

Article

Machine Learning-Driven Prediction of Quality Water in the Ebro River from the Ascó Water Quality Monitoring Station Data, Northeast Spain

Paulino José García-Nieto ^{1,*}, Esperanza García-Gonzalo ¹ , José Ramón Alonso Fernández ², Cristina Díaz Muñiz ³ and Luis Alfonso Menéndez-García ¹ 

¹ Department of Mathematics, Faculty of Sciences, University of Oviedo, C/Leopoldo Calvo Sotelo, 18, 33007 Oviedo, Spain; espe@uniovi.es (E.G.-G.); menendezalfonso@uniovi.es (L.A.M.-G.)

² Cantabrian Basin Authority, Spanish Ministry for the Ecological Transition and Demographic Challenge, 33071 Oviedo, Spain; jralonsof@gmail.com

³ Asturias Water Consortium, C/Santa Susana 15, 33007 Oviedo, Spain; cdingva@gmail.com

* Correspondence: pjgarcia@uniovi.es; Tel.: +34-985103417

Abstract

Continuous monitoring of river water quality is essential for understanding hydro-chemical dynamics and supporting adaptive water resource management. This study develops a Gaussian Process Regression (GPR) model with a Matern32 kernel, optimized via the Nelder–Mead (NM) algorithm, to predict turbidity, electrical conductivity, and UV254 absorbance in the Ebro River (Spain), without considering temporal dependencies. The Matern 32 kernel was selected for its ability to model the non-smooth, heterogeneous variability of water quality parameters, and its hyperparameters were tailored to each indicator, revealing unique spatial scales of variability. Using 46,287 high-frequency observations, the model achieves Nash–Sutcliffe Efficiency (NSE) values of 0.8956, 0.9503, and 0.9180 for turbidity, conductivity, and absorbance, respectively. Feature importance analysis reveals flow rate and nitrates as dominant drivers. The GPR framework outperforms multilayer perceptron (MLP), multivariate linear regression (MLR), and M5 model tree (M5P), demonstrating its potential for real-time water quality monitoring.

Keywords: turbidity; conductivity and absorbance; Gaussian process regression (GPR); multilayer perceptron (MLP); multivariate linear regression; M5 model tree; river water quality



Academic Editor: Dibyendu Sarkar

Received: 26 May 2026

Revised: 25 June 2026

Accepted: 26 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Monitoring water quality is an essential component of environmental science [1,2]. This is crucial for sustainable development and an important indicator of the health of aquatic ecosystems [1,3]. However, problems with water pollution have become more severe due to the acceleration of global industrialization and urbanization [4], and the construction of dams and reservoirs has exacerbated the issue.

Understanding the dynamics of water quality in freshwater systems is essential for grasping ecosystem functioning and for supporting evidence-based decision-making in environmental management. Conventional monitoring programs rely on a suite of physicochemical indicators that reflect different aspects of water condition and contamination pathways. Turbidity, electrical conductivity, and ultraviolet absorbance at 254 nm (UV254) are widely recognized as sensitive indicators of suspended particulate matter, dissolved

ionic content, and dissolved organic matter, respectively [5–7]. Each parameter reflects different hydrological, geochemical, and biological processes within aquatic environments. Their combined analysis provides a comprehensive view of the ecosystem's status and potential pollution sources.

Globally, public agencies are increasingly concerned about water quality. In response, the United Nations (UN) introduced the 2030 Agenda in 2015, aiming to advance a more resilient and sustainable world. Goal 6 of the 17 Sustainable Development Goals (SDGs) focuses on ensuring everyone has access to clean water and sanitation and can sustainably manage these resources. Economic activity, human health, and the preservation of ecosystem services and biodiversity depend on the quality of surface water [8,9]. Increased mortality rates have been linked to pollution and water scarcity, particularly among vulnerable populations, such as children and newborns [10].

Discharge is a primary driver of sediment mobilization and solute transport [11]. Nutrients such as nitrates and ammonium influence primary productivity and redox processes [12], while pH, dissolved oxygen, and temperature modulate chemical equilibria and biological activity [13]. Trace metals, including mercury, can interact with particulate and dissolved fractions; however, their influence on bulk optical and conductivity metrics is often limited.

Developing predictive models for these parameters based on routinely monitored environmental variables, such as streamflow, ammonium, dissolved oxygen, pH, water temperature, nitrates, and dissolved mercury, offers a powerful approach for elucidating the processes that control spatiotemporal variability in water quality [14,15].

In this study, we constructed three independent statistical models, each of which targeted one of the aforementioned water quality indicators as the response variable. By analyzing turbidity, conductivity, and UV254 separately, we aim to quantify how each indicator responds to the same set of environmental predictors, thereby revealing common and distinct drivers of water quality dynamics. This approach allows us to identify dominant controlling factors, potential threshold behaviors, and covariate interactions that may otherwise remain obscured in multivariate observational datasets.

The development of such models has practical implications beyond scientific understanding. Predictive relationships between environmental variables and key water quality indicators can support real-time forecasting systems, enabling the early detection of contamination events, sediment mobilization episodes, and shifts in the characteristics of dissolved organic matter. Furthermore, these models can inform adaptive management strategies, guide the design of monitoring networks, and improve water authorities' ability to anticipate the effects of climate variability and human activities on freshwater ecosystems [16]. Integrating predictive modeling with routine environmental monitoring is a valuable step toward more resilient and data-driven water resource management.

Partial differential equations (PDEs) are fundamental to modeling various physical phenomena, such as water quality in bodies of water. They are also the heart of many engineering and science applications. However, solving these equations is challenging, especially as the complexity of the problems increases. Traditionally, classical numerical methods based on well-established physical and mathematical principles have been used [17,18]. While these methods are robust, they have notable limitations in high-dimensional scenarios or with complex geometries. Furthermore, water quality assessment based on Laplacian mathematical models requires extensive datasets that are often difficult to obtain or unavailable, as these models depend on the internal physicochemical processes occurring in rivers, reservoirs, and lakes. The computational implementation of these models is complex and demands substantial processing time. Conversely, the advent of

machine learning (ML) techniques, such as the Gaussian Process Regression (GPR), has opened new avenues for studying water quality in rivers, reservoirs, and lakes.

