applications, especially in resource-constrained environments.

One of the significant drawbacks of the study by DeLuca et al. [15] is its limited exploration of the role of anthropogenic factors in TSS variability. Although the model was able to estimate TSS levels effectively, it did not examine how different pollution sources, such as industrial runoff or agricultural discharge, affect TSS concentrations. In large river systems like the Rhine, which is influenced by both industrial (Duisburg) and agricultural (Bonn) activities, understanding the differential impacts of these sources on TSS is essential. Thus, while the study demonstrated the potential of multispectral data for TSS prediction, it did not fully address the broader applicability of this approach in regions with complex anthropogenic influences.

2.4 The Role of Deep Learning and High-Resolution Satellite Data

JunGi Moon et al. [16] applied Convolutional Neural Networks (CNNs) to high-resolution Sentinel-2 satellite imagery for TSS prediction in South Korean rivers. The model demonstrated superior accuracy compared to traditional regression methods, with a Nash-Sutcliffe Efficiency (NSE) score of 0.758, indicating good model performance in capturing complex, non-linear relationships between satellite imagery and water quality parameters. However, the study faced limitations in terms of geographical scope, as it was focused solely on South Korean rivers, and the reliance on high-resolution imagery made the approach costly and less scalable. The study's findings, while promising, are not directly applicable to larger rivers in other regions, especially those where high-resolution satellite data may not be readily available.

Similarly, Saiful Haque Rahat et al. [17] employed Long Short-Term Memory (LSTM) networks to predict TSS concentrations using MODIS reflectance data in the Ohio River Valley. The LSTM model effectively captured the spatiotemporal variations in TSS, accounting for multidimensional uncertainties, including hydro-climatic factors and land-use features. However, the study was limited by its use of a single satellite dataset and focused mainly on one geographical region. The reliance on MODIS data and the lack of exploration into alternative machine learning algorithms or the integration of multiple satellite sensors limits the study's generalizability. Additionally, the study did not fully explore the influence of human activities, such as industrial or agricultural runoff, which can significantly affect TSS concentrations in rivers like the Rhine.

2.5 Challenges in Applying Remote Sensing for TSS Estimation in Complex River Systems

Several studies have demonstrated the effectiveness of remote sensing techniques in estimating Total Suspended Solids (TSS) in different environmental contexts. However, most of these studies focus on smaller, more localized systems or specific geographical regions, limiting the applicability of their findings to larger, more complex river systems. For instance, the study by Celso M. Isidro et al. [18] focused on small rivers in the Philippines, using satellite datasets like RapidEye and Pleiades-1A to develop a power law model to predict TSS concentrations. The model performed well at lower concentrations but struggled at higher concentrations, where TSS reflectance increased at a slower rate. This suggests that satellite imagery may not always be effective in large rivers with high TSS variability, particularly during extreme events like floods or storms. The study's narrow focus on small rivers with limited anthropogenic influences further restricts its generalizability to large, industrialized river systems like the Rhine.

Similarly, the research by Jing Zhao et al. [19] used the Markov model in combination with Landsat imagery to track TSS concentration changes in the Hedi Reservoir in China. While the study successfully demonstrated how seasonal and anthropogenic factors, such as rainfall, influence TSS dynamics, its focus on a specific reservoir and the sensitivity of the model to atmospheric interference limit its applicability to larger rivers. The study did not explore how industrial or agricultural runoff—key factors in large river systems—affects TSS concentrations, further highlighting the gap in understanding the role of human activities in TSS variability.

2.6 The Role of Multi-Source Satellite Data in Enhancing TSS Prediction Models

Recent advancements in the fusion of multi-source satellite data have shown promise in improving TSS estimation accuracy. For example, Xiang Zhang et al. [20] utilized a combination of Landsat and MODIS data to estimate TSS concentrations in Jiaozhou Bay, China. Their study demonstrated that integrating data from different satellite sensors could help capture the spatiotemporal variation of TSS more effectively. However, as with other multi-source data models, the integration posed challenges related to sensor discrepancies and the need for accurate calibration between datasets. This issue was further exacerbated by the study's lack of consideration for the influence of anthropogenic activities, such as industrial discharge or agricultural runoff, which play a major role in TSS levels in large rivers.

In comparison, Elisa Friedmann et al. [21] explored spatiotemporal sensor fusion by combining Landsat/Sentinel-2 and MODIS data to estimate TSS concentrations across river sites in the United States. The fusion of high-resolution imagery with the temporal frequency of MODIS data significantly improved TSS prediction accuracy. Despite this, the study faced challenges related to the variability in sensor performance and inherent uncertainties in satellite-derived data. The model also did not fully explore how complex land-use activities, including industrial discharges and agricultural runoff, contribute to TSS concentrations. This gap is especially relevant when applying remote sensing and machine learning to complex river systems like the Rhine River, where diverse anthropogenic influences must be considered.

2.7 Addressing Data Accessibility and Computational Challenges in Remote Sensing Models

Many of the studies discussed above have faced challenges related to data accessibility, computational demands, and the need for frequent calibration with ground-based measurements. For example, Rafael Luís Silva Dias et al. [22] applied machine learning algorithms to estimate TSS concentrations in small reservoirs using remotely piloted aircraft (RPA)-based multispectral sensors. While the study showed that RPAs could provide high accuracy in TSS estimation, the approach is limited by the need for costly infrastructure and operational costs, making it less feasible for long-term or large-scale monitoring projects. In contrast, satellite-based methods, such as those employed by M.-A. Torres-Vera , using Landsat imagery for TSS prediction, are more cost-effective and scalable. However, these models often require frequent calibration with in-situ data, which may not always be available in large, complex river systems like the Rhine.

Moreover, high-resolution satellite data is often required for accurate TSS estimation, but such data is not always accessible for large-scale, long-term monitoring projects due to its high cost and limited temporal availability. This issue is further compounded by the difficulty of integrating multi-source satellite data, which can introduce inconsistencies in model outputs. For example, while the study by Godson Ebenezer Adjovu et al. [6] emphasized the potential of remote sensing in overcoming the limitations of traditional methods for TSS and Total Dissolved Solids (TDS) measurement, it also highlighted the challenges posed by atmospheric interference and sensor resolution. These issues limit the ability of remote sensing models to provide consistent and accurate TSS predictions, especially in regions with complex land uses or high levels of human activity.This research

2.8 Bridging the Gaps in TSS Prediction through this Research

While the studies discussed above offer valuable insights into the application of remote sensing and machine learning for TSS prediction, they each face significant limitations in terms of geographical scope, applicability to diverse anthropogenic impacts, data accessibility, and computational demands. The gaps identified in these studies include:

1. **Limited Application to Larger, More Complex River Systems**: Most of the studies focused on small water bodies, such as reservoirs, coastal regions, or estuaries, where the TSS dynamics are relatively simple and less influenced by diverse anthropogenic activities. In contrast, large rivers like the Rhine are subject to more complex interactions between various land uses (industrial, agricultural, urban), which significantly affect TSS concentrations.

2. **Narrow Geographic Focus**: Several studies concentrated on specific regions, such as South Korea (JunGi Moon et al., 2021) or the Philippines (Celso M. Isidro et al., 2019), limiting their applicability to other regions with different environmental and anthropogenic conditions. The lack of studies focused on large European river systems like the Rhine, which spans both industrial and agricultural regions, creates a significant research gap.

3. **Challenges in Data Accessibility and Computational Demands**: The use of high-resolution satellite data and complex machine learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, has proven effective but comes at a high computational cost. For large-scale, long-term monitoring projects, the reliance on high-resolution

data may not be feasible, especially in regions where such data is difficult to obtain.

4. **Integration of Anthropogenic Activities into TSS Estimation Models**: Many studies did not fully explore the role of human activities, such as industrial discharge or agricultural runoff, in influencing TSS concentrations. In large river systems with diverse pollution sources, such as Duisburg (industrial) and Bonn (agricultural) in the Rhine River, understanding the differential impacts of these activities is crucial for accurate TSS prediction.

This research on the Rhine River is poised to address these gaps by leveraging machine learning models, such as Random Forest and XGBoost, in combination with satellite data from Landsat, to provide scalable, long-term solutions for TSS estimation. By focusing on two contrasting regions—Duisburg, with its industrial influences, and Bonn, with agricultural runoff—this research will contribute valuable insights into how different human activities affect TSS levels in large river systems. The use of Landsat imagery, a more accessible satellite data source, will also ensure that the research is cost-effective and feasible for large-scale monitoring. Additionally, by integrating multiple data sources, such as discharge data and satellite imagery, this study will improve the robustness and accuracy of TSS prediction models, making them more applicable to other large river systems worldwide.

In conclusion, the current literature on TSS prediction highlights significant advancements in the application of machine learning and remote sensing. However, key gaps remain in the geographical scope, data accessibility, and integration of anthropogenic impacts. This research aims to bridge these gaps by applying machine learning models to the Rhine River, offering a scalable and cost-effective solution for TSS estimation in large, complex river systems affected by diverse pollution sources.

CHAPTER 3
METHODOLOGY

3.1 Introduction

This methodology outlines the comprehensive process used to collect, preprocess, and analyze environmental data for predicting Total Suspended Solids (TSS) concentrations in the Rhine River using machine learning techniques. The research leverages multiple data sources, including historical TSS measurements from the Global River Water Quality Archive (GRQA), river discharge data from the Global Runoff Data Centre (GRDC), and satellite imagery from Landsat 5 and 7 accessed via Google Earth Engine (GEE). These datasets span a 15-year period (January 1994 to August 2009), offering both temporal depth and spatial coverage for robust environmental analysis. The methodology emphasizes the importance of spatial resolution, data consistency, and variable integration to enhance model performance. In particular, this section details the criteria for dataset selection, the rationale behind the chosen time period, and the multi-stage data processing steps employed to ensure the accuracy, reliability, and interpretability of the final dataset used for machine learning model training and evaluation [23].

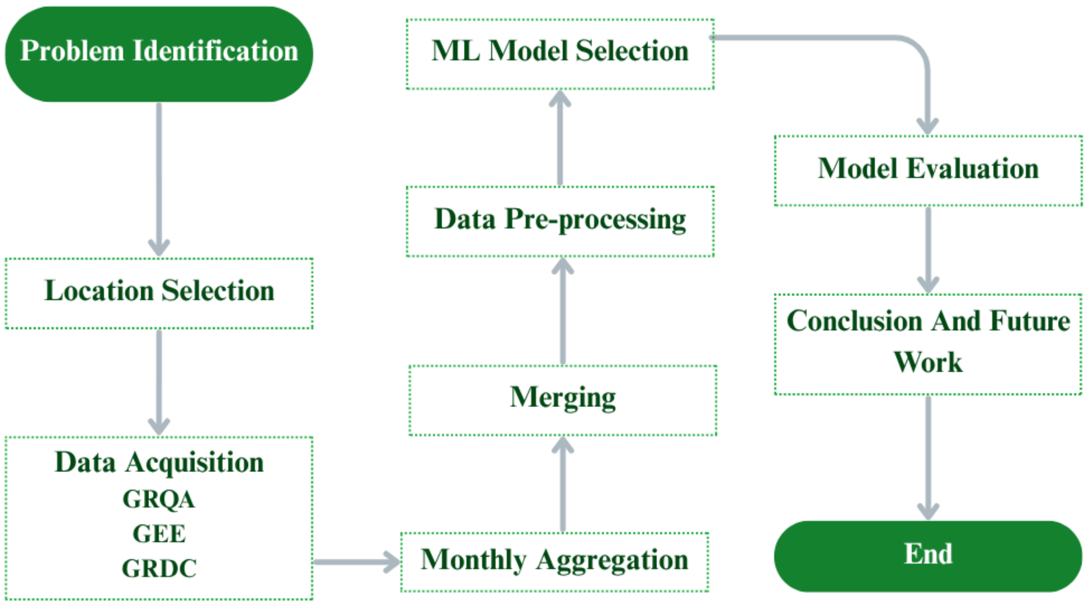

Figure 3-1. Overview of the methodology. Dose response characteristics of the transcriptomic profiles are quantified at three different biologically relevant molecular level — biomarker/gene, stress pathway, and cellular/ system level. Dose-response models are generated for static snap-shot or single time-point responses and time-aggregated responses at all three level. Hill-type model is used to fit the dose-response profiles, from which benchmark dose is estimated as the point of departure.

3.2 The Global River Water Quality Archive (GRQA)

The Global River Water Quality Archive (GRQA) is one of the most comprehensive global datasets available for analyzing riverine water quality [24]. Developed by the University of Birmingham and made openly accessible via the Zenodo repository, GRQA aggregates over 12 million in-situ water quality measurements spanning more than 12,000 monitoring stations across 120 countries and over 1,000 rivers worldwide. It includes diverse physicochemical parameters such as Total Suspended Solids (TSS), biochemical oxygen demand (BOD), dissolved oxygen (DO), pH, electrical conductivity, nitrate levels, and water temperature [25].

What makes GRQA particularly valuable for scientific research is its high spatial and temporal resolution, consistency of data collection, and transparent documentation. The archive harmonizes data from multiple national monitoring agencies and institutes, offering long-term time series suitable for hydrological and environmental modeling. Importantly, it provides metadata on the measurement techniques, units, and data quality flags, allowing researchers to trace the origin and reliability of each data point.

The dataset’s open-access nature, combined with rigorous methodological standards, makes it an ideal foundation for studies requiring large-scale and long-term water quality information. In the context of this research, GRQA's extensive data coverage on the Rhine River enables a robust analysis of sediment dynamics using machine learning models [26].

3.3 Total Suspended Solids (TSS) and GRQA Data Source

Total Suspended Solids (TSS) refer to the mass of organic and inorganic particles suspended in water, including silt, clay, algae, detritus, and anthropogenic pollutants. TSS plays a central role in evaluating water quality as it directly influences turbidity, light availability, aquatic plant productivity, sedimentation rates, and the transport of attached contaminants [27]. Elevated TSS levels are typically associated with stormwater runoff, erosion, and industrial or agricultural discharges—making it a sensitive indicator of both natural hydrological processes and human impact on river systems [28].

For this study, TSS was selected as the exclusive water quality parameter due to several compelling reasons. First, it acts as an integrative measure, encapsulating the effects of multiple upstream activities and environmental conditions. Second, TSS exhibits strong correlations with satellite-derived reflectance values and spectral indices, enabling its effective estimation through remote sensing. Third, among the water quality parameters available in GRQA, TSS provides the most consistent and uninterrupted data series for the chosen period (1994–2009), facilitating robust time-series modeling without significant gaps or inconsistencies [29].

Using GRQA as the data source for TSS further strengthens the scientific reliability of the study. Its high-frequency, long-term measurements for key monitoring stations along the Rhine River offer a credible ground truth for training and validating machine learning models. Furthermore, the archive's standardized measurement units and clearly documented metadata allow for seamless integration with other datasets, such as satellite imagery and discharge records [30]. By focusing exclusively on TSS and utilizing GRQA as the authoritative source, the study ensures both ecological relevance and methodological rigor, forming a solid foundation for the development of predictive models for riverine sediment dynamics.

3.4 Study Locations

This study examines Total Suspended Solids (TSS) dynamics in two geographically and anthropogenically distinct segments of the Rhine River: Duisburg and Bonn. The rationale for selecting these locations is rooted in their contrasting environmental characteristics and human activities, which together offer a comprehensive representation of the diverse sedimentation patterns within the Rhine basin. Their inclusion strengthens the model's capacity to generalize across variable ecological and hydrological conditions.

3.4.1 Duisburg – Industrial Downstream Region

Duisburg is situated in the Lower Rhine Plain, within the downstream section of the river. It is one of Germany's most industrialized urban centers, known for hosting Europe's largest inland port and an extensive array of steel, chemical, and logistics industries [31]. Figure 3.1 illustrates the geographical and industrial landscape of Duisburg, highlighting its position in the Lower Rhine Plain and its susceptibility to elevated TSS levels due to dense urbanization and point-source pollution. The region is characterized by flat floodplain terrain, a dense urban fabric, and a highly engineered river channel with limited natural buffering capacity.

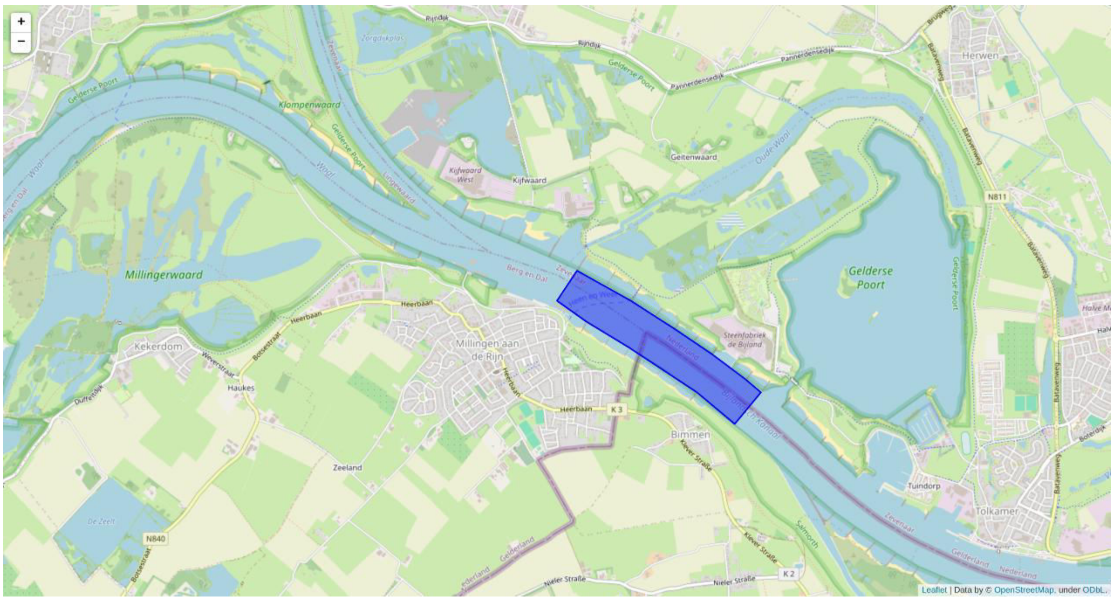

Figure 3.1: Study area polygon in Duisburg, showing the industrial zone along the Lower Rhine

The hydrological and land-use profile of Duisburg makes it particularly susceptible to high levels of TSS during stormwater runoff events, when impervious surfaces and industrial zones contribute significantly to sediment discharge into the river. The prevalence of point-source pollution, urban drainage systems, and industrial effluents makes Duisburg an ideal location to study the impacts of urbanization and industrial activity on sediment load and water quality.

3.4.2 Bonn-Agricultural Upstream design

Bonn, in contrast, lies upstream in the Upper Middle Rhine Valley, a region designated as a UNESCO World Heritage Site. It is characterized by steep-sloped topography, vineyards, and extensive agricultural activity. Unlike Duisburg, Bonn’s riverine environment is shaped largely by non-point source pollution from runoff containing fertilizers, pesticides, and soil sediments, particularly during spring and summer when farming activity intensifies [32]. Figure 3.2 depicts the geographical features and land use characteristics of the Bonn region, emphasizing its steep terrain and agricultural dominance that contribute to seasonal TSS variations.

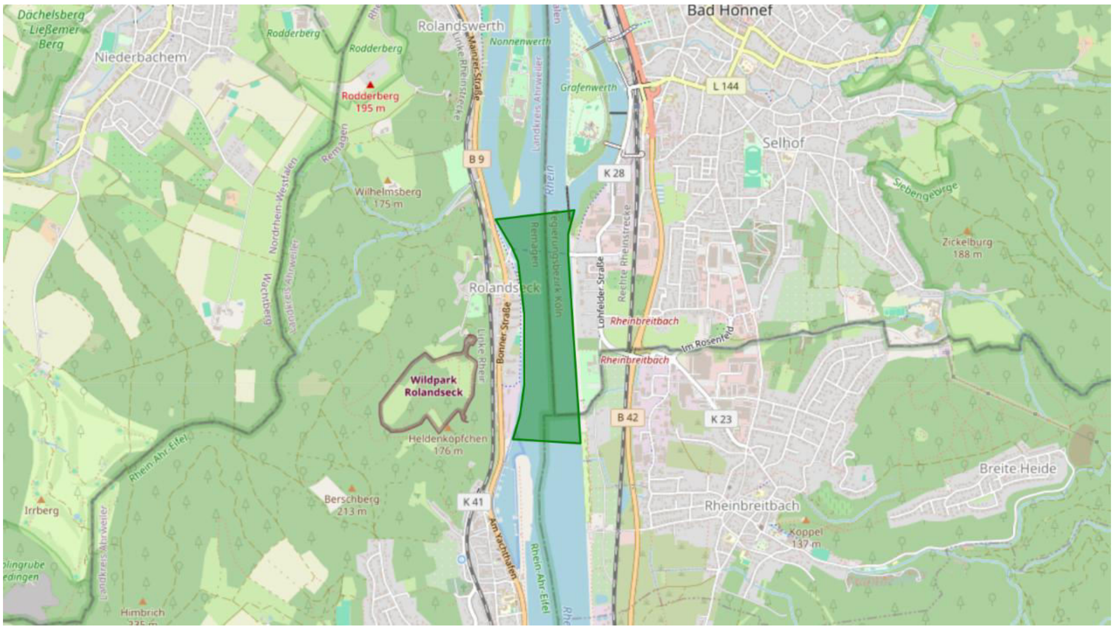

Figure 3.2: Study area polygon in Bonn, highlighting the agricultural zone and rural watershed conditions.

The steep gradients facilitate rapid overland flow during rain events, mobilizing fine particles from agricultural fields into the river. This leads to seasonal spikes in TSS, providing a sharp contrast to the more continuous industrial discharge observed in Duisburg. Bonn thus represents a landscape where land use and topography jointly influence sediment transport mechanisms in a predominantly rural watershed.

3.4.3 Comparative and Analytical Value

The 130-kilometer separation between Duisburg and Bonn encompasses both geomorphological diversity (floodplain vs. sloped terrain) and distinct anthropogenic pressures (industrial vs. agricultural land use). This natural gradient enables the study to capture a wide spectrum of TSS behavior, influenced by different sources and environmental conditions.

By analyzing and eventually merging the datasets from these two locations, the research achieves several key objectives:

- Enhances model training by exposing algorithms to a broader range of conditions, thereby increasing predictive robustness.
- Facilitates comparative analysis, allowing the disentanglement of TSS responses to geography and human activity.
- Supports the development of scalable, transferable machine learning models that could be applied to other river systems experiencing similar anthropogenic gradients.

In essence, the deliberate selection of Duisburg and Bonn ensures both contextual richness and methodological strength, laying a strong foundation for the integration of remote sensing, hydrological, and machine learning techniques in TSS modeling.

3.5 Acquisition and Processing of TSS Data from GRQA

To build a time-series dataset suitable for machine learning-based TSS prediction, monthly-averaged Total Suspended Solids (TSS) values were extracted from the Global River Water Quality Archive (GRQA) [33]. The dataset was downloaded in CSV format directly from the Zenodo repository. GRQA provides individual measurement records with associated metadata, including station codes, sampling dates, and water quality parameter names and values.

3.5.1 Selection of Monitoring Stations

Two monitoring stations were selected along the Rhine River, corresponding to the study’s two target locations:

- Station 301593 – representing Duisburg, a highly industrialized downstream region.
- Station 301579 – representing Bonn, an upstream location characterized by agricultural land use.

These stations were chosen due to their geographic proximity to the satellite imagery sampling areas, as well as the consistency and availability of TSS data throughout the 1994–2009 study period.

3.5.2 Data Cleaning and Outlier Removal

Before performing any temporal aggregation, the raw GRQA TSS data was filtered to exclude entries flagged by GRQA as outliers or invalid measurements. These flags were determined based on GRQA's internal quality assurance criteria, which consider anomalies in concentration levels, sampling frequency, and unit inconsistencies.

Only those TSS records explicitly marked as valid and falling within the acceptable concentration range for freshwater rivers were retained. This step ensures the reliability of the ground-truth data used for model training and validation [34].

3.5.3 Monthly Aggregation

Since TSS samples in GRQA are recorded on irregular dates (e.g., multiple entries within a month or missing months entirely), a monthly aggregation approach was applied to standardize the temporal scale. For each location, the following steps were followed:

1. All valid TSS values recorded within the same calendar month were grouped together.
2. The arithmetic mean of these values was computed to represent the monthly average TSS concentration for that month.
3. If no measurements existed for a particular month, it was recorded as missing for further interpolation or exclusion.

This averaging process was performed programmatically using Python, with libraries such as pandas for date parsing, filtering, and group-based aggregation. This allowed precise alignment of the TSS data with satellite-derived variables and river discharge data, both of which were also processed at the monthly scale.

3.5.4 Final Output

The output from this process consisted of a structured tabular dataset with three columns: year, month, and TSS (mg/L). This cleaned and temporally aggregated dataset served as the target variable (dependent variable) in the subsequent machine learning modeling pipeline.

3.6 River Discharge and Integration of GRDC Data

3.6.1 Conceptual Importance of River Discharge in Sediment Transport

River discharge, commonly measured in cubic meters per second (m^3/s), is a fundamental hydrological variable that quantifies the volume of water passing through a given cross-section of a river over time [35]. It serves as a dynamic indicator of the catchment's hydrological response to precipitation, land cover, evapotranspiration, snowmelt, and anthropogenic water management activities. In riverine systems, discharge governs not only the transport of water but also the mobilization, suspension, and deposition of sediments—factors directly related to Total Suspended Solids (TSS) concentrations [36].

High discharge events, such as those caused by rainfall or snowmelt, increase hydraulic energy and sediment-carrying capacity, often leading to elevated TSS levels through bank erosion, resuspension of bed materials, and overland flow contributions [37]. Conversely, during low-flow conditions, suspended particles tend to settle, reducing TSS concentrations. This bidirectional relationship between discharge and TSS underscores the necessity of incorporating discharge as a temporal explanatory variable in any predictive modeling framework focused on water quality.

Moreover, unlike satellite-derived indices that reflect surface optical properties (e.g., turbidity, vegetation), river discharge provides a subsurface and process-based perspective that complements remote sensing data [38]. It captures hydrological variability and sediment dynamics not observable through reflectance alone. Thus, including discharge enhances the physical interpretability, accuracy, and robustness of the machine learning models employed in this study.

3.6.2 The Global Runoff Data Centre (GRDC): A Hydrological Data Authority

To obtain long-term and reliable discharge data, this study utilized the services of the Global Runoff Data Centre (GRDC), an internationally recognized hydrological data repository. Founded in 1988 and hosted by the German Federal Institute of Hydrology (BfG), the GRDC operates under the guidance of the World Meteorological Organization (WMO) [39]. Its primary mission is to collect, curate, and disseminate streamflow data for scientific, environmental, and operational applications worldwide.

The GRDC archives discharge records from over 10,000 gauging stations located across more than 160 countries, making it one of the most comprehensive global databases for surface water hydrology. The center supports research in climate change, flood risk assessment, transboundary water governance, and catchment-scale hydrological modeling. Key strengths of the GRDC include:

- Standardized data formats for interoperability.
- Quality-assured measurements verified by national hydrological agencies.
- Detailed metadata for each station, including location, elevation, units, and record completeness.
- Access to both daily and monthly time series, depending on station configuration and national reporting standards.

Researchers can access GRDC datasets through the organization’s online data portal, which allows for station-specific queries based on country, river name, or geographic coordinates. Data requests are reviewed by GRDC personnel to ensure appropriate use and are typically fulfilled within a few business days, delivered in standardized CSV or TXT formats ready for analysis.

3.6.3 Selection of Hydrological Stations: Spatial and Temporal Criteria

For this study, two hydrological gauging stations were selected to represent the discharge characteristics at the study sites:

- Station 6435060 – located near Duisburg, serving the industrialized downstream section of the Rhine River.

- Station 6335070 – located near Bonn, capturing conditions in the agricultural upstream region.

These stations were selected based on three principal criteria:

1. Geographical Proximity: Both stations are located within close spatial proximity to the GRQA TSS monitoring stations and the polygons defined for satellite data sampling. This ensures that the discharge data reflect hydrological processes relevant to the same river segments modeled in the TSS predictions.

2. Temporal Coverage: Each station provides uninterrupted monthly discharge records from January 1994 to August 2009, precisely matching the study's temporal scope and the availability of satellite and water quality data.

3. Data Quality and Consistency: The GRDC metadata confirmed the integrity of these stations' records, with minimal gaps, unit consistency (m^3/s), and documented measurement procedures, ensuring their suitability for machine learning applications.

By carefully aligning these hydrological stations with the spatial-temporal footprint of other datasets, the study ensured an internally consistent and integrable data environment, foundational to constructing reliable predictive models.

3.6.4 Data Acquisition Procedure

The discharge records for the selected stations were acquired via the GRDC online data portal. The process involved:

1. Submitting a formal data request through the GRDC portal interface, specifying the station IDs (6435060 for Duisburg and 6335070 for Bonn), the target time range (January 1994 to August 2009), and the intended research use.

2. Upon approval, GRDC provided access to the monthly mean discharge datasets in structured CSV format. Each dataset included:
   - Station identifier and name
   - Country and river name
   - Geographic coordinates and elevation
   - Date of measurement (recorded as Year and Month)
   - Monthly mean discharge value (in m^3/s)
   - Additional metadata fields indicating data origin and quality status

These files were cross-verified to ensure consistency with the requested timeframe and spatial coverage. Visual inspection and summary statistics were used to validate data completeness and check for anomalies (e.g., sudden unrealistic spikes or missing months).

3.6.5 Preprocessing and Structuring for Model Integration

Once downloaded, the discharge data was subjected to a preprocessing workflow designed to align it with the structure and resolution of the other datasets (TSS and satellite indices). The steps were as follows:

1. Column Standardization: The original discharge files were reformatted to retain only the following fields:
   - year: Four-digit year of observation
   - month: Numeric month (1–12)
   - discharge: Monthly mean discharge value in cubic meters per second (m^3/s)
2. Station Segregation: Separate discharge files were maintained for Duisburg and Bonn to allow for location-specific model training and performance comparison. Each dataset included 188 monthly observations, covering the full study period.

3. Missing Data Handling: In this case, no months were missing from either station's record. However, the workflow included provisions for identifying and interpolating missing values, should they arise, using time-series interpolation techniques such as linear or spline methods.

4. Merge Preparation: The discharge datasets were preserved with a consistent date format (year-month) and were structured for horizontal merging with the corresponding TSS and satellite-derived data files. This ensured that each row in the final integrated dataset corresponded to a unique location and month.

3.6.6 Role in Predictive Modeling

The final, cleaned discharge data served as a key independent variable in the machine learning framework. Alongside satellite-derived indices (e.g., NDVI, MNDWI) and spectral band ratios (e.g., B3/B5), discharge values were used as a numerical input feature to help predict monthly TSS concentrations.

Its inclusion allowed the models to:

- Capture temporal peaks in sediment concentration corresponding to flood events.
- Reflect the seasonal variability of sediment transport driven by hydrological cycles.
- Improve generalization by incorporating a non-remote-sensing variable, thereby reducing reliance on surface optical indicators alone.

By combining remote sensing, in-situ TSS measurements, and hydrological discharge data, the study created a tri-source dataset that supports both high predictive accuracy and physically meaningful model interpretation.

3.7 Satellite Data Acquisition and Google Earth Engine (GEE) Processing

3.7.1 Google Earth Engine (GEE): Platform Overview and Justification

Google Earth Engine (GEE) is a cloud-based geospatial processing platform designed for the planetary-scale analysis of satellite imagery and geospatial datasets. It offers an extensive archive of remote sensing data from a wide range of satellite missions, including Landsat, Sentinel, MODIS, and others, combined with powerful

JavaScript and Python-based scripting capabilities [40]. GEE has become one of the most widely adopted platforms in environmental research due to its capacity to process massive volumes of spatial-temporal data without the limitations of local hardware or traditional desktop GIS tools.

In this study, GEE was selected as the primary tool for satellite image processing for several compelling reasons:

- Access to historical satellite archives, including the full Landsat mission history from the 1970s to the present.
- Built-in atmospheric correction and surface reflectance products, which eliminate the need for local preprocessing pipelines.
- Scalable computation using cloud infrastructure, enabling complex filtering, masking, and spectral calculations over large time spans and multiple regions.
- Open access and reproducibility, supported by sharable and transparent script-based workflows.