Regression problems can be effectively addressed using the GPR framework [19–27], a supervised learning methodology characterized by robustness and the ability to model complex nonlinear relationships. Based on previous studies, GPR has demonstrated broad applicability across diverse scientific and engineering domains, including lithium-ion battery hybrid modeling [28], an effective way to forecast wind loads on cargo ships [29], and springback estimation in incremental sheet metal forming [30]. In particular, we can mention two significant earlier studies using the GPR approach in hydrology: the first successfully predicted soil moisture [31], and the second quickly anticipated the maximum area and depth of flood inundation [32]. Several arguments can be made to support the efficacy and applicability of the proposed GPR methodology [19–27]. (1) The model's cost function is computed exclusively from the training dataset, so the GPR methodology only requires a subset of the complete data for effective model development. (2) GPR offers a complete distribution of probable values and a well-founded confidence interval for each prediction. (3) The GPR technique adapts well to small- to medium-sized datasets, leveraging priors and kernels to model complex relationships without requiring many examples. (4) GPR captures structures such as periodicity, smoothness, and linear trends, allowing it to model complex functions without increasing complexity. (5) GPR is a nonparametric method, meaning that the model's complexity automatically adjusts to the amount of data, unlike linear or polynomial models. (6) Regularization arises naturally from the probabilistic definition, helping avoid overfitting. (7) GPR is supported by a well-defined probabilistic framework that facilitates integration with other Bayesian models. (8) GPR methodology exhibits robustness to outliers when hyperparameter optimization is appropriately performed. Consequently, its predictions are minimally affected by anomalous data points, resulting in high accuracy and reliability.

The objective of this study is to develop and validate a Gaussian Process Regression (GPR) model, optimized via the Nelder–Mead algorithm [33–37], for predicting turbidity, conductivity, and UV254 absorbance in the Ebro River using high-frequency physicochemical data. Specifically, we aim to: (1) assess the predictive performance of GPR compared to other machine learning methods, such as multilayer perceptron (MLP) [38–42], multivariate linear regression (MLR) [43–47], and M5 model tree (M5P) [48–57]; (2) identify the dominant drivers of each water quality indicator; and (3) demonstrate the potential of GPR for real-time water quality monitoring.

Although machine learning has been applied to water quality prediction, most studies focus on single indicators or use simpler models (e.g., linear regression). The application of Gaussian Process Regression (GPR) to high-frequency, multi-parameter datasets in river systems remains understudied. This study addresses this gap by developing a GPR model with a Matern32 kernel, optimized via Nelder–Mead, to predict turbidity, conductivity, and UV254 absorbance. GPR and NM have been applied individually in various domains; however, the combination for multi-parameter water quality prediction in regulated river systems is a novel methodological approach. This study does not introduce a new algorithm but rather presents a novel application of established machine learning techniques to address a research gap in water quality modeling. Using the Matern32 kernel—optimized via Nelder–Mead—for multi-parameter water quality prediction represents a methodological advance. The Matern32 kernel was selected for its ability to model non-smooth, heterogeneous variability characteristic of river water quality parameters, which exhibit abrupt changes associated with hydrological events and sensor noise. While alternative kernels (e.g., RBF) could potentially yield comparable performance, a systematic comparison was beyond the scope of this study. Future work could explore the relative performance of

different kernel functions across various water quality indicators and hydrological settings. To our knowledge, no prior study has modeled turbidity, conductivity, and UV254 absorbance using this framework with such high predictive accuracy ($NSE > 0.89$) for all three indicators. This approach enables the identification of common and unique drivers of water quality dynamics, offering a more holistic understanding of river systems.

The structure of this article is as follows: Section 2 describes the experimental design, the GPR, MLP, MLR, and M5 model tree approaches in depth, as well as a thorough summary of every variable taken into account in the study. Section 3 integrates the variable importance rankings, observed measurements, and GPR predictions to present a comprehensive summary of the results obtained using this approach. Lastly, Section 4 concludes the investigation by outlining the key findings.

2. Materials and Methods

2.1. Study Area

Data from Station 906/A163, an automated water quality monitoring station along the lower course of the Ebro River in Ascó, Tarragona, Catalonia (Northeast Spain), was used for the investigation (see Figure 1a–c). The Ebro River is the largest river basin in Spain, drains an area of approximately 85,000 km², and plays a crucial role in the region's water supply, agriculture, and ecosystem services. At Ascó, the river is characterized by a regulated flow regime due to upstream reservoirs that modulate seasonal discharge variability while allowing episodic high-flow events associated with intense precipitation.

The Ascó monitoring station is located in a section of the Ebro River subjected to significant human-caused pressures, including intensive agricultural activities, urban settlements, and the Ascó Nuclear Power Plant located upstream. The plant's thermal discharges and potential chemical inputs, combined with agricultural runoff, contribute to the river's physicochemical variability. The watershed's land use is dominated by agriculture, which is a primary source of nitrates, alongside urban and industrial inputs.

The study area experiences prolonged summer droughts and mild winter precipitation under a Mediterranean climatic regime, resulting in significant intra-annual variability of water temperature, nutrient loading, and dissolved oxygen levels. Unusual algal growth can pose a serious environmental risk in aquatic systems like the Ebro River. The entire region of Aragón relies on the Ebro River as a primary source of drinking water. Therefore, preventing serious hazards, such as toxic algal blooms, is critical, and turbidity, conductivity, and absorbance measurements can serve as early warning indicators.

➤ Instrumentation and technical specifications

The station uses an integrated multi-parametric probe assembly designed for continuous automated sampling in lotic environments. The core technical attributes include:

- **Parameter suite:** Thermodynamic temperature, electrometric pH, electrical conductivity (EC), dissolved oxygen (DO), and nephelometric turbidity. All of these parameters are measured simultaneously.
- **Structural resilience:** The sensors are rated under IP68 standards, which allow for total immersion up to 60 m and ensure stability against the Ebro's hydraulic pressure and flow fluctuations.
- **Data acquisition (DAQ):** The system features a high-capacity internal logger that can buffer up to 45,000 data points, ensuring data continuity in the event of telemetry failure.

➤ Analytical methodologies

The station employs several analytical principles to minimize uncertainty and drift.

- Potentiometric and conductometric sensing: pH is determined via hydrogen-ion selective electrodes, while salinity and ion concentration are derived from four-electrode conductivity cells that compensate for polarization effects.
 - Optical luminescence (oxygenometry): Unlike traditional Clark-type polarographic sensors, this station uses optical sensors to measure dissolved oxygen. This method relies on the quenching of luminescence and offers superior long-term stability and reduced calibration frequency.
 - Piezoresistive level detection: Hydrostatic pressure is monitored using piezoresistive or capacitive transducers. These sensors measure the mechanical deformation of a ceramic diaphragm to calculate water column height with high precision.
 - Nephelometric turbidity: Measured using 90-degree light scattering (ISO 7027) [58] and calibrated to detect variations in suspended solids ranging from 0 to 1000 NTU.
- Data telemetry and environmental significance

The station operates via GPRS/satellite telemetry and transmits real-time datasets to the Ebro Basin Authority's control center. This is particularly important for monitoring the thermal plume and potential chemical discharge associated with the Ascó Nuclear Power Plant. It ensures that industrial activity remains within the ecological thresholds established by the Water Framework Directive (WFD) [59–62].