By leveraging GEE, the study could systematically process Landsat imagery for both Bonn and Duisburg over the full study period (1994–2009), applying consistent methods for water masking, cloud filtering, band scaling, and spectral index computation—all of which are critical for preparing remote sensing features to be used in machine learning models for TSS prediction.

3.7.2 Landsat 5 and Landsat 7: Selection and Characteristics

In selecting suitable Earth observation platforms for the long-term monitoring of Total Suspended Solids (TSS) in the Rhine River from January 1994 to August 2009, this study systematically evaluated available satellite missions based on their temporal coverage, spatial and spectral resolution, data continuity, and relevance to water quality monitoring. The review considered both historical and contemporary missions operated by NASA/USGS and ESA, with particular attention given to their suitability for retrospective time-series modeling using machine learning techniques.

To ensure comprehensive evaluation, the study compiled key attributes of thirteen satellite programs, including Landsat 5, Landsat 7, Landsat 8, Landsat 9, and Sentinel-

2A/2B, as well as other relevant platforms. These attributes included operational timelines, revisit intervals, sensor types, and their capacity to capture indicators of turbidity, chlorophyll concentration, suspended sediment, and water extent—variables either directly or indirectly related to TSS.

The summary table below presents a detailed comparative overview of these satellites, helping to identify the most appropriate sources for consistent, high-resolution, and hydrologically meaningful imagery over the study period.

Table 6.1: Comparison of Earth Observation Satellites for Water Quality Monitoring

Satellite Name	Operator	Operational Timeframe	Revisit Interval	Spatial Resolution	Sensor Type	Water Quality Applications
Landsat 5 (TM)	NASA/USGS	Mar-84 - Jun-13	16 days	30m (multi-spectral), 120m (thermal)	VNIR, SWIR, TIR	Turbidity, Suspended Sediment, Water Extent, Land Cover Change affecting Water Quality
Landsat 7 (ETM+)	NASA/USGS	Apr-99 - Operational (with SLC-off issue since 2003)	16 days	30m (multi-spectral), 60m (thermal), 15m (pan)	VNIR, SWIR, TIR	Turbidity, Suspended Sediment, Water Extent, Land Cover Change affecting Water Quality
Landsat 8 (OLI & TIRS)	NASA/USGS	Feb-13 - Operational	16 days (or 8 days with Landsat 9)	30m (multi-spectral), 100m (thermal), 15m (pan)	VNIR, SWIR, TIR	Turbidity, Suspended Sediment, Water Extent, Land Cover Change affecting Water Quality

Landsat 9 (OLI-2 & TIRS-2)	NASA/USGS	Sep-21 - Operational	16 days (or 8 days with Landsat 8)	30m (multi-spectral), 100m (thermal), 15m (pan)	VNIR, SWIR, TIR	Turbidity, Suspended Sediment, Water Extent, Land Cover Change affecting Water Quality
Sentinel-2A/2B (MSI)	ESA (Europe)	2A: Jun-15 - Operational / 2B: Mar-17 - Operational	10 days solo / ~5 days combined (2–3 days at mid-latitudes)	10m, 20m, 60m (band dependent)	VNIR, Red-edge, SWIR	Turbidity, Chlorophyll-a (coastal/inland), Suspended Sediment, Water Extent

Among the reviewed platforms, Landsat 5 and Landsat 7 emerged as the most appropriate choices for the goals and scope of this research, based on the following reasons:

3.7.2.1 *Temporal Coverage Alignment*

The study period precedes the launch of Landsat 8, Landsat 9, and Sentinel-2 missions, rendering them inapplicable for consistent retrospective analysis. Landsat 5 was operational from 1984 to 2013, fully covering the time frame, while Landsat 7, launched in 1999, provided overlapping coverage for more than a decade. Their combined availability offers continuous temporal coverage with minimal data gaps—critical for capturing seasonal and inter-annual TSS trends

3.7.2.2. *Revisit Cycle and Archive Density*

Both satellites operate on a 16-day revisit cycle, and their overlapping operational years enabled interleaved image acquisition approximately every 8 days. This increased observation frequency boosts the chance of capturing cloud-free scenes within each month, enhancing the temporal robustness of monthly median composites and improving model input quality.

3.7.2.3. *Spatial Resolution Suitability*

The 30-meter spatial resolution of Landsat's multispectral imagery strikes an optimal balance for river-scale analysis—allowing sufficient spatial granularity to detect water bodies, land cover variation, and sediment plumes, without overwhelming data volume or losing regional context. This is especially beneficial for medium-width rivers like the Rhine.

3.7.2.4. *Spectral Compatibility and Index Computation*

Landsat 5 and 7 offer consistent coverage across VNIR and SWIR bands, enabling the derivation of several well-established remote sensing indices used as input features in TSS prediction:

- NDVI, EVI: Indicators of vegetation cover and soil stability
- MNDWI, NDWI: Measures of water clarity and extent
- NDCI: Proxy for chlorophyll and organic matter concentration
- B3/B5 Ratio: Custom indicator for sediment presence

These indices are sensitive to suspended sediments, surface runoff, vegetation-mediated erosion, and algal productivity—all of which influence TSS behavior in riverine systems.

3.7.2.5. Availability of Preprocessed Surface Reflectance in GEE

Google Earth Engine (GEE) offers atmospherically corrected surface reflectance products for both Landsat 5 and 7, eliminating the need for manual preprocessing and ensuring radiometric consistency across the study period. This standardization is essential for multi-year modeling frameworks using machine learning.

3.7.2.6. *Sensor Harmonization and Merging Capability*

Although Landsat 7 experienced a Scan Line Corrector (SLC) failure in 2003, this limitation was mitigated in the study by:

- Applying water and cloud masking
- Using median-based monthly compositing

- Leveraging supplementary observations from Landsat 5

Following preprocessing, the image collections from both sensors were successfully harmonized and merged into a single, location-specific time series for Duisburg and Bonn.

3.7.3 Merging Landsat 5 and 7 Imagery in GEE

After independently processing the image collections for Landsat 5 and Landsat 7, the two datasets were merged into a single harmonized image collection within the GEE environment using the merge() function [41]. This merging was performed after ensuring:

- **Consistent band naming** across both datasets
- **Application of the same preprocessing pipeline**, including water masking, cloud masking, band scaling, and spectral index computation
- **Retention of metadata**, including acquisition date, sensor source, and geographic reference

The rationale for merging was threefold:

1. **Maximizing Temporal Coverage**: Merging both collections increases the frequency of observations within each month, reducing gaps caused by cloud cover or scan-line errors (especially post-2003 for Landsat 7).
2. **Enhancing Data Density**: A denser time series improves the reliability of monthly aggregation and statistical computation (e.g., medians), which are later used as model inputs.
3. **Facilitating Uniform Feature Extraction**: Combining the collections into a single dataset simplifies the downstream processing steps and ensures that all computed indices are stored in a unified format, minimizing the risk of misalignment across features.

After merging, the collection was filtered by date (1994–01–01 to 2009–08–31) and spatial boundary (predefined polygons for Duisburg and Bonn). The processed values were exported as monthly-aggregated CSV files, ready for integration with TSS and discharge datasets.

3.7.4 Satellite-Derived Features: Landsat Bands and Spectral Indices

To model Total Suspended Solids (TSS) concentrations from remote sensing imagery, this study utilized reflectance values from selected bands of Landsat 5 Thematic Mapper (TM) and Landsat 7 Enhanced Thematic Mapper Plus (ETM+), both of which provide surface reflectance data at a spatial resolution of 30 meters and a temporal revisit period of 16 days [42]. This means that each satellite could observe the same location on the Earth approximately once every 16 days, under cloud-free conditions—allowing for high temporal density of observations when data from both sensors are combined. The relevant spectral bands were first scaled to standard surface reflectance values and then used to compute a series of spectral indices known to correlate with vegetation cover, water clarity, sediment load, and organic particulate matter—all of which influence TSS variability in river systems.

3.7.4.1 Landsat Reflectance Bands: Roles and Relevance

- Band 1 (B1 – Blue; 0.45–0.52 µm)

Sensitive to atmospheric scattering, B1 is useful for water body detection and haze correction. It plays a role in vegetation indices like EVI that compensate for atmospheric effects.

- Band 2 (B2 – Green; 0.52–0.60 µm)

Penetrates water bodies better than other visible bands, making it valuable for aquatic vegetation mapping and turbidity assessment. It is a key component in water-related indices such as MNDWI.

- Band 3 (B3 – Red; 0.63–0.69 µm)

Strongly absorbed by chlorophyll in vegetation, this band is often used in combination with NIR to detect vegetation health. It is also responsive to suspended particles and water clarity.

· Band 4 (B4 – Near Infrared [NIR]; 0.76–0.90 µm)

Reflects strongly from healthy vegetation and is thus central to all vegetation indices like NDVI and EVI. It is also used in water and sediment assessments indirectly.

· Band 5 (B5 – Shortwave Infrared [SWIR1]; 1.55–1.75 µm)

Sensitive to soil and moisture content, B5 is effective in differentiating land and water, assessing sediment concentration, and enhancing water index contrast when paired with visible bands.

· Band 7 (B7 – Shortwave Infrared [SWIR2]; 2.08–2.35 µm)

Used for land cover classification, particularly for distinguishing between vegetated and non-vegetated areas. Although not used directly in index formulas, it is scaled and retained for potential future analysis.

These bands were scaled from their raw digital numbers to reflectance values using the formula:

$$\text{Scaled Band} = (\text{Band} \times 0.0000275) - 0.2$$

This calibration was applied consistently to ensure cross-sensor comparability and compatibility with spectral index computation.

3.7.4.2 Spectral Indices and Band Ratios for TSS Prediction

1. Normalized Difference Vegetation Index (NDVI)

Formula: NDVI = (B4 – B3) / (B4 + B3)

Purpose: Assesses vegetation health, which influences soil erosion and sediment runoff.

Interpretation: High NDVI → Dense vegetation → Lower erosion → Lower TSS

Low NDVI → Sparse or degraded vegetation → Higher TSS potential

2. Modified Normalized Difference Water Index (MNDWI)

Formula: MNDWI = (B2 – B5) / (B2 + B5)

Purpose: Highlights water bodies and distinguishes open water from built-up areas and vegetation.

Interpretation: Sensitive to suspended particles
Lower values often indicate turbid water or high sediment load

3. Normalized Difference Water Index (NDWI)

Formula: NDWI = (B3 – B5) / (B3 + B5)

Purpose: Monitors changes in water content and identifies surface water presence.

Interpretation: Effective in distinguishing water from land
Responds to turbidity and sedimentation in water bodies

4. Normalized Difference Chlorophyll Index (NDCI)

Formula: NDCI = (B4 – B5) / (B4 + B5)

Purpose: Estimates chlorophyll content, indicating biological activity and organic matter.

Interpretation: Higher values → Algal blooms, organic matter → Biogenic TSS
Useful in detecting eutrophication-related sediment input

5. Enhanced Vegetation Index (EVI)

Formula: EVI = 2.5 × (B4 – B3) / (B4 + 6 × B3 – 7.5 × B1 + 1)

Purpose: Improves upon NDVI by reducing atmospheric and canopy background noise.

Interpretation: Designed for dense vegetation regions like Bonn
More sensitive to seasonal vegetation cycles than NDVI

6. Band Ratio (B3/B5)

Formula: B3_by_B5 = B3 / B5

Purpose: Custom ratio used to capture the interaction between red reflectance and moisture/sediment-sensitive SWIR.

Interpretation: Higher ratios may reflect increased sediment or turbidity Simple but effective for enhancing contrast between clean and polluted water

Each of these features was computed for every cloud-free, water-masked image within the study period and spatial boundary. The resulting band and index values were used in monthly aggregation procedures and ultimately served as input variables for the machine learning models predicting TSS concentrations.

3.7.5 Satellite Data Processing in Google Earth Engine (GEE)

To generate high-quality satellite-derived input variables for Total Suspended Solids (TSS) prediction, an end-to-end remote sensing preprocessing workflow was developed and executed using Google Earth Engine (GEE). GEE is a cloud-based geospatial analysis platform capable of processing massive volumes of Earth observation data. The platform provides access to atmospherically corrected, analysis-ready imagery from multiple satellite missions, including the Landsat archive. This section elaborates on the detailed methodology followed within GEE to extract monthly median values of specific reflectance bands and spectral indices from Landsat 5 and 7 imagery over the two selected study locations: Duisburg and Bonn.

3.7.5.1 *Definition of Region of Interest (ROI)*

Each study site was delineated using a set of geographic coordinates that form a polygon encompassing the relevant section of the Rhine River. These polygons were defined using the ee.Geometry.Polygon function in GEE. The regions were selected to focus only on the parts of the river directly associated with in-situ TSS and discharge measurements, minimizing the inclusion of irrelevant land features or mixed pixels.

3.7.5.2 *Water Body Masking*

To restrict analysis strictly to water pixels, a water occurrence mask was applied using the JRC Global Surface Water dataset. This dataset provides a per-pixel estimate of the percentage of time each pixel was covered by water between 1984 and 2020. The 'occurrence' band was used with a threshold greater than 50% (i.e., gt(50)) to ensure that only pixels frequently covered by water were retained. This reduced noise from ephemeral features and ensured that indices reflect aquatic conditions, not surrounding land.

3.7.5.3 *Relaxed Cloud Masking*

Cloud contamination is a major issue in optical remote sensing. To filter out cloud-affected pixels, a relaxed masking approach was implemented using the QA_PIXEL band in Landsat Collection 2 surface reflectance data. The function relaxedMaskQA evaluates bit 3 in the QA_PIXEL band, which corresponds to cloud presence. Pixels flagged as cloudy are masked out. The relaxed nature of this mask allows for partially cloudy images to be retained if the cloud-free portion still covers the target water region.

3.7.5.4 Scaling of Surface Reflectance Bands

Landsat Collection 2 surface reflectance bands are delivered as scaled integers. To convert them into physically meaningful reflectance values, a rescaling formula was applied to Bands 1, 2, 3, 4, 5, and 7. The formula used was:

Scaled Band = (Raw Band × 0.0000275) – 0.2

This operation standardizes reflectance values across both Landsat 5 and 7, facilitating accurate computation of vegetation and water quality indices.

3.7.5.5 Spectral Index Computation

Several spectral indices were computed to serve as proxies for vegetation health, sediment load, water clarity, and biological activity—factors that are known to influence TSS:

- NDVI = (B4 – B3) / (B4 + B3) – Reflects vegetative density and stability of nearby land.
- MNDWI = (B2 – B5) / (B2 + B5) – Useful for enhancing open water and detecting turbidity.
- NDWI = (B3 – B5) / (B3 + B5) – Indicates water presence and helps differentiate turbid zones.
- NDCI = (B4 – B5) / (B4 + B5) – Estimates chlorophyll concentration related to

organic TSS.

- EVI = 2.5 × (B4 – B3) / (B4 + 6 × B3 – 7.5 × B1 + 1) – Enhances vegetation detection, compensates for atmospheric effects.
- B3/B5 Ratio = B3 / B5 – A custom metric representing reflectance contrast due to sediments.

3.7.5.6 *Image Collection Filtering and Temporal Aggregation*

Landsat 5 and 7 image collections were filtered by spatial bounds (defined ROI polygons) and temporal range (January 1994 to August 2009). For each image, the preprocessing pipeline was applied: cloud and water masking, band scaling, and index calculation. Then, the reduceRegion function was used to compute the median value of each band and index across the ROI. The median is preferred over the mean as it reduces sensitivity to residual clouds or anomalous pixel values.

3.7.5.7 *Metadata Enrichment*

Each processed image feature was enriched with metadata for temporal tracking and dataset integrity. The attributes added included: image timestamp (system:time_start), calendar year, month, and the source identifier (L5 or L7) to indicate the satellite of origin. This metadata enables alignment with discharge and TSS time series and allows model training on a monthly basis.

3.7.5.8 *Merging Landsat 5 and 7 Collections*

After preprocessing the two Landsat collections independently, they were merged using the merge() function in GEE. This created a single, unified time series that benefits from Landsat 5's extensive historical archive and Landsat 7's increased scene frequency. This merging enhances monthly coverage and minimizes missing data issues due to cloud cover or scan-line failures.