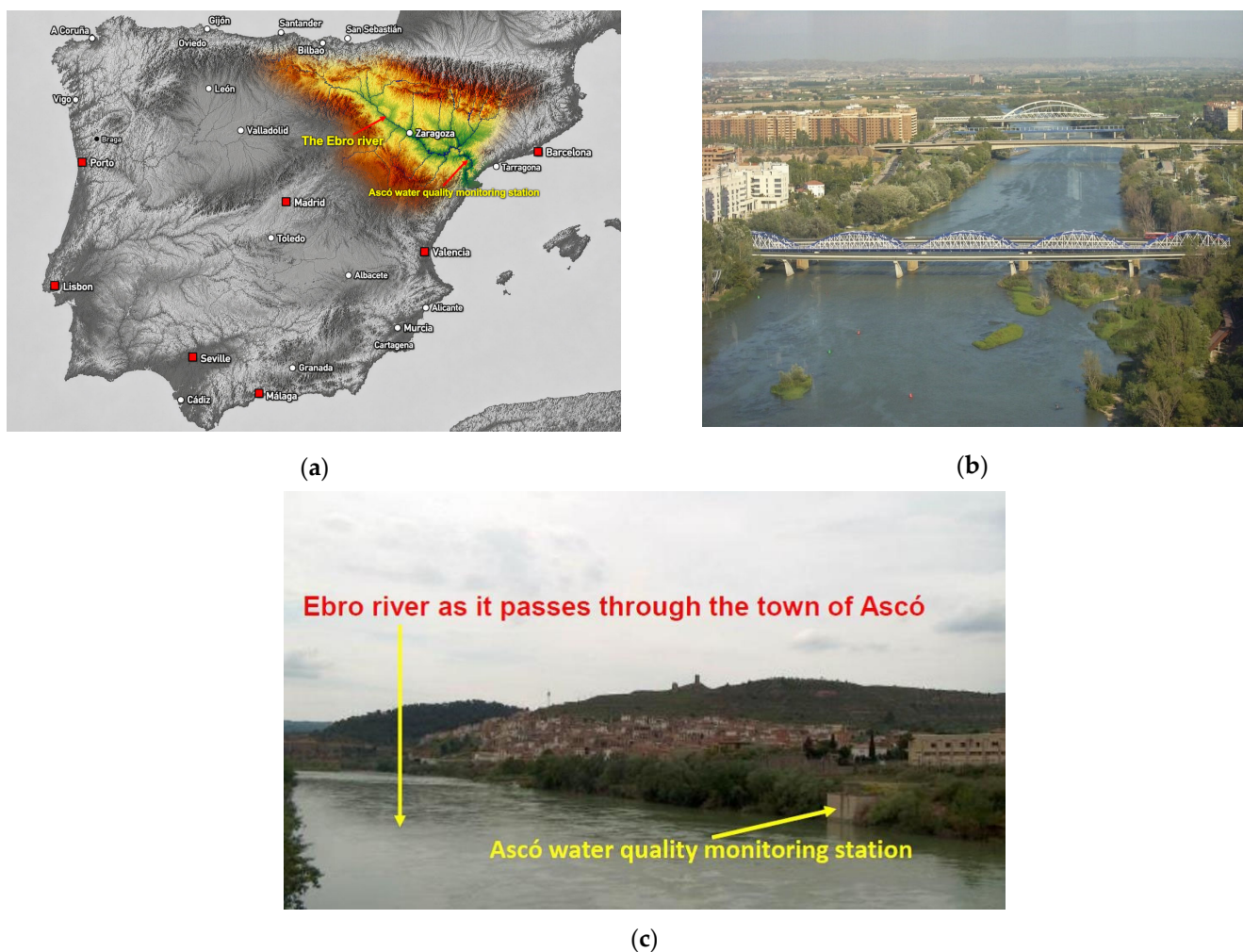


Figure 1. (a) Spain's physical map with the Ebro River's position and the Ascó water quality monitoring station (province of Tarragona, Catalonia, Northeast Spain); (b) the Ebro River as it passes through Saragossa; and (c) a closer look at the Ascó water quality monitoring station.

2.2. Experimental Dataset

The initial dataset included observations collected every 15 min from 2019 to 2024. Due to intermittent sensor failures, approximately 75% of observations had missing values for at least one variable. To avoid bias, we performed a complete-case analysis and retained 46,287 observations (25% of the total) with no missing values for any variable. The missing data primarily resulted from intermittent sensor failures and calibration issues that occurred randomly throughout the study period, rather than being systematically associated with specific seasons or hydrological conditions.

The original data's high temporal resolution (15 min intervals) ensured that the retained observations were temporally well distributed across all seasons and hydrological conditions. Using complete-case analysis ensures that the model is trained exclusively on consistent, reliable measurements, avoiding the potential biases that imputation methods could introduce. Furthermore, the established predictive relationships serve as a contingency tool for estimating water quality parameters in the event of sensor failures.

Environmental samples were collected at the Ascó automated station via a high-precision multi-parametric probe integrated into the SAICA network (Automatic Water Quality Information System). The system autonomously collects samples in situ at a fixed depth within the river's main flow. It uses piezoresistive transducers to sense discharge and level, alongside optical sensors to measure dissolved oxygen. Every sample includes a wide range of water quality indicators, such as water temperature, pH, electrical conductivity, nephelometric turbidity, and absorbance. Additionally, the probe is equipped with ion-selective electrodes for the high-frequency quantification of nitrate (NO_3^-) and ammonium (NH_4^+), as well as dissolved mercury, in every recorded sample.

2.3. Variables of the Model

2.3.1. Dependent Variables

Three water quality indicators were selected as dependent variables due to their widespread use as integrative proxies for river conditions:

- **Turbidity:** It was measured in nephelometric turbidity units (NTUs) and used as an indicator of suspended particulate matter. In large rivers, such as the Ebro, turbidity is primarily associated with sediment resuspension and transport during periods of high flow. However, it may also reflect algal biomass or fine particulate organic matter during w periods of low flow.
- **Electrical conductivity (EC):** Expressed in $\mu\text{S cm}^{-1}$, it is a proxy for the total dissolved ionic content of the water. EC integrates contributions from major ions and nutrients, and it is sensitive to dilution effects, agricultural runoff, and geochemical background conditions.
- **Ultraviolet absorbance at 254 nm (UV254):** It was selected as an optical indicator of dissolved organic matter, particularly aromatic organic compounds. Expressed in dimensionless absorbance units (AUs), UV254 accounts for the path length of the measurement cuvette and is commonly used as a surrogate for dissolved organic carbon. It is relevant for both ecological processes and drinking water treatment due to its association with disinfection by-product formation.