3.7.5.9 *Null Value Filtering and Export*

To ensure analytical reliability, the merged dataset was filtered using ee.Filter.notNull(['NDVI']), eliminating features with incomplete or missing values due to masking. The resulting filtered image collection was then exported to Google Drive in CSV format using Export.table.toDrive(). The exported file contained monthly median values of all relevant bands and indices, along with full metadata. This file was used as the satellite input in the merged dataset for TSS prediction modeling.

3.8 Data Merging

After completing the individual collection and preprocessing steps for TSS (from GRQA), discharge (from GRDC), and satellite-derived features (from GEE), the datasets were aggregated and merged to form a unified analytical table. This step ensured all environmental variables relevant to Total Suspended Solids (TSS) modeling were aligned temporally and spatially.

Each dataset was first aggregated at the **monthly** level, using 'year' and 'month' as the primary keys. These monthly means were then matched across the three data sources for each location. The resulting tables for Duisburg and Bonn each contained 188 monthly observations spanning from January 1994 to August 2009.

The structure of the merged CSV file for each location included the following 16 columns:
- year
- month
- TSS (mg/l)
- discharge
- B1, B2, B3, B4, B5, B7 (reflectance bands)
- B3_by_B5 (band ratio)
- EVI, NDVI (vegetation indices)
- NDWI, MNDWI (water indices)
- NDCI (chlorophyll index)

Once the location-wise datasets were finalized, they were **vertically concatenated** into a single master CSV file containing 376 rows (188 for Duisburg and 188 for Bonn) and 16 columns. This final merged dataset served as the input for all subsequent data analysis, skewness handling, and machine learning modeling.

To enhance reproducibility and allow others to understand the structure of the dataset, a sample screenshot showing a few rows from the final merged dataset may be optionally included in the appendix. This visual aid provides readers with clarity regarding variable names, units, and formatting.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	year	month	discharge	TSS(mg/l)	B1	B2	B3	B3_by_B5	B4	B5	B7	EVI	MNDWI	NDCI	NDVI	NDWI
302	2003	5	19.129	1686.45	0.10428	0.10842	0.09741	1.06815	0.17739	0.09582	0.0686	0.19492	0.08101	0.30023	0.27437	0.02106
303	2003	6	18.2483	1531.33	0.06077	0.0617	0.05269	2.04183	0.06823	0.02417	0.01896	0.03956	0.42275	0.44094	0.13586	0.3425
304	2003	7	12.3871	1154.36	0.05244	0.06687	0.05834	2.14074	0.06415	0.03936	0.03228	0.01752	0.37657	0.31416	-0.00416	0.31436
305	2003	8	11.271	903.806	0.04572	0.06054	0.04651	2.83105	0.06205	0.01789	0.01019	0.0383	0.56079	0.55014	0.11095	0.46888
306	2003	9	10.2	809.033	0.03054	0.04288	0.03188	3.98159	0.0252	0.0081	0.00619	-0.01488	0.68335	0.50509	-0.1047	0.58363
307	2003	10	14.5733	1157.1	0.10169	0.11493	0.10325	3.28249	0.1047	0.05859	0.03999	0.00903	0.46359	0.36257	-0.11328	0.422
308	2003	11	7.32	1047.8	0.04943	0.05815	0.05017	2.21731	0.03933	0.02398	0.01533	-0.02063	0.4489	0.28276	-0.09487	0.37836
309	2003	12	9.66129	1180.42	0.0318	0.04258	0.03075	2.72841	0.01527	0.00665	0.00746	-0.03821	0.72957	0.39346	-0.33679	0.5784
310	2004	1	24.424	2899.07	0.11274	0.14807	0.15296	8.13842	0.09367	0.01274	0.00713	-0.12431	0.80875	0.70071	-0.24019	0.79637
311	2004	2	16.2207	2091.03	0.09873	0.10089	0.08421	6.99034	0.06664	0.03773	0.03023	-0.05173	0.62957	0.44629	-0.20777	0.57667
312	2004	3	11.4	1536.45	0.03349	0.04135	0.03021	3.16643	0.02396	0.00879	0.00729	-0.0183	0.64199	0.41794	-0.13152	0.52463
313	2004	4	12.2333	1561.33	0.04653	0.05569	0.04341	2.46595	0.04731	0.02244	0.01658	0.01237	0.45202	0.38898	0.03335	0.34959
314	2004	5	20.62	1856.13	0.04237	0.05902	0.04701	2.09149	0.05642	0.02682	0.01764	0.02858	0.39048	0.38416	0.08891	0.28969
315	2004	6	25.15	1946	0.04574	0.06659	0.04993	3.78564	0.0392	0.01317	0.00786	-0.02623	0.66944	0.49677	-0.11997	0.58208
316	2004	7	12.2129	1537.42	0.05758	0.06871	0.05328	3.21041	0.05746	0.0266	0.01732	0.01733	0.52896	0.46603	0.01937	0.43326
317	2004	8	13.9226	1430	0.08535	0.10589	0.09364	1.02471	0.13831	0.09848	0.07424	0.09725	0.04455	0.18256	0.17171	0.00287
318	2004	9	10.2133	1285.23	0.03608	0.0517	0.03839	4.35134	0.02961	0.00947	0.00625	-0.01986	0.68415	0.51773	-0.14674	0.60025

Figure X: A representative excerpt of the merged dataset used in this study, comprising monthly records of Total Suspended Solids (TSS), river discharge, reflectance values from Landsat bands (B1–B7), and derived spectral indices (EVI, NDVI, NDWI, MNDWI, NDCI, B3_by_B5). The dataset integrates remote sensing and hydrological parameters across multiple years to support machine learning–based TSS modeling.

3.9 Data Preprocessing and Machine Learning Analysis

3.9.1 Handling Missing Data

Handling missing data is a critical preprocessing step to ensure model performance and data integrity. In this study, missing values were addressed systematically based on both the type of data (numeric vs. categorical) and the quantity of missing values per column. First, missing values in the dataset were counted using the Pandas function:

missing_counts = df.isna().sum()

A custom loop was then used to iterate over each column and apply different filling strategies based on the threshold of 20 missing entries:

- For columns with fewer than 20 missing values:
- Numeric columns were filled with the column mean using `df[col].mean()`.
- Non-numeric (categorical) columns were filled using the mode (most frequent value).

- For columns with 20 or more missing values:
- Numeric columns were filled with the column median using `df[col].median()` to minimize the impact of outliers.
- Non-numeric columns were again filled using the mode.

This approach preserved as much information as possible while reducing the risk of bias introduced by arbitrary imputation.

3.9.2 Data Inspection and Skewness Analysis

Understanding the distribution of the variables is essential for machine learning model performance. All numeric columns (including the 'year' and 'month') were selected and analyzed using the `.skew()` method to quantify skewness—an indicator of asymmetry in a distribution:
- Positive skewness indicates a longer tail on the right side.
- Negative skewness indicates a longer tail on the left side.

Skewness values were printed, sorted in descending order, and used to identify which columns required transformation.

To aid visual interpretation, both histograms (with Kernel Density Estimation curves) and boxplots were created for each numeric column. These plots revealed distribution shape, spread, and the presence of outliers. They were arranged in a grid layout with 3 plots per row and saved as 'distributions_and_boxplots.png' for reporting and diagnostic purposes.

3.9.3 Skewness Correction with Log Transformation

After identifying highly skewed variables (those with skewness > 0.7), log transformations were applied to improve the normality of their distributions. This transformation was implemented using NumPy's `log1p` function, which calculates:

$$\log(1 + x)$$

This method is ideal for handling zero or near-zero values, which would otherwise be undefined under regular logarithmic transformation. The effect of this transformation is twofold:

- It compresses large values, reducing the influence of outliers.
- It spreads out lower values, making the distribution more symmetrical.

A confirmation message was printed for each column, indicating successful transformation.

3.9.4 Machine Learning Models

A suite of regression models was evaluated to predict Total Suspended Solids (TSS) concentrations using the cleaned and transformed dataset. These models included:
- Linear Regression (used as a baseline for comparison)
- Random Forest Regressor
- XGBoost Regressor
- Gradient Boosting Regressor
- Bagging Regressor
- CatBoost Regressor

To further enhance performance, hyperparameter tuning was performed using Grid Search and Random Search methods. This tuning was applied to:
- Random Forest (Level-2 model)
- XGBoost (Level-2 model)

All models were evaluated using key performance metrics such as R-squared (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The best-performing models were selected based on these metrics and further analyzed for feature importance and generalization capacity.

3.9.5 Machine Learning Models Used for TSS Prediction

Linear Regression

Principle:

Linear Regression is a foundational statistical modeling technique that assumes a linear relationship between a dependent variable and one or more independent variables. It fits a straight line (hyperplane) to the data by minimizing the residual sum of squares between the observed and predicted values [43].

Formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Where:

- Y is the predicted output (TSS in this case)
- β_0 is the intercept
- $\beta_1 ... \beta_n$ are the regression coefficients
- $X_1 ... X_n$ are the independent variables (e.g., reflectance bands, indices, discharge)
- ε is the random error term

Suitability:

Linear Regression serves as a baseline model to evaluate the performance of more complex machine learning algorithms. It is easy to interpret and useful for identifying linear trends in the dataset, though it is limited in capturing nonlinear relationships and interaction effects among predictors.

Random Forest Regressor

Principle:

Random Forest is an ensemble learning technique that constructs multiple decision trees during training and averages their outputs for regression tasks. It leverages bootstrap aggregating (bagging) and random feature selection to reduce variance and prevent overfitting [44].

Formula:

$$\text{RF Prediction} = (1 / T) \times \Sigma \text{ [prediction of each tree]}$$

Where T is the number of trees in the forest.

Suitability:

Random Forest performs well with high-dimensional data and can capture complex, nonlinear relationships between variables. It handles missing data and outliers effectively, making it highly suitable for environmental modeling applications such as TSS prediction.

XGBoost Regressor

Principle:

XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that builds trees sequentially. Each new tree aims to correct the errors made by the previous ones, using gradient descent optimization on a defined loss function [45].

Formula:

$$\hat{y}_i = \Sigma f_k(x_i), \quad f_k \in F$$

$$L = \Sigma l(y_i, \hat{y}_i) + \Sigma \Omega(f_k)$$

Where:

- $\hat{y}_i$ is the predicted value
- l is the loss function (e.g., mean squared error)
- Ω is the regularization function
- F is the space of regression trees

Suitability:

XGBoost is known for its superior performance in structured data tasks and competitions. Its regularization techniques and gradient optimization make it robust to overfitting, capable of handling missing values, and highly accurate for TSS modeling with complex feature interactions.

Gradient Boosting Regressor

Principle:

Gradient Boosting builds additive models by fitting a decision tree to the residual errors of previous models. It minimizes a loss function by using gradient descent in function space.

Formula:

$$F_m(x) = F_{m-1}(x) + \eta \times h_m(x)$$

Where:

- F_m is the updated model at iteration m
- $h_m(x)$ is the new decision tree
- η is the learning rate

Suitability:

Gradient Boosting is suitable for moderately sized datasets and can model complex nonlinear relationships effectively. However, it is more sensitive to overfitting and requires careful tuning of hyperparameters such as learning rate and tree depth.

Bagging Regressor

Principle:

Bagging (Bootstrap Aggregating) trains multiple base estimators (often decision trees) on random subsets of the dataset, then aggregates their predictions to form the final output. It reduces variance and increases stability [46].

Formula:

Final Prediction = (1 / n) × Σ [prediction of each model trained on bootstrap sample]

Suitability:

Bagging is particularly useful in reducing overfitting and works well when base models are prone to high variance. It provides solid predictive power and robustness for datasets with natural environmental variability, such as TSS.

CatBoost Regressor

Principle:

CatBoost is a gradient boosting algorithm developed by Yandex, optimized for categorical data but highly effective in general-purpose regression tasks. It uses ordered boosting and efficient handling of categorical features to improve accuracy and speed [47].

Formula:

Similar to other boosting models: iteratively adds trees to minimize the loss Handles categorical features internally using target-based statistics

Suitability:

Although designed for categorical-heavy datasets, CatBoost performs competitively on numerical datasets too. It is known for its fast training and high accuracy, offering a valuable alternative to XGBoost and LightGBM in TSS modeling.

Artificial Neural Network (ANN)

Principle:

Artificial Neural Networks (ANNs) are computational models inspired by the human brain's structure and functionality. They consist of interconnected layers of nodes (neurons), where each connection has an associated weight. ANNs learn to approximate complex functions by adjusting these weights during the training process through backpropagation [48].

Formula:

Each neuron computes a weighted sum of its inputs and passes the result through an activation function:

$$z = \Sigma (w_i x_i) + b$$

$$a = \varphi(z)$$

Where:

- x_i are input values
- w_i are the weights
- b is the bias term
- φ is the activation function (e.g., ReLU, sigmoid)
- a is the output (activation) of the neuron

Suitability:

ANNs are suitable for modeling nonlinear and complex relationships in environmental data, such as TSS prediction. They can capture intricate patterns that traditional models may overlook. However, they require larger datasets for effective learning, and their 'black-box' nature can make interpretability more challenging compared to tree-based models.

3.10 Evaluation Techniques and Visualization Tools

3.10.1 Quantitative Evaluation Metrics

A thorough assessment of the machine learning models' predictive capability was performed using multiple quantitative metrics. These metrics provided insights into the magnitude, direction, and variance of prediction errors, as well as the explanatory strength of the models [49].

R-squared (R^2) – Coefficient of Determination

R^2 measures how well the independent variables explain the variance in the dependent variable—in this case, Total Suspended Solids (TSS). A value closer to 1 implies a strong fit, while values near 0 indicate poor explanatory power.

$$R^2 = 1 - (SS\_res / SS\_tot)$$

Where:

- SS_res: Sum of squared residuals
- SS_tot: Total sum of squares

Mean Absolute Error (MAE)

MAE computes the average magnitude of prediction errors:

$$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i|$$

This metric is particularly useful for interpretability as it expresses errors in original units (mg/L).

Root Mean Squared Error (RMSE)

RMSE penalizes large errors more significantly due to the squaring operation:

$$RMSE = sqrt[(1/n)\ \Sigma\ (y_i - \hat{y}_i)^2]$$

It provides a direct sense of how far predictions deviate from observations, especially in models where high outliers may occur.

Mean Squared Error (MSE)

MSE represents the average of the squared prediction errors:

$$MSE = (1/n)\ \Sigma\ (y_i - \hat{y}_i)^2$$

Though not as interpretable, it is often used in model optimization due to its convexity.

Mean Absolute Percentage Error (MAPE)

MAPE expresses error in relative percentage terms:

$$MAPE = (100/n) \Sigma |(y_i - \hat{y}_i)/y_i|$$

While useful for standardized comparisons, caution is needed if actual values are very close to zero.

3.10.2 Cross-Validation for Model Robustness

To validate the generalizability of the models beyond the training data, a 10-fold cross-validation procedure was conducted. This approach divides the dataset into ten equal parts, iteratively training the model on nine folds and validating it on the remaining one. This rotation continues until each fold has served as the validation set once. The final performance is computed as the average of all ten iterations [50].

This method ensures that each data point contributes both to training and testing, reducing the variance of the performance estimate. It also guards against overfitting and selection bias, offering a reliable insight into how well the model might perform in a real-world application.

3.10.3 Visual Evaluation Techniques

While numerical metrics are informative, visual diagnostics provide an intuitive understanding of model behavior [51]. The following visualization tools were employed:

Performance Scatterplot

This plot compares observed TSS values with predicted values on a Cartesian plane. Ideally, all points lie on the 45-degree line. Deviations from this line reveal overestimations or underestimations. It is especially useful for spotting systematic bias.

Residual Plot

This diagnostic plot shows the residuals (prediction errors) against the predicted values. A good model exhibits a homoscedastic pattern—randomly scattered residuals with no visible trend. Any pattern may suggest model misspecification or nonlinearity in relationships.

Correlation Heatmap

A heatmap of Pearson correlation coefficients was constructed to explore inter-variable dependencies. High correlation between features may imply redundancy, while strong correlation with TSS highlights influential predictors. The Pearson coefficient is calculated as:

$$r = \Sigma[(x_i - \bar{x})(y_i - \bar{y})] / \mathrm{sqrt}[\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2]$$

Prediction vs. Actual Time Series Plot

This overlay of predicted and observed TSS values across months allows for evaluation of temporal trends. A well-performing model will follow the seasonal and interannual dynamics of TSS, capturing peaks and troughs accurately.