2.3.2. Independent Variables

Seven physicochemical variables were considered as potential drivers of the selected water quality indicators (see Table 1):

- **River discharge ($\text{m}^3 \text{s}^{-1}$):** It is the primary hydrological force in the system. It regulates the movement of sediment, the dilution of dissolved components, and the movement of organic matter and nutrients from the watershed to the river channel. Due to

the regulated nature of the Ebro River, discharge variability reflects both natural hydrological events and reservoir management operations.

- Nitrate concentration (NO_3^-): It was included as a key nutrient indicator because it is strongly influenced by agricultural activities within the basin. Nitrates directly contribute to the ionic strength of river water and are closely linked to biogeochemical processes, such as primary production and nitrification–denitrification dynamics.
- pH reflects the river’s acid–base status and incorporates carbonate buffering, biological activity, and anthropogenic inputs. Variations in pH influence chemical speciation, nutrient availability, and microbial processes and may indirectly affect turbidity, EC, and UV254.
- Dissolved oxygen (DO): It is a fundamental indicator of metabolic balance in aquatic systems. DO concentrations are controlled by temperature, flow conditions, reaeration, and biological respiration and production. Including DO allows for the assessment of links between biogeochemical activity and water quality indicators.
- Water temperature plays a central role in controlling chemical reaction rates, biological activity, and gas solubility. Seasonal temperature fluctuations in the Mediterranean climate of the study area are expected to influence nutrient cycling and organic matter dynamics.
- Ammonium (NH_4^+): It is considered a reduced form of inorganic nitrogen. It is often associated with wastewater inputs or internal biogeochemical recycling. Although it is typically present at lower concentrations than nitrate, ammonium can indicate localized pollution sources and influence oxygen dynamics through nitrification.
- Dissolved mercury (Hg): It was included as a trace contaminant of interest due to its environmental relevance and potential links to industrial activities. Although mercury concentrations are generally low, they allowed for the assessment of whether trace metals exert any detectable influence on bulk water quality indicators.

Table 1. List of the input and output variables used in this study, along with their means and standard deviations.

Input Variables	Parameter Name	Mean	Standard Deviation
Water temperature ($^{\circ}\text{C}$)	Water_temp	19.14	6.12
pH values	pH_values	8.27	0.15
Dissolved oxygen (mg/L)	DO	8.32	1.97
Ammonium (mg/L)	NH_4^+	0.03	0.02
Nitrate (mg/L)	NO_3^-	10.22	2.65
Dissolved mercury ($\mu\text{g/L}$)	Hg	0.001	0.01
Flow rate (m^3/s)	FR	272.12	197.96
Output variables			
Turbidity (NTU)	Turb	12.41	9.58
Conductivity ($\mu\text{S}/\text{cm}$)	Conductivity	1022.1	291.07
UV254 absorbance (AU/m)	Absorb	6.05	2.07

2.4. Mathematical Modeling

2.4.1. Gaussian Process Regression (GPR)

A Gaussian Process (GP) is a nonparametric probabilistic model for functions. For each finite subset, a GP is defined as a collection of random variables with a multivariate normal distribution [21,27,28]:

$$f(\mathbf{x}) \sim \text{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \quad (1)$$

where:

- $m(x) = E[f(x)]$ is the mean function (or expected value) of the function $f(x)$ at the input point x ($m(x) = 0$ is often assumed for simplicity);
- $k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))^T]$. Also known as the covariance or kernel function, it conveys assumptions about the underlying function, such as smoothness and periodicity, and quantifies the similarity between points. The covariance function is $k(x, x')$, which must allow for a positive definite kernel and can be thought of as a measure of the confidence in $m(x)$.

Since it reflects the latent regression model's underlying presumption, the Gaussian Process depends on the choice of covariance function and is also known as the prior assumption [23,26]. The function employed as kernel in this study, which, if we call $r = \|x_i - x_j\|_2$, is expressed as [21–23,26]:

- Matern32-type kernel:

$$k(x_i, x_j) = a \left(1 + \frac{\sqrt{3}r}{\ell} \right) e^{-\frac{\sqrt{3}r}{\ell}} \quad (2)$$

Therefore, the following parameters determine the kernel's type [21,23]:

- Characteristic length scale (ℓ): it regulates the horizontal scale on which the function fluctuates;
- Parameter a : it is the amplitude or magnitude of the kernel function.

The choice of parameters directly impacts the efficiency of the Gaussian Process. Hyperparameter optimization was conducted using the Nelder–Mead (NM) algorithm to maximize the marginal likelihood [33–37,63].

Assume that we have a training set $D = \{(x_i, y_i) / i = 1, 2, \dots, N\}$ for the Gaussian method, where σ_n is the standard deviation of the Gaussian noise $\varepsilon \sim N(0, \sigma_n^2)$. So, the observed values $y = \{y_i\}_{i=1}^N$ and a predicted value f^* at a new test point x^* follow a joint normal distribution [22–27]:

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N \left(0, \begin{bmatrix} k(X, X) + \sigma_n^2 I & k(X, x^*) \\ k(x^*, X) & k(x^*, x^*) \end{bmatrix} \right) \quad (3)$$

where:

- $k(X, X)$: is the covariance matrix of the training dataset X ;
- $k(x^*, x^*)$: is the covariance matrix of the test dataset;
- $k(X, x^*)$: depicts the covariance matrix obtained from the training dataset X and the new point x^* . Furthermore, $k(x^*, X) = k(X, x^*)^T$.

The function values $f(x)$ are usually not readily visible in most applications. Actually, the only observations that are available are the noisy ones, which are expressed as follows:

$$y = f(x) + \varepsilon \quad (4)$$

so we assume that ε is the additive white noise, which can be represented by a normal distribution with a mean of 0 and a standard deviation of σ_n . We also assume that Gaussian noise is independent and uniformly distributed. As some references [19–27] explain, a Gaussian Process can be represented by any restricted subset of the observed values:

$$y \sim \text{GP} \left(m(x), k(x, x') + \sigma_n^2 \delta_{ij} \right) = \text{GP} \left(0, k(x, x') + \sigma_n^2 \delta_{ij} \right) \quad (5)$$

where δ_{ij} , which is defined as follows, represents the Kronecker delta function:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The aim of the GPR model is to predict the function value $\bar{f}^*$ and its variance $cov(f^*)$ at a new test point $\mathbf{x}^*$ is. It is feasible to condition the prior assumption on the observed data and assess the likelihood of predictions for f^* because $\mathbf{y}$ and f^* are jointly distributed. This can be stated as follows [21–23]:

$$f^* | \mathbf{x}^*, X, \mathbf{y} \sim N(\bar{f}^*, cov(f^*)) \quad (6)$$

where

$$\bar{f}^* = E[f^* | \mathbf{x}^*, X, \mathbf{y}] = k(\mathbf{x}^*, X) [k(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \quad (7)$$

$$cov(f^*) = K(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X) [k(X, X) + \sigma_n^2 I]^{-1} k(X, \mathbf{x}^*) \quad (8)$$

New test input points can then be predicted using the resulting distribution. Specifically, $\bar{f}^*$ is the expected output value of the GPR model for a test point.

Additionally, the confidence interval (CI) for the expected output can be computed using the variance $cov(f^*)$. For instance, the 95% confidence interval (CI) for GPR predictions can be computed using $[\bar{f}^* - 2 \times \sqrt{cov(f^*)}, \bar{f}^* + 2 \times \sqrt{cov(f^*)}]$. Thus, the GPR model not only provides predicted values but also measures the degree of confidence in them. Furthermore, since the GPR model's predicted outputs depend solely on the inputs and observed values $\mathbf{y}$, it is a nonparametric model.

2.4.2. Multilayer Perceptron-Type Artificial Neural Network (MLP)

Minsky and Papert demonstrated [38–40] that nonlinear problems, such as the XOR function, could not be solved by a simple perceptron or an adaptive linear element (ADALINE) [38–42]. While several simple perceptrons can solve certain nonlinear tasks, there was no automated way to modify the weights in the hidden layers. By propagating errors backward through the hidden layers, the Generalized Delta Rule (GDL), initially introduced by Rumelhart and colleagues in 1986 [39], allows weight updates. This advancement enabled the use of nonlinear activation functions across multiple layers. It has been shown that the resulting multilayer perceptron (MLP) serves as a universal approximator [39–42]. Therefore, an MLP can be used to model nonlinear interactions between independent and dependent variables. This architecture has become one of the most widely adopted forms of artificial neural networks.

A particular kind of multilayer artificial neural network (ANN), the multilayer perceptron (MLP), can address issues that cannot be resolved by linear techniques [39–42]. This limitation is the main restriction of the simple perceptron. In contrast, an MLP can use either full or local connectivity. In a fully connected layer, each neuron is linked to every neuron in the adjacent layer, whereas an MLP with local connections does not satisfy this requirement.

Figure 2 shows an MLP consisting of three types of layers [38–42,64]:

- Input layer: This layer processes the data related to the independent variables after receiving them;
- Output layer: This layer generates the network's final output and represents the dependent variables;
- Hidden layers: These middle layers process and propagate information throughout the network, situated between the input and output layers.

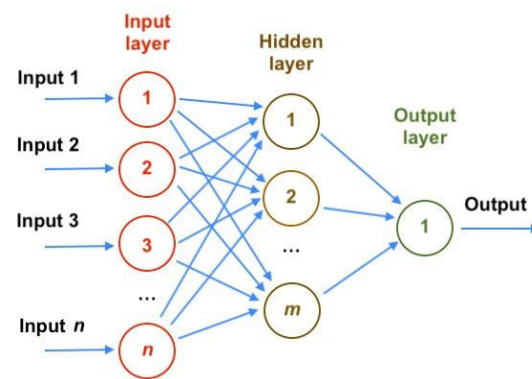


Figure 2. Example of an artificial neural network inspired by the MLP architecture. The hidden layer consists of n neurons, the input layer comprises m neurons, and the output layer contains a single neuron.

The backpropagation mathematical framework, also known as error backpropagation or the extended delta rule, is used to train these ANNs [38–42,64]. Consequently, the term “backpropagation artificial neural network” is commonly used to describe an MLP. The use of differentiable activation (transfer) functions by these ANN processing units, or neurons, is a basic feature. MLP uses this type of learning, adjusting its connection weights based on differences between the expected and observed output values. This adjustment is implemented through backpropagation, which extends the least mean square (LMS) approach used in the linear perceptron. The error at node j for the n -th data point is defined as $e_j(n) = d_j(n) - y_j(n)$, where d denotes the observed value and y represents the value predicted by the MLP.

2.4.3. Multivariate Linear Regression (MLR)

Let $y_i \in \Re$ be the scalar response variable, $x_i \in \Re^p$ be a vector of explanatory factors, and $\{x_i, y_i\}_{i=1}^n$ be a sample of independent observations. The definition of the multivariate linear regression model is [43–47]:

$$y_i = \beta_0 + x_i^T \beta + \varepsilon_i, i = 1, 2, \dots, n \quad (9)$$

where:

- The independent term (termed the intercept) is denoted by $\beta_0 \in \Re$;
- The parameter vector is called $\beta \in \Re^p$;
- Random mistake is referred to as ε_i .

Determining the design matrix in matrix notation [43,44]:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \in \Re^{n \times (p+1)} \quad (10)$$

the parameter vector is $\theta = (\beta_0, \beta^T)^T$ and the response vector is $y = (y_1, y_2, \dots, y_n)^T$, and the model is given by the expression [43,44]:

$$y = X\theta + \varepsilon \quad (11)$$

The answer to the optimization issue is known as the ordinary least squares (OLS) estimator [43–47]:

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{p+1}} \|y - X\theta\|_2^2 \quad (12)$$

The objective function is differentiable and convex. The *normal equations* result from the first-order condition of relative or local extremum [45,46]:

$$X^T X \hat{\theta} = X^T y \quad (13)$$

The estimator has a closed form under the full-range assumption [43–47]:

$$\hat{\theta} = (X^T X)^{-1} X^T y \quad (14)$$

Figure 3 offers a multivariate linear regression model example.

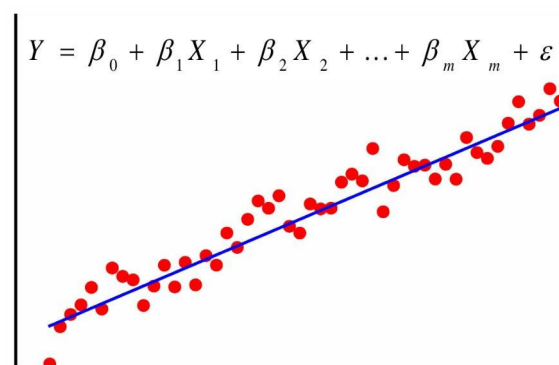


Figure 3. Illustrative representation of the MLR model.

2.4.4. M5 Model Tree

Decision trees are nonparametric predictive models that minimize an error criterion by recursively dividing the feature space into disjoint parts [48–57]. Classification trees assign discrete labels, while regression trees estimate continuous values, typically using the mean of the observations in each leaf.

This approach may be limited when the relationship between the predictor variables and the response variable is approximately linear in local regions of the input space, but not globally.