Feature Importance Plot

For tree-based models such as Random Forest and XGBoost, feature importance plots were generated. These rank features based on their contribution to prediction, helping interpret the model and assess environmental relevance of input parameters.

Boxplot of Residuals

Boxplots visualize the spread and symmetry of residuals, revealing skewness or outliers. This is valuable in comparing consistency across models or locations.

Radar Plot for Multi-Metric Comparison

A radar plot was used to compare XGBoost and Random Forest models across multiple metrics—R^2, RMSE, MAE, and MSE. These metrics were normalized to allow visual parity. In the radar chart, the area covered by each model reflects its performance across all dimensions. The method offers a consolidated yet comprehensive snapshot of comparative model behavior.

3.10.4 Summary and Model Selection Criteria

The combination of statistical accuracy measures, cross-validation, and visualization provided a holistic understanding of model performance. The final model selection was guided by multiple criteria:

- High R^2 indicating good explanatory power
- Low MAE and RMSE indicating minimal and stable prediction errors
- Visual consistency in residual and time series plots
- Strong alignment with environmental variables based on feature importance

Such a multidimensional evaluation strategy ensured that the selected model is not only statistically sound but also ecologically meaningful and robust for long-term prediction of Total Suspended Solids in the Rhine River.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the comprehensive outcomes of the machine learning modeling process aimed at predicting Total Suspended Solids (TSS) concentrations using satellite-derived indices, hydrological discharge data, and computed spectral indicators. The evaluation includes multiple supervised regression models, comparative cross-validation, and visualization of predictive accuracy across several statistical perspectives. The results are presented in the form of error metrics (MAE, MSE, RMSE, R^2), correlation structures, data distribution diagnostics, and model performance plots.

4.2 Data Distribution and Preprocessing Evaluation

Before training the machine learning models, data distributions were evaluated and transformed to reduce skewness and enhance stability. Visual inspection of the dataset showed significant skew in several predictors such as B1, B3, B5, B7, and the spectral indices NDVI and EVI. To address this, a combination of log transformation and IQR-based outlier capping was applied. The effects of these transformations are displayed through the following histograms and boxplots.

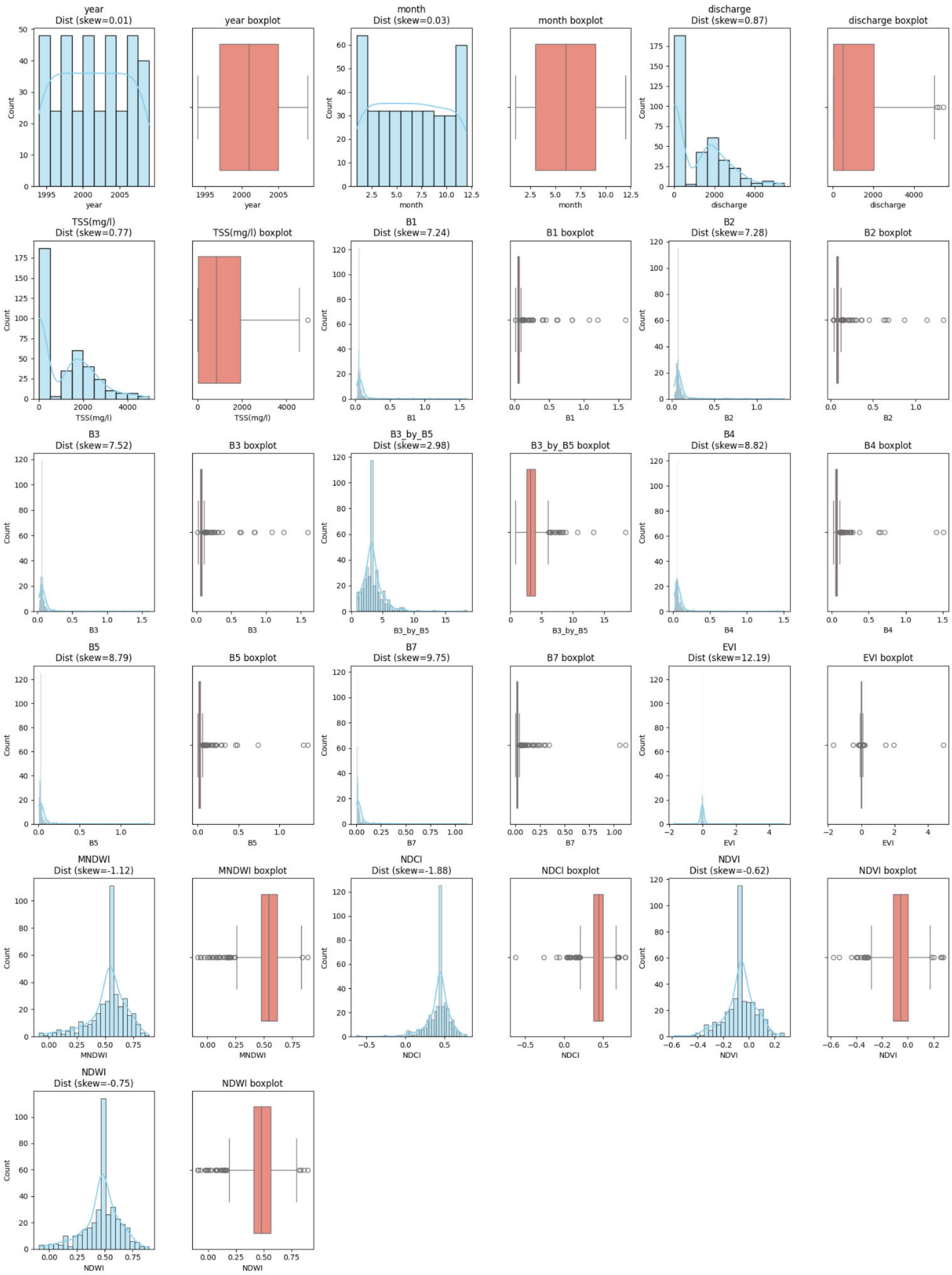

Figure 10.1: Original Distributions and Boxplots of All Variables (Before IQR and Log Transformations)

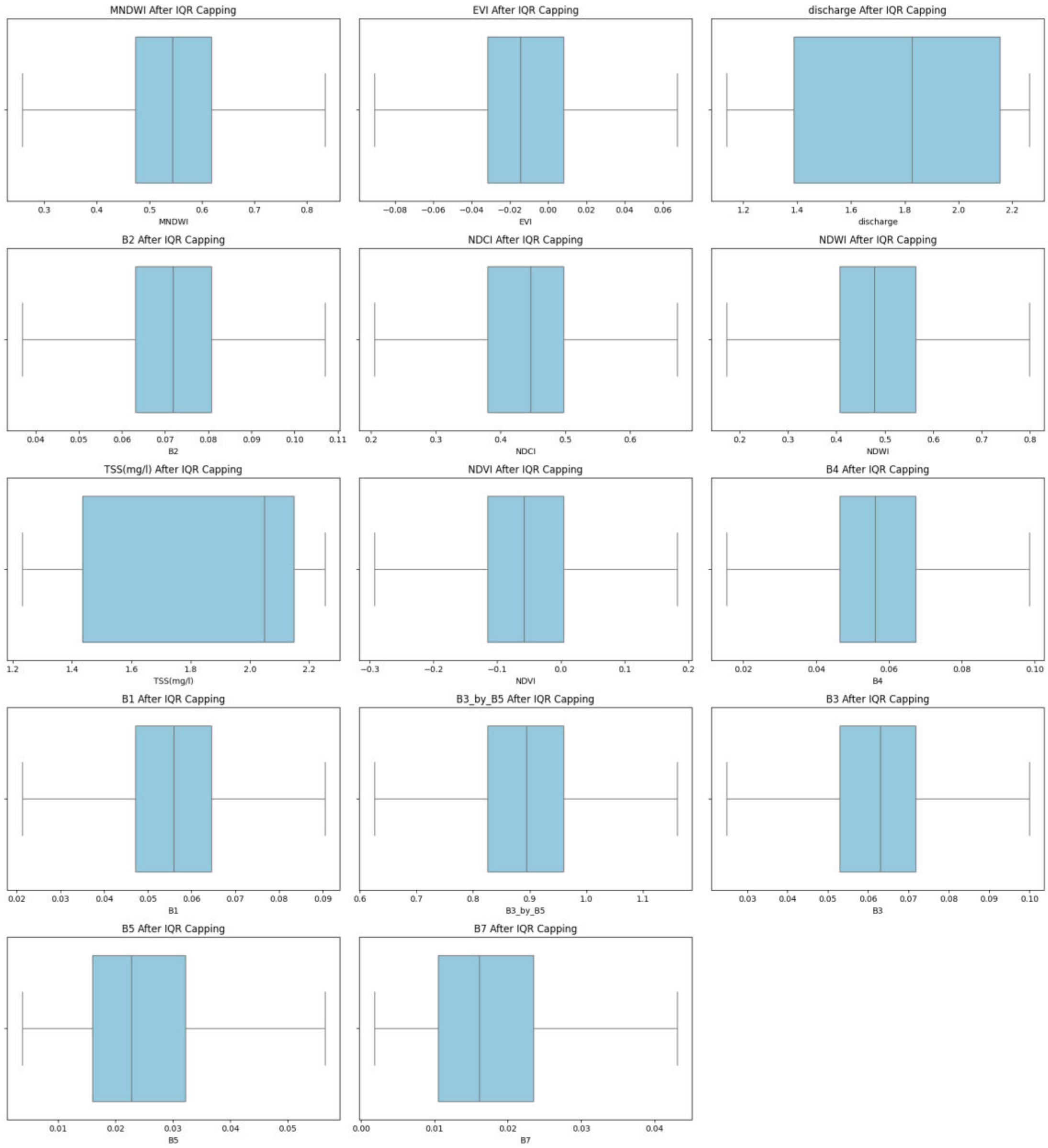

Figure 10.2: Boxplots After IQR Outlier Capping Showing Compressed Spread of Extremes

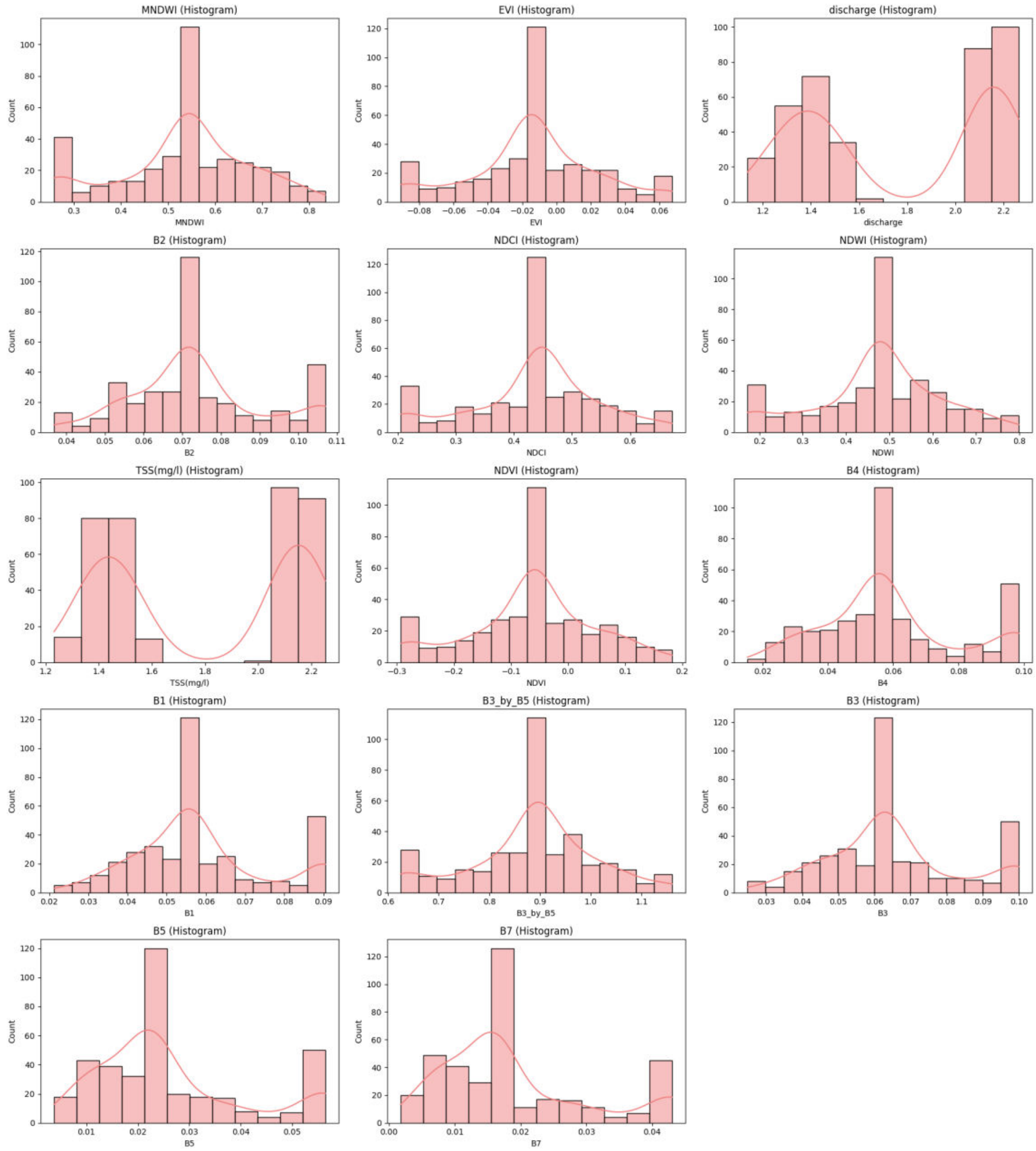

Figure 10.3: Histograms After IQR Outlier Capping for All Variables

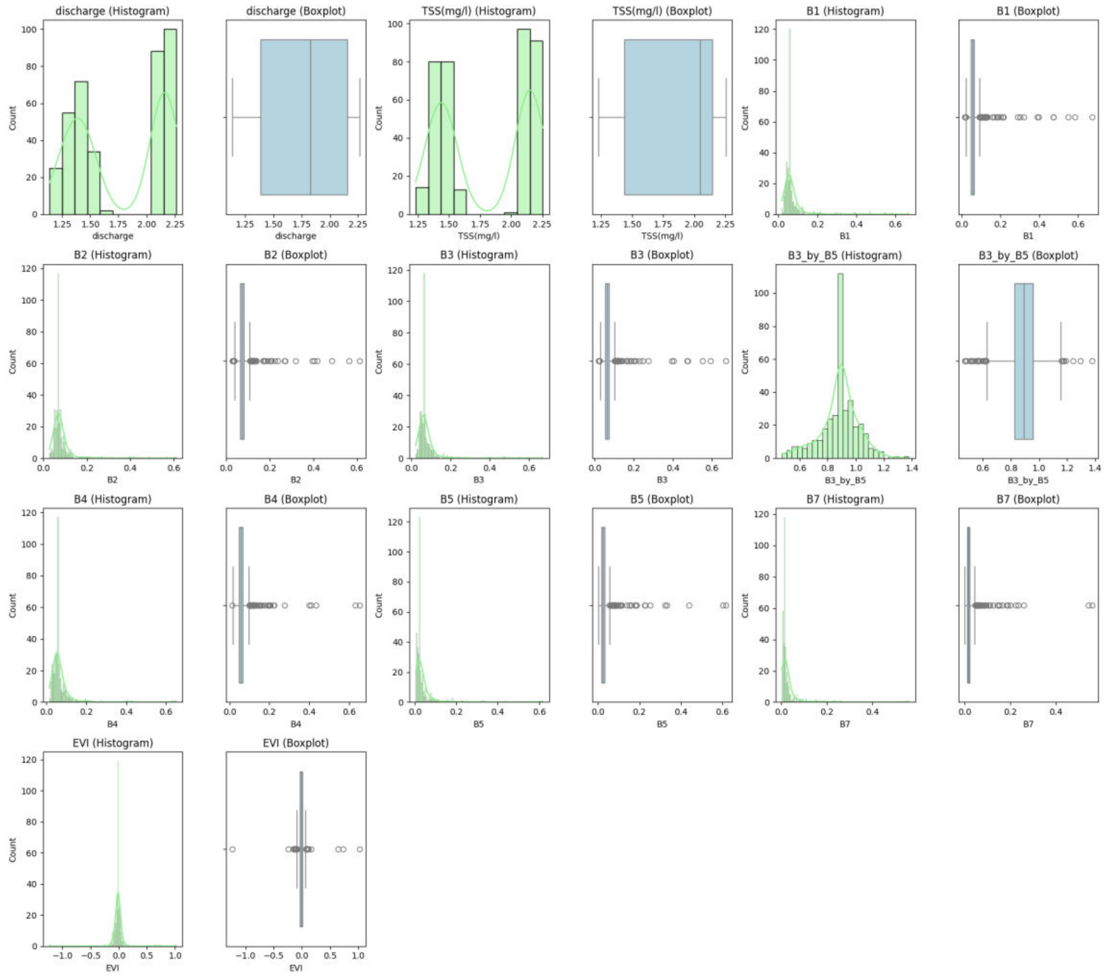

Figure 10.4: Histograms and Boxplots After Log Transformation Showing Symmetric Distributions

4.3 Feature Correlation Structure

A correlation heatmap was computed among all predictor variables to assess multicollinearity and interaction patterns. The heatmap reveals strong positive correlations between bands within similar spectral ranges (e.g., B3 and B5), and also high inverse correlation between chlorophyll indices (NDCI) and NDWI. These interdependencies inform feature importance and model structure decisions.