The M5 method introduces model trees, which were first presented by Quinlan in 1992 [48] and then refined [49,50]. Unlike traditional regression trees, M5 captures local linear relationships more accurately by replacing multivariate linear regression models with constant values in the leaves. The main objective of the M5 model tree is to combine the prediction accuracy of linear models with the interpretability of decision trees. The components of an M5 model tree are [48–57]:

- Binary divisions based on an explanatory variable are represented by internal nodes;
- Leaves, which have the following linear regression model [48–57]:

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i x_i \quad (15)$$

so that:

- $\hat{y}$ is the variable of interest (or target variable);
- x_i are the predictor variables;
- β_i are the coefficients estimated using least squares.

As a result, each leaf designates an area of the input space where it is expected that the relationship is roughly linear. Rather than using the traditional variance, the M5 algorithm chooses the splittings based on the projected decrease in the target variable's standard deviation [48–57]:

$$\Delta SD = SD(T) - \sum_k \frac{|T_k|}{|T|} SD(T_k) \quad (16)$$

so that:

- T is the node's dataset;
- T_k are the partitions' resultant subsets;
- The standard deviation is denoted by $SD(\cdot)$.

In terms of anticipated value, this criterion favors divisions that produce more homogeneous subgroups. After building the tree, the observations reaching each leaf are used to fit a linear regression model to that leaf. To prevent overfitting:

- Backward selection is used to remove irrelevant variables;
- A small number of observations penalizes the model's complexity.

The M5 uses an estimated error-based pruning technique, contrasting:

- The subtree's overall error;
- The node's single linear model's inaccuracy.

The subtree is swapped out for the linear model if its estimated error is lower. This procedure enhances the generalizability of the tree. M5 uses a smoothing procedure that combines the prediction of the leaf model with those of its predecessor nodes to prevent sudden discontinuities between neighboring leaves [48–57]:

$$\hat{y}_{smooth} = \frac{n \cdot \hat{y}_{leaf} + k \cdot \hat{y}_{parent}}{n + k} \quad (17)$$

so that:

- n is the number of instances of the leaf;
- The smoothing parameter is k .

An illustration of a simple M5 model tree is displayed in Figure 4.

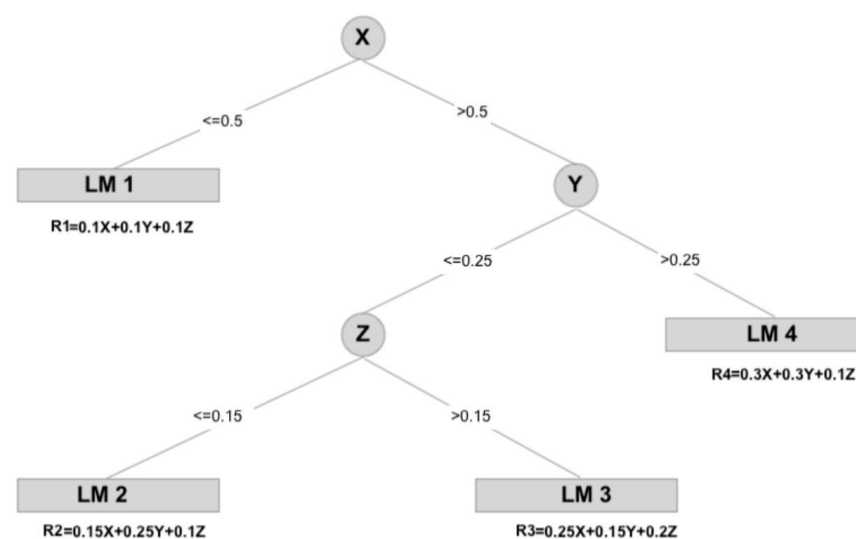


Figure 4. An M5 model tree example.

2.4.5. Nelder–Mead (NM) Algorithm

Selecting the right combination of hyperparameters is crucial. A process of optimization that minimizes a specific loss function is required to select the optimal hyperparameters, which govern how algorithms learn. Cross-validation is commonly used to evaluate the performance of the resulting approximations [33–37].

The Nelder–Mead method [33,63] is a derivativeless numerical optimization algorithm designed to minimize real functions of multiple variables $f: \mathfrak{R}^n \rightarrow \mathfrak{R}$. It is a direct search method that operates on a simplex formed by $n + 1$ vertices in the space $\mathfrak{R}^n$. The algorithm iteratively updates the simplex's position through geometric transformations intended to approximate the objective function's local minimum. Let the simplex at iteration k be [33–37]:

$$S^{(k)} = \{x_1^{(k)}, x_2^{(k)}, \dots, x_{n+1}^{(k)}\}, x_i^{(k)} \in \mathfrak{R}^n \quad (18)$$

ordered according to the function values:

$$f(x_1^{(k)}) \leq f(x_2^{(k)}) \leq \dots \leq f(x_{n+1}^{(k)}) \quad (19)$$

It is defined as follows:

- $x_b = x_1$: best point;
- $x_g = x_n$: worst good point;
- $x_w = x_{n+1}$: worst point.

The centroid excluding the worst point is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (20)$$

The NM optimizer applies four transformations according to the behavior of f [33–37]: (a) reflection, $x_r = \bar{x} + \alpha(\bar{x} - x_w)$, with typical reflection coefficient $\alpha = 1$; (b) expansion. If $f(x_r) < f(x_b)$, then evaluate $x_e = \bar{x} + \gamma(\bar{x} - x_w)$ with expansion coefficient $\gamma = 2$. (c) Contraction: (1) outer contraction (if $f(x_b) \leq f(x_r) < f(x_w)$) $x_{oc} = \bar{x} + \rho(x_r - \bar{x})$; (2) inner contraction (if $f(x_r) \geq f(x_w)$) $x_{ic} = \bar{x} - \rho(\bar{x} - x_w)$, with contraction coefficient $\rho = 0.5$. (d) Simplex reduction. If the contraction does not improve, we shall take $x'_i = x_b + \sigma(x_i - x_b), i = 2, \dots, n + 1$, with $\sigma = 0.5$.

Therefore, the NM optimizer is a heuristic method that works well with continuous, non-differentiable, or noisy functions. Its practical success stems from its ability to model the local topologies of f using simple geometric operations [33–37,65].

2.5. Model Validation and Performance Evaluation

The NM/GPR-based methodology was developed using the seven input variables listed in Section 2.3. The model targets the estimation of turbidity, conductivity, and absorbance. The optimal model, which exhibited the best fit to the observed dataset, was selected to predict these three output variables with excellent precision and reliability. The main assessment metric used in this investigation was the Nash–Sutcliffe Efficiency (NSE) index. The NSE is a widely used statistical metric in hydrology and environmental modeling that evaluates a model's predictive ability. It measures the relative improvement of the model's predictions over a simple benchmark, such as the mean of the observed data. Mathematically, the NSE is equivalent to the coefficient of determination (R^2) [66–70], as both metrics are calculated using the same formula. In hydrological studies, however, the NSE is preferred because it is more interpretable in the context of comparing model performance to a baseline. Values greater than 0.75 are typically considered indicative of good predictive skill.

The observed values t_i are compared with the model's projected values y_i . Consequently, the following sums of squares are described below [66–70]:

- The explained sum of squares is denoted by $SS_{reg} = \sum_{i=1}^n (y_i - \bar{t})^2$.
- $SS_{tot} = \sum_{i=1}^n (t_i - \bar{t})^2$: this summation is directly correlated with the sample variance.
- The residual sum of squares is denoted by $SS_{err} = \sum_{i=1}^n (t_i - y_i)^2$.

where $\bar{t}$ is the experimental data's mean value, which is defined as follows:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i \quad (21)$$

The NSE index (or its mathematical equivalent, the coefficient of determination) is then provided by [66–70]:

$$NSE \equiv 1 - \frac{SS_{err}}{SS_{tot}} \quad (22)$$

The difference between the experimental and planned data decreases as the NSE index (R^2 statistic) gets closer to 1.0.

Two other performance metrics considered in this study were the mean absolute error (MAE) and root mean square error (RMSE) [67,69]. These metrics are widely utilized for evaluating the predictive accuracy of mathematical models. In particular, the following expressions define the RMSE and MAE [66–70]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (t_i - y_i)^2}{n}} \quad (23)$$

$$MAE = \frac{\sum_{i=1}^n |t_i - y_i|}{n} \quad (24)$$

Perfect concordance between the expected and observed data is shown by a root mean square error (RMSE) value of zero. The mean absolute error (MAE) measures the average vertical departure of each data point from the identity line and can also be understood as the average horizontal distance between each point and the identity line. In a broader sense, the MAE accurately measures the average absolute difference between the observed and projected values t_i and y_i .

We have also used the Percent Bias (PI) index, which is a statistical metric that evaluates systematic overestimation or underestimation in a predictive model. It quantifies the average tendency of predicted values to deviate from observed values, expressed as a percentage. A PI value of 0% indicates perfect agreement between predictions and observations, while positive values suggest overestimation and negative values indicate underestimation.

$$PI = \frac{\sum_{i=1}^n (t_i - y_i)}{\sum_{i=1}^n t_i} \times 100\% \quad (25)$$

Several techniques have been developed to precisely forecast turbidity, conductivity, and absorbance, including the NM/GPR-based approach, MLP models, the MLR technique, and the M5 model tree algorithm. The effectiveness of these approaches has been evaluated using the NSE index.

As previously mentioned, the characteristic length scale (ℓ), the amplitude or magnitude of the Matern32 kernel function (a parameter), and the noise variance (σ^2) are all essential components of the GPR approximation. Additionally, the Nelder–Mead (NM) optimizer was used in this study to identify the optimal parameters for GPR models. This approach uses the concept of a simplex, which is a polytope with $N + 1$ vertices in N dimensions. Examples include a triangle in a plane, a tetrahedron in three dimensions,

a line segment in a line, and so forth [33–37]. It represents a systematic exploration of a predefined subset of the parameter space using the specified statistical machine learning algorithm. The approach looks for a local optimal solution approximation to a problem involving N variables of a function that needs to be minimized.

Although the dataset exhibits high temporal resolution, time was not included as a predictor variable because the study aims to model instantaneous relationships between physicochemical drivers and water quality indicators. The training set makes up 50% of the data, and the testing set makes up the remaining 50. The NM/GPR model is constructed using the training subset. The NM algorithm is employed to calibrate the GPR hyperparameters via a 10-fold cross-validation procedure [71,72]. To ensure robustness against potential autocorrelation, 10-fold cross-validation with random splitting was employed to ensure that both the training and testing subsets captured the full range of hydrological and seasonal conditions. After identifying the ideal hyperparameters, the entire training dataset is used to build the model. The model is then used to predict the values in the test set. We then assess the goodness-of-fit of the approximation by comparing these predictions with the observed values. Figure 5 illustrates the process diagram for the NM/GPR methodology implemented in this study.

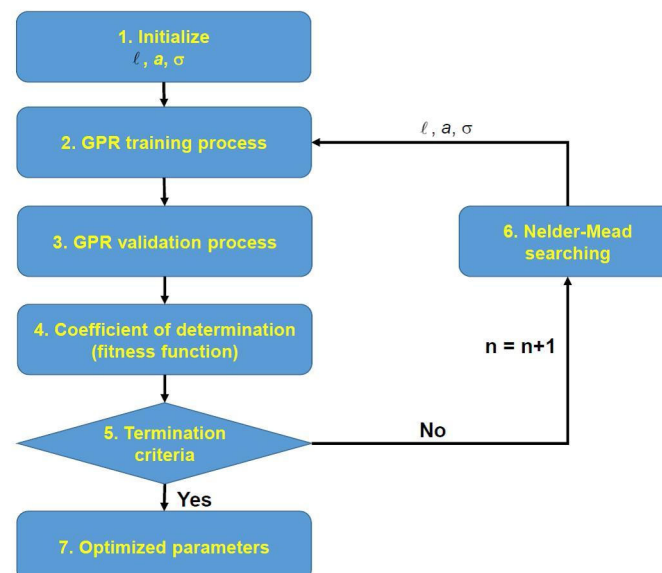


Figure 5. Process diagram illustrating the NM/GPR-based approach utilizing a Matern32-type kernel function.

Furthermore, the true NSE index is often estimated via cross-validation [71,72]. In this study, we assessed the predicted performance of the NM/GPR-based approximation using a 10-fold cross-validation procedure [72]. The open-source machine learning package gplite (v0.13.0) from the R Project was used to implement regression modeling based on the GPR framework [73]. Table 2 displays the range of the hyperparameters used in this investigation.

Table 2. Initial values of the three NM/GPR-based approach’s hyperparameters in this investigation.

GP Hyperparameters	Initial Value
ℓ	0.3
a	1
σ^2 (noise variance)	0

Similarly, the NSE index for MLP models, MLR technique, and the M5 model tree were determined using the 10-fold cross-validation process [71,72]. The MLP, MLR, and M5 model tree algorithms were implemented employing the open-source machine learning platform WEKA (v3.8.6) [74,75].

3. Results and Discussion

Table 1 displays all seven independent input factors, which correspond to the seven physicochemical parameters. A total of 240,765 samples were used in this investigation, meaning that 240,765 experimental measurements were taken and analyzed at the Ascó water quality monitoring station on the Ebro River in accordance with the European Water Framework Directive [59].