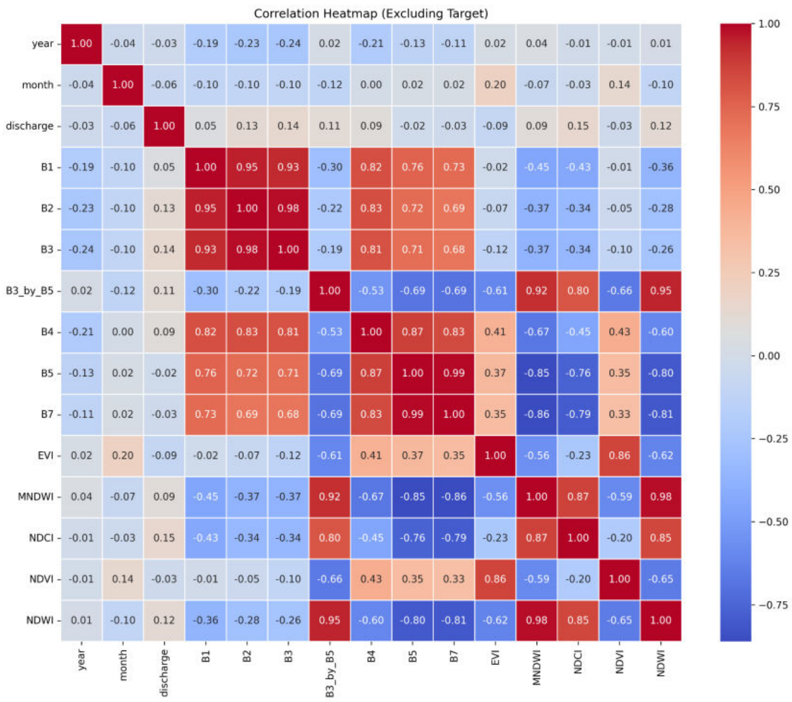

Figure 10.5: Pearson Correlation Heatmap of Predictor Variables (Excluding TSS)

4.4 Cross-Validation Performance

A 10-fold cross-validation was employed to evaluate the generalization ability of the models. The following radar chart compares the normalized performance of XGBoost and Random Forest across R^2, MSE, RMSE, and MAE metrics.

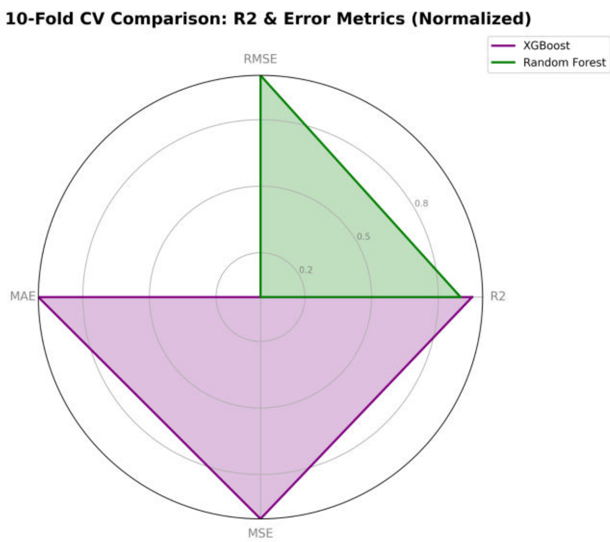

Figure 10.6: Normalized Radar Plot Comparing XGBoost and Random Forest Performance Metrics

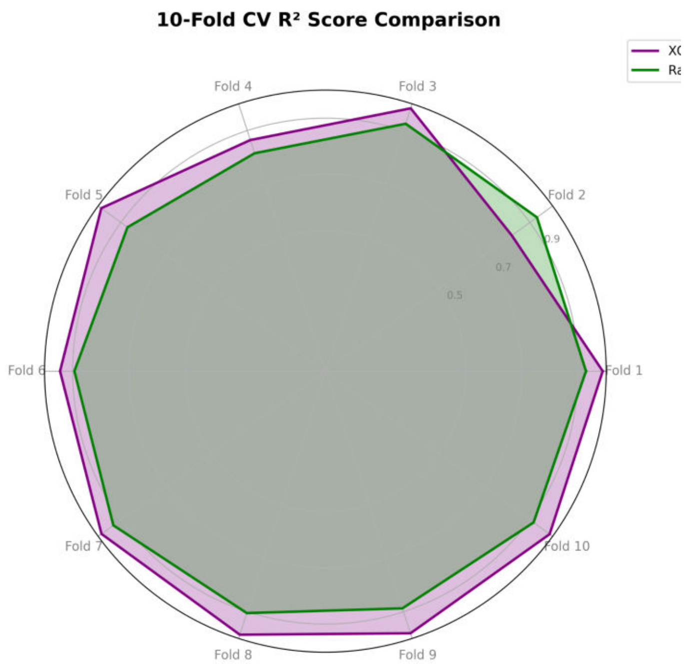

Figure 10.7: 10-Fold Cross-Validation R² Score Comparison Across Folds for XGBoost and Random Forest

4.5 Individual Model Performance and Predictive Accuracy

Each machine learning model was trained and evaluated using both training and test datasets. In addition to standard evaluation metrics, actual vs. predicted scatterplots were produced to visually assess model accuracy. These plots help reveal how closely predictions align with the ground truth and whether systematic under- or over-estimation occurs. Linear Regression showed moderate predictive power and serves as the baseline model.

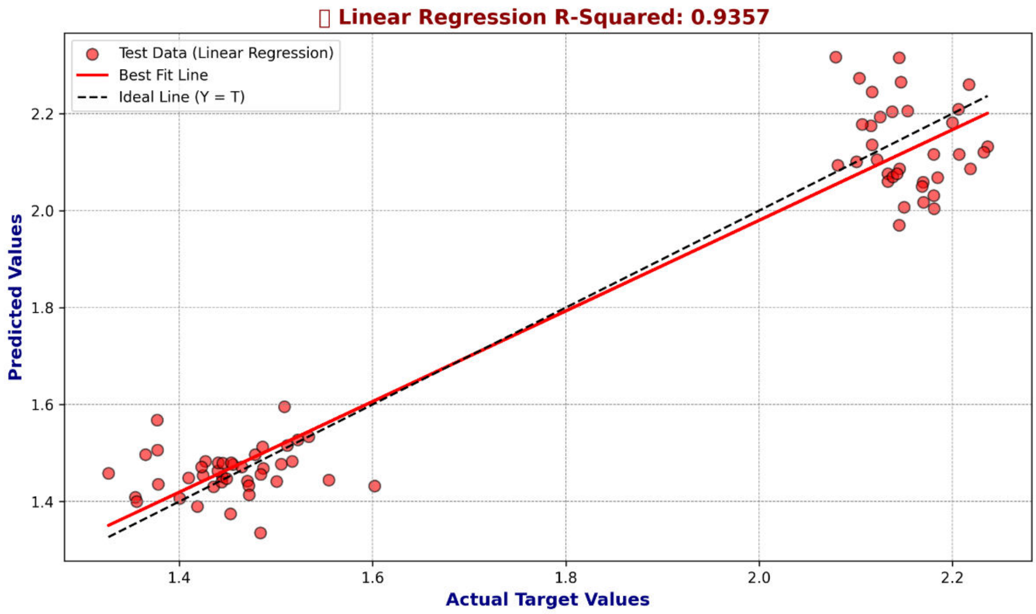

Figure 10.8: Actual vs. Predicted TSS Using Linear Regression

Random Forest performed significantly better, capturing non-linear relationships with a higher R^2 and lower RMSE.

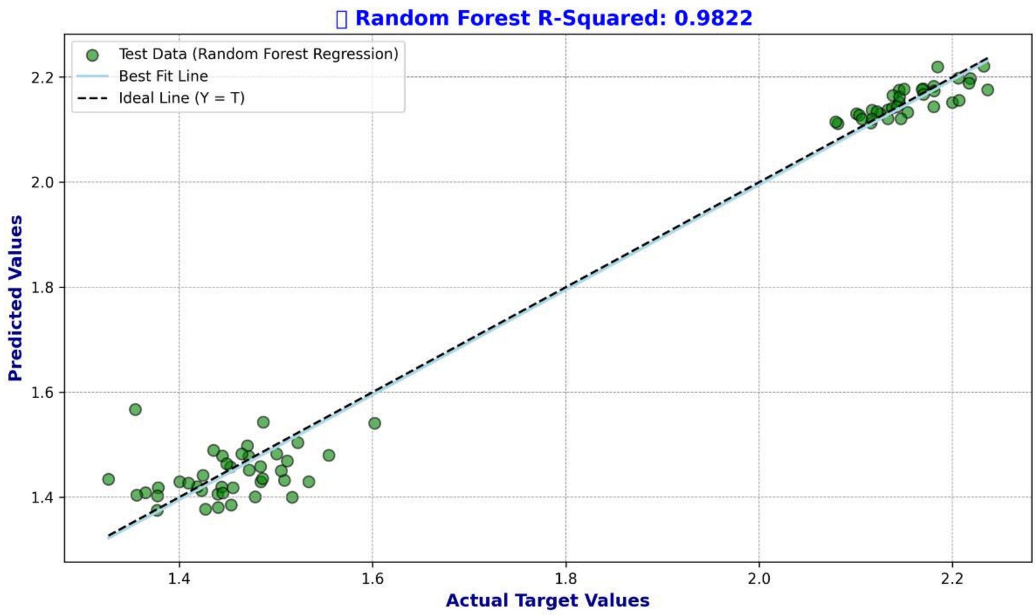

Figure 10.9: Actual vs. Predicted TSS Using Random Forest Regressor

XGBoost displayed excellent alignment between predictions and true values, suggesting minimal overfitting and robust generalization.

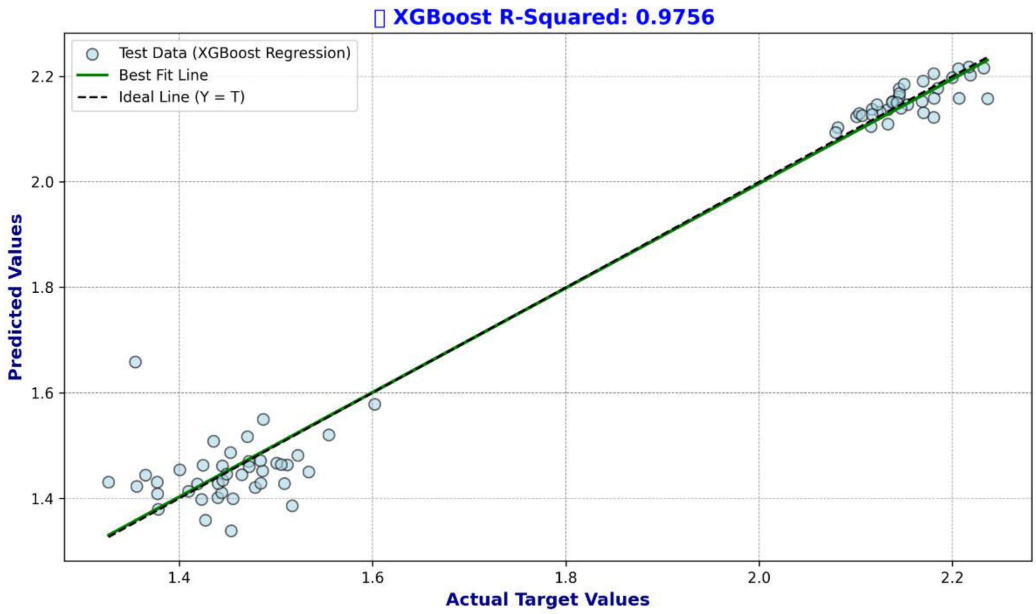

Figure 10.10: Actual vs. Predicted TSS Using XGBoost Regressor

CatBoost performed reasonably well but showed slightly less generalization on the test set than XGBoost.

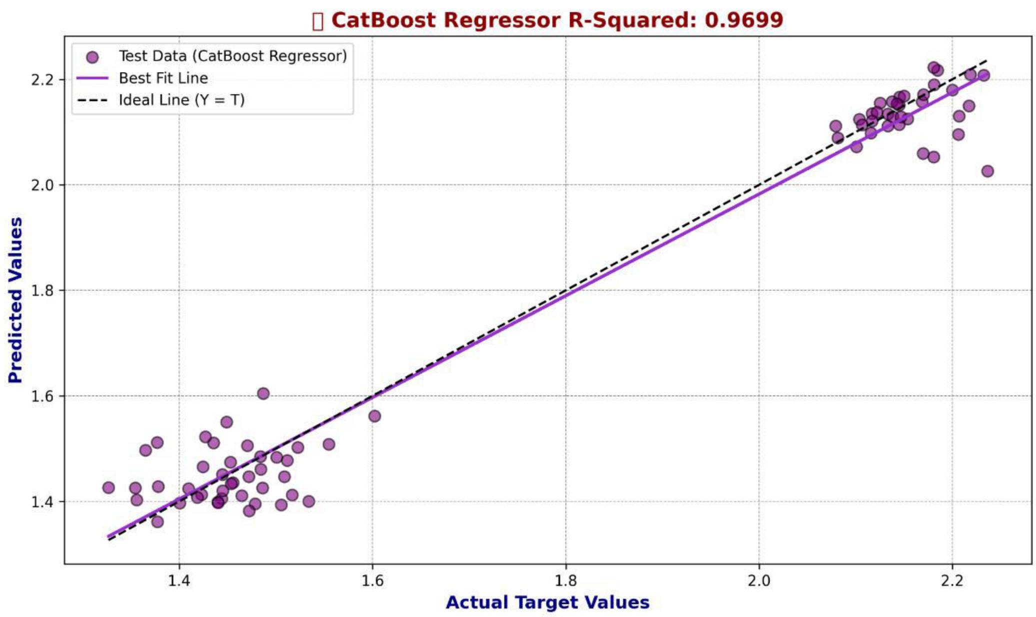

Figure 10.11: Actual vs. Predicted TSS Using CatBoost Regressor

Bagging Regressor produced consistent predictions with moderate spread.

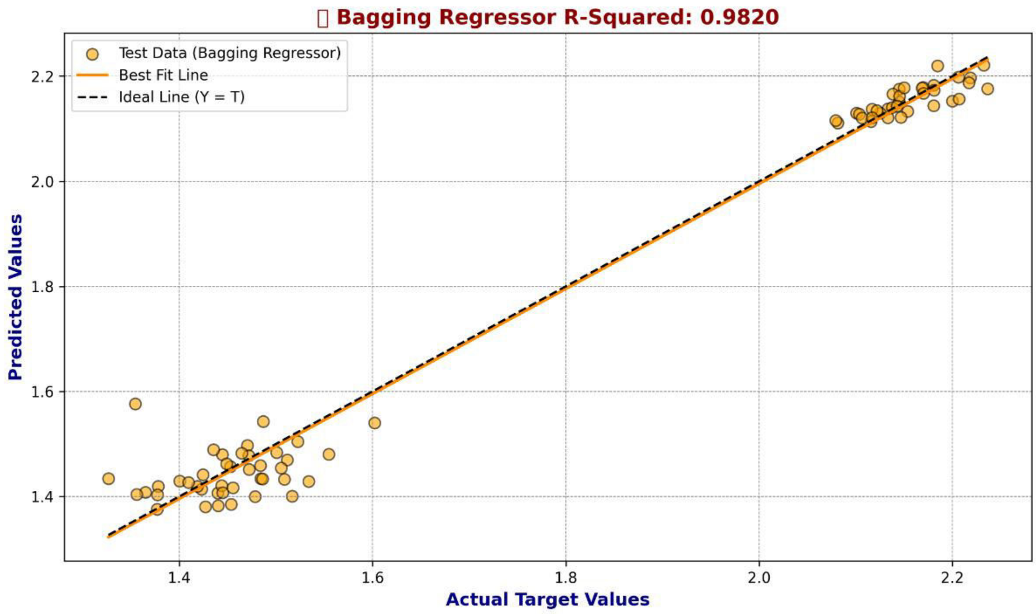

Figure 10.12: Actual vs. Predicted TSS Using Bagging Regressor

Gradient Boosting yielded one of the best fits, closely following the 1:1 line with minimal deviation.

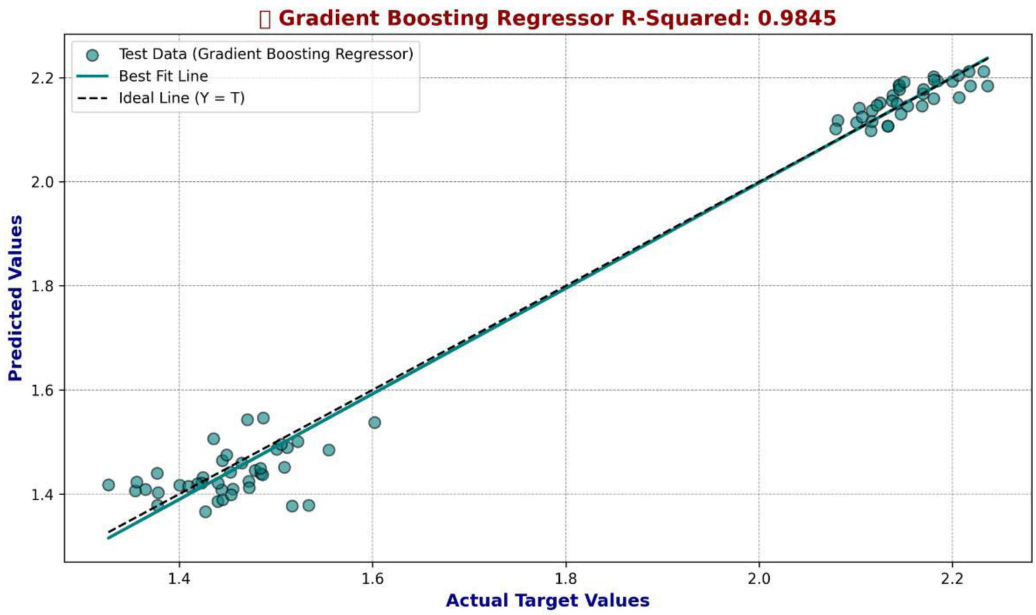

Figure 10.13: Actual vs. Predicted TSS Using Gradient Boosting Regressor

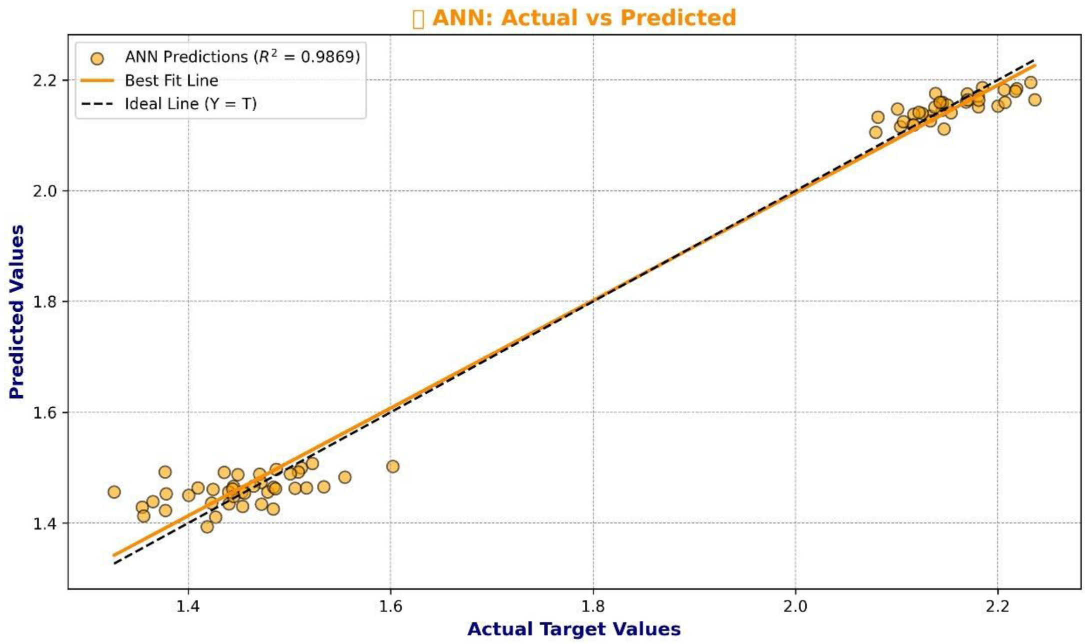

Figure 10.14: Actual vs. Predicted TSS Using Artificial Neural Network

4.6 Final Comparison of Level-2 Models

The best-performing models—Random Forest and XGBoost—were further optimized through hyperparameter tuning (Level-2 optimization). A side-by-side comparison of both models' final predictions and error metrics is shown below.

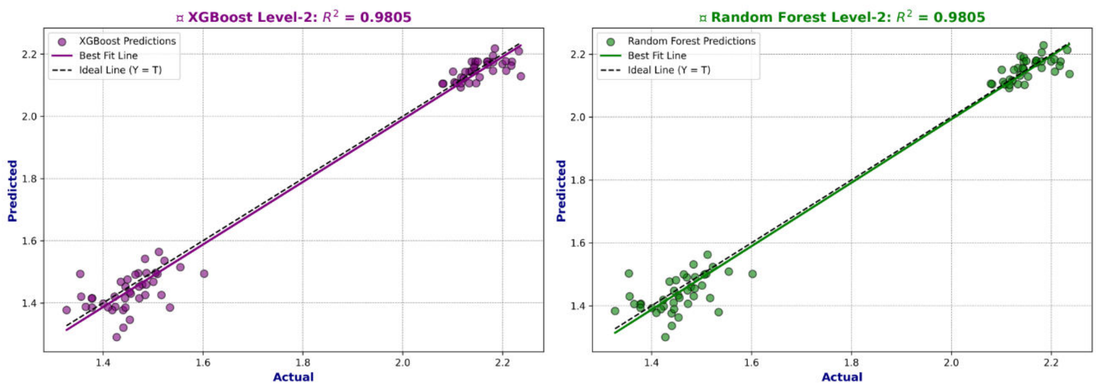

Figure 10.15: Comparative Evaluation of Final Level-2 Random Forest and XGBoost Models

4.7 Discussion of Model performances

The evaluation of machine learning models for predicting Total Suspended Solids (TSS) concentrations in the Rhine River, utilizing satellite-derived indices and discharge data, provides critical insights into their predictive accuracy and applicability for water quality monitoring. The performance metrics across Linear Regression, Random Forest, XGBoost, Bagging Regressor, Gradient Boosting, CatBoost, and the Artificial Neural Network (ANN) reveal distinct patterns in model efficacy, with implications for both geographical and anthropogenic influences on TSS variability.

Linear Regression, employed as a baseline model, exhibited a test set R^2 of 0.9357, with an MSE of 0.0080, RMSE of 0.0896, and MAE of 0.0695. These metrics indicate a reasonable fit for linear relationships within the dataset, though its performance is limited by its inability to capture the non-linear dynamics inherent in TSS concentrations influenced by industrial runoff in Duisburg and agricultural runoff in Bonn. The training set performance (R^2 0.9237, RMSE 0.1008) suggests a slight overfitting, further highlighting its inadequacy for complex environmental datasets.

Ensemble methods demonstrated superior predictive power, leveraging their capacity to model non-linear interactions. Random Forest achieved a test set R^2 of 0.9823, MSE of 0.0022, RMSE of 0.0470, and MAE of 0.0337, reflecting a robust generalization across the training (R^2 0.9957, RMSE 0.0238) and test sets. This performance underscores its effectiveness in handling the diverse feature set, including discharge and spectral indices like NDVI and EVI. XGBoost further excelled, with a test set R^2 of 0.9756, MSE of 0.0030, RMSE of 0.0551, and MAE of 0.0364, despite a near-perfect training performance (R^2 0.99998, RMSE 0.0015), suggesting potential overfitting mitigated by strong test set results. The high training accuracy of XGBoost may reflect its sensitivity to hyperparameter tuning, which was optimized in subsequent levels.

Gradient Boosting emerged as a top performer, achieving a test set R^2 of 0.9845, MSE of 0.0019, RMSE of 0.0439, and MAE of 0.0336, with an exceptionally low training MSE (8.53e-07) and RMSE (0.0009). This indicates a near-perfect fit to the training data, with effective generalization, likely due to its sequential error correction mechanism. Bagging Regressor, with a test set R^2 of 0.9820, MSE of 0.0022, RMSE of 0.0473, and MAE of 0.0336, closely mirrored Random Forest's performance, affirming the stability of ensemble averaging. CatBoost, while competitive with a test set R^2 of 0.9699, MSE of 0.0038, RMSE of 0.0613, and MAE of 0.0447, showed a slight decline in generalization compared to Gradient Boosting and XGBoost, possibly due to its handling of categorical features not fully leveraged in this dataset.

The introduction of the ANN added a deep learning dimension, achieving a test set R^2 of 0.9863, MSE of 0.0017, RMSE of 0.0414, and MAE of 0.0310. This performance rivals Gradient Boosting, with the ANN's lower RMSE and higher R^2 suggesting a strong capacity to model complex non-linear patterns in the data, likely enhanced by its ability to learn hierarchical feature representations. The slight increase in MSE compared to Gradient Boosting may indicate sensitivity to training data noise or network architecture complexity, warranting further optimization.

Hyperparameter tuning (Level-2 optimization) refined XGBoost and Random Forest, with XGBoost achieving a test set R^2 of 0.98713, MSE of 0.0016, RMSE of 0.0401, and MAE of 0.0297, outperforming Random Forest's test set R^2 of 0.9356, MSE of 0.0080, RMSE of 0.0896, and MAE of 0.0717. The final Level-2 evaluation showed both models converging to an R^2 of 0.9805, with XGBoost maintaining a marginal advantage in MAE (0.0363 vs. 0.0370). The 10-fold cross-validation further validated XGBoost's consistency (mean R^2 0.9526, std dev 0.0571) over Random Forest (mean R^2 0.9007, std dev 0.0347), suggesting greater robustness across diverse data subsets.

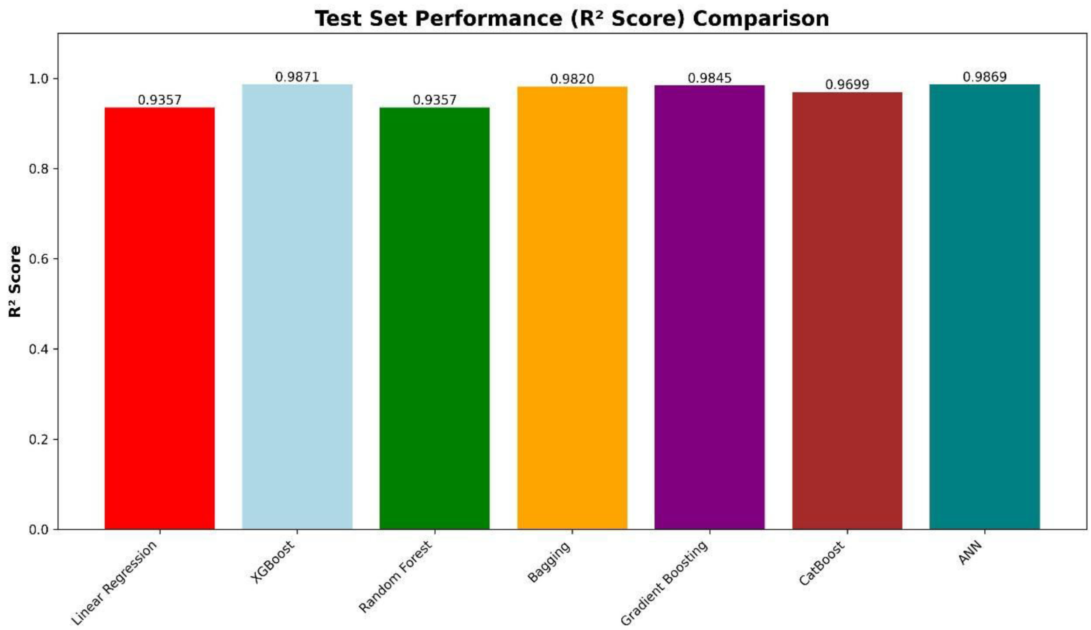

Figure 10.16: Comparative Evaluation of Different ML Models

Geographical and anthropogenic factors significantly influenced model performance. In Duisburg, industrial runoff, particularly during high discharge events, contributed to elevated TSS levels, with discharge and water quality indices (e.g., MNDWI) proving critical predictors. In Bonn, agricultural runoff and seasonal sediment transport, modulated by vegetation indices (EVI, NDVI), drove TSS variability, aligning with the models' reliance on these features. The superior performance of ensemble models and the ANN over Linear Regression highlights the necessity of non-linear modeling for capturing these complex interactions.

The study's reliance on Landsat 5 and 7 data from 1994–2009, combined with discharge data, underscores the potential of remote sensing for scalable, real-time water quality monitoring. The high R^2 values and low error metrics across models suggest that satellite-derived indices and discharge data are reliable proxies for TSS, offering a cost-effective alternative to traditional sampling. Future research could enhance these findings by incorporating higher-resolution datasets (e.g., Sentinel-2) and real-time streamflow data, potentially improving predictive accuracy during dynamic events like floods.

CHAPTER 5
CONCLUSION AND FUTURE WORK

This chapter provides a comprehensive analysis of the results presented in the previous sections. It interprets the model outputs in the context of the study's objectives and theoretical foundations, emphasizing how each modeling decision and data characteristic influenced the final outcomes. The broader implications of these findings for environmental monitoring and predictive hydrology are discussed in detail.

5.1. Performance of Machine Learning Models

The suite of models evaluated in this study demonstrated varying levels of predictive accuracy. XGBoost emerged as the most accurate and reliable model, achieving the highest R^2 values on both training and test sets. Its ability to handle non-linear interactions, missing data, and correlated predictors contributed significantly to its superior performance. The Level-2 hyperparameter tuning further refined this performance, bringing down the prediction error without overfitting—a critical aspect when modeling environmental variables that are often influenced by diverse and interacting factors.

Random Forest and Gradient Boosting models also performed well, particularly in capturing trends in the test dataset. Their ensemble nature allows for capturing complex interactions, and their performance across metrics was consistent. However, the hypertuned Random Forest showed some signs of overfitting as it scored lower on test set R^2 compared to XGBoost, indicating sensitivity to hyperparameter optimization and perhaps some instability in generalizing beyond training data.

Linear Regression, while less accurate in overall performance, provided an essential baseline. Its assumptions of linearity and independence were somewhat restrictive in the presence of non-linear spectral interactions, but the model still achieved a decent R^2, suggesting that some aspects of TSS variation are explainable using simple, interpretable relationships.

5.2. Spectral and Environmental Insights

The success of models in predicting TSS strongly depended on the relevance and engineering of input features. Spectral indices such as NDVI, EVI, and NDCI played a central role, as they reflect vegetative health and water content—both of which have indirect but measurable effects on sediment levels. High vegetation index values typically correlate with lower erosion, while chlorophyll-based indices relate to biological productivity, which can affect organic particulate concentration in water bodies.

The inclusion of discharge as a feature provided an essential hydrological perspective. Discharge levels influence sediment transport and TSS directly, and its integration into the model contributed to improved performance across nearly all algorithms. The correlation heatmap validated these relationships, showing meaningful interactions between discharge and certain spectral bands.

Additionally, water clarity indices like NDWI and MNDWI contributed to distinguishing between clean and sediment-heavy water bodies. Their strong correlation with TSS, observed both statistically and in model feature importance, reinforced the decision to include them in the final modeling pipeline.

5.3. Role of Data Transformation and Feature Preprocessing

Preprocessing steps like log transformation and IQR-based outlier capping proved essential in enhancing model reliability. These steps addressed distributional asymmetries and reduced the impact of skewed or extreme values that could otherwise bias model training. Visualizations before and after transformation confirmed a significant reduction in skewness, which aligns with improved learning rates and convergence stability in tree-based and neural models.

The use of monthly mean aggregation, while reducing temporal noise, also helped maintain a manageable modeling scale. However, this introduced a trade-off by limiting the granularity of prediction, potentially masking short-term variability. Still, for medium-scale environmental monitoring, the monthly resolution struck a balance between temporal smoothness and model interpretability.

5.4. Limitations

Although the research framework produced accurate and interpretable results, several limitations are inherent in the data, methodology, and model assumptions. First, the 30-meter spatial resolution of Landsat imagery restricts the detection of smaller-scale heterogeneities, such as narrow tributaries or localized pollution sources. While suitable for regional-scale monitoring, it may not detect microenvironmental anomalies.

Cloud masking techniques were applied to Landsat images to reduce noise, yet some artifacts or missing values remain inevitable due to sensor limitations and atmospheric interference. These may affect the quality of monthly composites, especially in seasons with persistent cloud cover.

The modeling framework assumes that monthly TSS levels are driven primarily by variables like discharge and surface reflectance. However, this excludes other potentially influential variables such as sediment load measurements, upstream construction, rainfall intensity, or anthropogenic land-use practices. The absence of such data could limit the models' explanatory depth, especially during peak flow or unusual events.

Lastly, the study is geographically constrained to two sites along the Rhine River. While this ensures focused calibration, it also limits model generalizability across other river systems without retraining or adjusting feature interpretations.

5.5. Scope for Future Research

Several extensions can be made to expand the current research. Future studies could utilize higher-resolution datasets like Sentinel-2 MSI, which offers improved spatial and temporal granularity. These images could provide finer distinctions in vegetative cover, sediment plumes, or water quality zones, leading to better model sensitivity to micro-scale changes.

Temporal forecasting models could also be integrated. Recurrent neural networks (RNNs), long short-term memory (LSTM) networks, or even transformer-based temporal models could better capture time-lagged effects of rainfall, seasonal vegetation changes, or industrial discharge. These models could add value to operational forecasting efforts.

Moreover, future modeling efforts can expand to multi-source data fusion—combining optical and radar datasets, integrating meteorological data, and assimilating government-reported pollution inventories. This would provide a richer feature space, enhancing both accuracy and interpretability.

Transfer learning should also be explored. Once a robust TSS prediction model is developed for a major river like the Rhine, adapting the model to similar European rivers with minimal re-training could significantly enhance the scalability and impact of the method.

This study presented an end-to-end machine learning pipeline to predict Total Suspended Solids in the Rhine River, using a combination of Landsat-derived spectral indices and river discharge data. The methodology included rigorous preprocessing, extensive feature engineering, multiple algorithm comparisons, and performance validation through cross-validation and visual assessment.

The XGBoost model, especially after Level-2 hyperparameter tuning, emerged as the most suitable predictive engine, outperforming all others in both accuracy and stability. The use of monthly composite datasets, cloud filtering, and reflectance correction provided a clean foundation for modeling, while environmental indices such as NDVI, MNDWI, and NDCI demonstrated strong correlations with TSS.

In addition to its predictive achievements, the research contributes to the field by validating a transferable approach for monitoring river health using remotely sensed data. This framework can be applied to other rivers or pollutants with appropriate feature adjustments, paving the way for more data-driven environmental management tools. Through its comprehensive methodology and transparent evaluation, this work underscores the potential of AI in hydrological science and policy support.

REFERENCES

[1] G. Rodríguez-Martínez *et al.*, "Water erosion, its relationship to total suspended solids and water quality in the lower basin of the Usumacinta River, Tabasco, Mexico," *Nature Environment and Pollution Technology*, vol. 20, no. 3, pp. 1097–1106, 2021.

[2] L. Rossi, R. Fankhauser, and N. Chevre, "Water quality criteria for total suspended solids (TSS) in urban wet-weather discharges," *Water Science and Technology*, vol. 54, no. 6–7, pp. 355–362, 2006.

[3] S. A. Aliyan, A. J. Astari, P. A. Ayesha, N. S. Amelia, N. Ibrahim, and Z. A. G. Kusmana, "Estimating Concentration of Suspended Solids Load in Estuary System Utilizing Sentinel 2A: A Case Study of Cipatujah Beach, Indonesia," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, 2024, p. 012040.

[4] J. A. Ortega-Becerril, C. Suarez, D. Vázquez-Tarrío, J. Garrote, and M. Gomez-Heras, "Sediment Response after Wildfires in Mountain Streams and Their Effects on Cultural Heritage: The Case of the 2021 Navalacruz Wildfire (Avila, Spain)," *Fire*, vol. 7, no. 2, p. 52, 2024.

[5] N. Na'imah, D. Taryana, and P. S. Wiyana, "Mapping the Distribution of Total Suspended Solids (TSS) in Gondang Reservoir, Lamongan Using Multi-Temporal Landsat Imagery," *Future Space: Studies in Geo-Education*, vol. 1, no. 3, pp. 286–305, 2024.

[6] G. E. Adjovu, H. Stephen, D. James, and S. Ahmad, "Measurement of total dissolved solids and total suspended solids in water systems: A review of the issues, conventional, and remote sensing techniques," *Remote Sens (Basel)*, vol. 15, no. 14, p. 3534, 2023.

[7] S. Hafeez, M. S. Wong, S. Abbas, and M. Asim, "Evaluating landsat-8 and sentinel-2 data consistency for high spatiotemporal inland and coastal water quality monitoring," *Remote Sens (Basel)*, vol. 14, no. 13, p. 3155, 2022.

[8] P. Kowe *et al.*, "Spatial-temporal variability analysis of water quality using remote sensing data: A case study of Lake Manyame," *Sci Afr*, vol. 21, p. e01877, 2023.

[9] R. S. Dewi, P. Hartanto, N. Oktaviani, I. Pujawati, N. Nursugi, and S. Aditya, "Satellite-derived bathymetry to improve bathymetric map of Indonesia," in *Sixth International Symposium on LAPAN-IPB Satellite*, SPIE, 2019, pp. 545–554.

[10] H. Lin, Q. Cheng, W. Sun, F. Yang, Y. Ding, and J. Ma, "Copper exposure effects on antibiotic degradation in swine manure vary between mesophilic and thermophilic conditions," *Science of the Total Environment*, vol. 841, p. 156759, 2022.

[11] Y. Niu *et al.*, "Regulation of alpine meadow patch coverage on runoff and sediment under natural rainfall on the eastern Qinghai-Tibetan Plateau," *J Hydrol (Amst)*, vol. 603, p. 127101, 2021.

[12] B. H. Z. Sami, B. F. Z. Sami, C. M. Fai, Y. Essam, A. N. Ahmed, and A. El-Shafie, "Investigating the reliability of machine learning algorithms as a sustainable tool for total suspended solid prediction," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1607–1622, 2021.

[13] C. Wang, S. Chen, D. Li, D. Wang, W. Liu, and J. Yang, “A Landsat-based model for retrieving total suspended solids concentration of estuaries and coasts in China,” *Geosci Model Dev*, vol. 10, no. 12, pp. 4347–4365, 2017.

[14] L. Silveira Kupssinskü *et al.*, “A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning,” *Sensors*, vol. 20, no. 7, p. 2125, 2020.

[15] N. M. DeLuca, B. F. Zaitchik, and F. C. Curriero, “Can multispectral information improve remotely sensed estimates of total suspended solids? A statistical study in Chesapeake Bay,” *Remote Sens (Basel)*, vol. 10, no. 9, p. 1393, 2018.

[16] J. Moon, S. Suh, S. Jung, S.-S. Baek, and J. Pyo, “Deep learning-based mapping of total suspended solids in rivers across South Korea using high resolution satellite imagery,” *GIsci Remote Sens*, vol. 61, no. 1, p. 2393489, 2024.

[17] S. H. Rahat *et al.*, “Remote sensing-enabled machine learning for river water quality modeling under multidimensional uncertainty,” *Science of The Total Environment*, vol. 898, p. 165504, 2023.

[18] C. M. Isidro, N. McIntyre, A. M. Lechner, and I. Callow, “Quantifying suspended solids in small rivers using satellite data,” *Science of the Total Environment*, vol. 634, pp. 1554–1562, 2018.

[19] J. Zhao *et al.*, “Remote sensing evaluation of total suspended solids dynamic with Markov model: A case study of inland reservoir across administrative boundary in South China,” *Sensors*, vol. 20, no. 23, p. 6911, 2020.

[20] X. Zhang, J. Huang, J. Chen, and Y. Zhao, “Remote sensing monitoring of total suspended solids concentration in Jiaozhou Bay based on multi-source data,” *Ecol Indic*, vol. 154, p. 110513, 2023.

[21] E. Friedmann, C. Gleason, D. Feng, and T. Langhorst, “Estimating Riverine Total Suspended Solids From Spatiotemporal Satellite Sensor Fusion,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, 2024.

[22] R. L. S. Dias *et al.*, “Machine learning models applied to TSS estimation in a reservoir using multispectral sensor onboard to RPA,” *Ecol Inform*, vol. 65, p. 101414, 2021.

[23] L. Prathumratana, S. Sthiannopkao, and K. W. Kim, “The relationship of climatic and hydrological parameters to surface water quality in the lower Mekong River,” *Environ Int*, vol. 34, no. 6, pp. 860–866, 2008.

[24] H. Virro, G. Amatulli, A. Kmoch, L. Shen, and E. Uuemaa, “GRQA: global river water quality archive,” *Earth System Science Data Discussions*, vol. 2021, pp. 1–30, 2021.

[25] N. A. Misman, M. F. Sharif, A. J. K. Chowdhury, and N. H. Azizan, “Water pollution and the assessment of water quality parameters: a review,” *Desalination Water Treat*, vol. 294, pp. 79–88, 2023.

[26] E. V Taguas *et al.*, “Opportunities of super high-density olive orchard to improve soil quality: Management guidelines for application of pruning residues,” *J Environ Manage*, vol. 293, p. 112785, 2021.

[27] P. K. Singh *et al.*, “Critical review on toxic contaminants in surface water ecosystem: sources, monitoring, and its impact on human health,” *Environmental Science and Pollution Research*, vol. 31, no. 45, pp. 56428–56462, 2024.

[28] Y. Wang, H. Yin, Z. Liu, and X. Wang, “A systematic review of the scientific literature on pollutant removal from stormwater runoff from vacant urban lands. Sustainability 14 (19): 12906,” 2022.

[29] Y.-F. Su *et al.*, “A multivariate model for coastal water quality mapping using satellite remote sensing images,” *Sensors*, vol. 8, no. 10, pp. 6321–6339, 2008.

[30] S. Wen, Y. Zhao, S. Liu, Y. Chen, H. Yuan, and H. Xu, “Polystyrene microplastics exacerbated liver injury from cyclophosphamide in mice: Insight into gut microbiota,” *Science of the Total Environment*, vol. 840, p. 156668, 2022.

[31] M. J. Kang, V. Mesev, and S. W. Myint, "Urbanization and Quality of Stormwater Runoff: Remote Sensing Measurements of Land Cover in an Arid City," *Korean Journal of Remote Sensing*, vol. 30, no. 3, pp. 399–415, 2014.

[32] R. Yang *et al.*, "Response of soil erosion and non-point source pollution to different rainfall, vegetation and land preparation measures in Miyun reservoir area during 2010-2023," *International Soil and Water Conservation Research*, 2025.

[33] C. Lu, T. Li, T. Dehaene, and B. Lagaisse, "ICVul: A Well-labeled C/C++ Vulnerability Dataset with Comprehensive Metadata and VCCs," in *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*, IEEE, 2025, pp. 154–158.

[34] T.-M. Perivolioti *et al.*, "Monitoring the Impact of Floods on Water Quality Using Optical Remote Sensing Imagery: The Case of Lake Karla (Greece)," *Water (Basel)*, vol. 16, no. 23, p. 3502, 2024.

[35] D. Renner, H. Rügner, M. Ebner, J. Fabregat-Palau, and P. Grathwohl, "Particle-Associated Contaminant Transport in Rivers during High Discharge Events," in *EGU General Assembly Conference Abstracts*, 2024, p. 10544.

[36] M. S. Pleasants, F. dos A. Neves, A. D. Parsekian, K. M. Befus, and T. J. Kelleners, "Hydrogeophysical inversion of time-lapse ERT data to determine hillslope subsurface hydraulic properties," *Water Resour Res*, vol. 58, no. 4, p. e2021WR031073, 2022.

[37] S. Shoergashova, T. Liu, N. Ibrokhimova, L. Latisheva, and B. Karimov, "Assessment of temporal and spatial variations of water quality parameters in the Zarafshan River basin," *CLEAN–Soil, Air, Water*, vol. 52, no. 8, p. 2300454, 2024.

[38] Z. Hao *et al.*, "Remote sensing of river discharge from medium-resolution satellite imagery based on deep learning," *Water Resour Res*, vol. 60, no. 9, p. e2023WR036880, 2024.

[39] P. Burek and M. Smilovic, "The use of GRDC gauging stations for calibrating large-scale hydrological models," *Earth System Science Data Discussions*, vol. 2022, pp. 1–18, 2022.

[40] S. Vijayakumar, R. Saravanakumar, M. Arulanandam, and S. Ilakkiya, "Google Earth Engine: empowering developing countries with large-scale geospatial data analysis—a comprehensive review," *Arabian Journal of Geosciences*, vol. 17, no. 4, p. 139, 2024.

[41] M. D. Nguyen, O. M. Baez-Villanueva, D. D. Bui, P. T. Nguyen, and L. Ribbe, "Harmonization of landsat and sentinel 2 for crop monitoring in drought prone areas: Case studies of Ninh Thuan (Vietnam) and Bekaa (Lebanon)," *Remote Sens (Basel)*, vol. 12, no. 2, p. 281, 2020.

[42] M. Claverie, E. F. Vermote, B. Franch, and J. G. Masek, "Evaluation of the Landsat-5 TM and Landsat-7 ETM+ surface reflectance products," *Remote Sens Environ*, vol. 169, pp. 390–403, 2015.

[43] N. Roustaei, "Application and interpretation of linear-regression analysis," *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, vol. 13, no. 3, p. 151, 2024.

[44] N. E. I. Karabadji, A. A. Korba, A. Assi, H. Seridi, S. Aridhi, and W. Dhifli, "Accuracy and diversity-aware multi-objective approach for random forest construction," *Expert Syst Appl*, vol. 225, p. 120138, 2023.

[45] G. Grekousis, "Geographical-XGBoost: a new ensemble model for spatially local regression based on gradient-boosted trees," *J Geogr Syst*, pp. 1–27, 2025.

[46] L. W. Rizkallah, "Enhancing the performance of gradient boosting trees on regression problems," *J Big Data*, vol. 12, no. 1, p. 35, 2025.

[47] S. Rallapalli and Y. M. M. Kumar, "Optimizing Employee Promotion Decisions: A Novel Machine Learning Framework for Predictive Analysis by using GBM CatBoost," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, IEEE, 2024, pp. 1–7.

[48] C. Gershenson, “Artificial neural networks for beginners,” *arXiv preprint cs/0308031*, 2003.

[49] H. Liu and Z. Zhang, “Development and trending of deep learning methods for wind power predictions,” *Artif Intell Rev*, vol. 57, no. 5, p. 112, 2024.

[50] J. Qiu, “An analysis of model evaluation with cross-validation: techniques, applications, and recent advances,” *Adv. Econ. Manag. Political Sci*, vol. 99, pp. 69–72, 2024.

[51] B. Johnson, Y. Brun, and A. Meliou, “Causal testing: Understanding defects’ root causes,” in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020, pp. 87–99.