To address this complicated problem, the complete dataset was partitioned into two subsets: a training set comprising 50% of the data and a testing set consisting of the remaining 50%. The basic concept is to construct a GPR-reliant model using the training data and then use the model to test the data to produce predictions after determining the model's ideal parameters using the NM optimizer. The NM optimizer is indeed well-suited for optimizing the GPR hyperparameters. By analyzing the changes in the cross-validation error at each iteration, the NM algorithm determines the ideal values for the ℓ , a , and σ^2 parameters by examining the variations in the cross-validation error at each iteration. The search space is multidimensional, with each dimension corresponding to a distinct parameter. In this way, Table 3 shows the optimal parameters for the three NM/GPR-based models for the three water quality indicators: turbidity, conductivity, and absorbance.

Table 3. The best-fitted NM/GPR approximations' optimal hyperparameters for the turbidity, conductivity, and absorbance prediction in this investigation.

Dependent Variable	ℓ	a	σ
Turbidity	0.858	0.426	0.049
Conductivity	0.287	0.422	0.0000323
Absorbance	0.826	0.332	0.05

In light of the completed computations, the NM/GPR-based approximation has enabled the development of a model that can predict turbidity, conductivity, and absorbance with high accuracy using the test dataset. The dependent variables of turbidity, conductivity, and absorbance had correlation coefficients of $r = 0.9465$, $r = 0.9749$, and $r = 0.9583$, respectively, and R^2 values of 0.8956, 0.9503, and 0.9180, respectively, for the best GPR approaches. Table 4 displays the findings. For comparison, models were also constructed using other advanced techniques, such as the M5 model tree, multilayer perceptron (MLP), and multivariate linear regression (MLR). In all cases, the NM/GPR–Matern32 approach yielded the best performance. The outcomes are summarized in Table 4.

Table 4. NSE index, PI index, correlation coefficients (r), root mean square deviation (RMSE), and mean absolute error (MAE) over the testing set for the distinct techniques and dependent variables (turbidity, conductivity, and absorbance) fitted in this investigation.

Output Variable	Method	NSE	PI	r	RMSE	MAE
Turbidity	GPR	0.8956	0.0295	0.9465	3.0151	1.5737
	MLP	0.5292	−1.7295	0.7572	6.4015	4.3289
	MLR	0.3179	0.0958	0.5653	7.7051	4.9037
	M5P	0.2842	0.5037	0.5331	13.190	4.5601

Table 4. Cont.

Output Variable	Method	NSE	PI	r	RMSE	MAE
Conductivity	GPR	0.9503	0.1264	0.9749	64.789	32.064
	MLP	0.7167	−1.4338	0.8507	154.66	115.28
	MLR	0.5696	0.0963	0.7548	190.61	145.96
	M5P	0.1773	−5.1862	0.6077	315.26	132.49
Absorbance	GPR	0.9180	0.0035	0.9583	0.5882	0.3944
	MLP	0.6878	−0.5528	0.8345	1.1481	0.8594
	MLR	0.3581	0.0278	0.5984	1.6462	1.1024
	M5P	0.0712	4.9960	0.2668	6.8967	2.1096

Figures 6, 7, and 8, respectively, show the first- and second-order terms that comprise the GPR-based model using the Matern32 kernel for the turbidity, conductivity, and absorbance indicators.

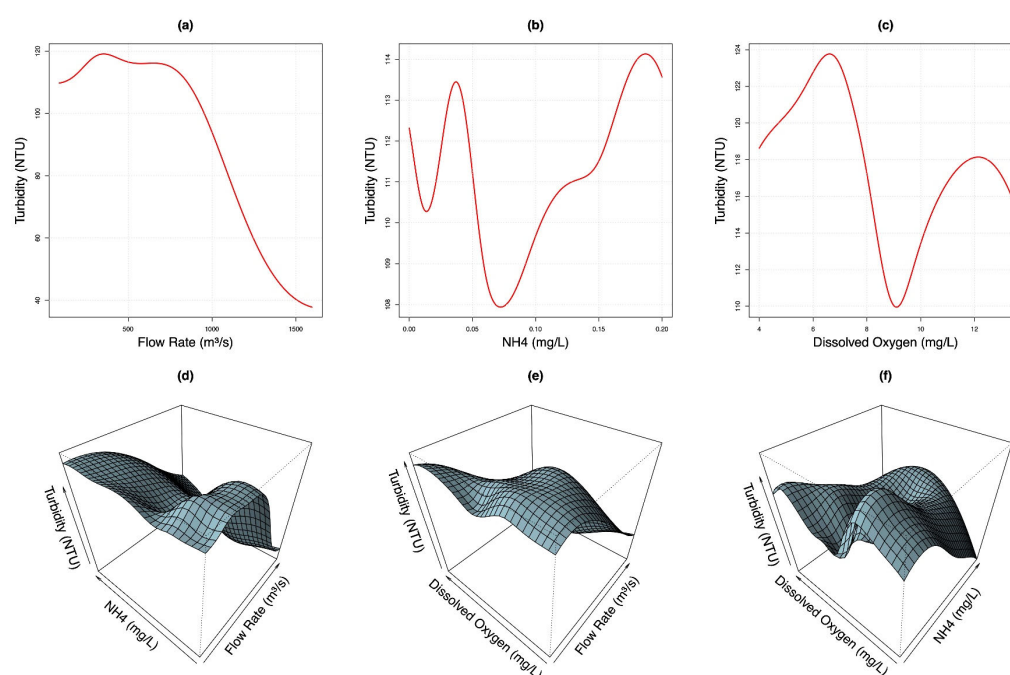


Figure 6. Visualization of the NM/GPR-based model's first- and second-order partial relationships using a Matern32 kernel for the turbidity indication, regarding its three most important independent variables: (a) turbidity vs. flow rate; (b) turbidity vs. NH_4^+ ion concentration; (c) turbidity vs. dissolved oxygen; (d) turbidity vs. flow rate and NH_4^+ ion concentration; (e) turbidity vs. flow rate and dissolved oxygen; and (f) turbidity vs. NH_4^+ ion concentration and dissolved oxygen.

Relevance of Variables

Another significant finding of this study is the crucial role that input variables play in predicting turbidity, conductivity, and absorbance for these intricate nonlinear models (see Figures 9–11). The optimal model corresponds to the NM/GPR-based approach that uses the Matern32 kernel function. The variance of the posterior latent mean was used to evaluate the relative importance of the variables in the GPR model [73–75]. This was achieved by systematically varying each variable across its observed range while holding all other variables at their respective mean values. This approach isolates the effect of each variable on the model predictions. The variance of the posterior predictive was used to measure each variable's significance, indicating how much each variable affects the model's predictions.

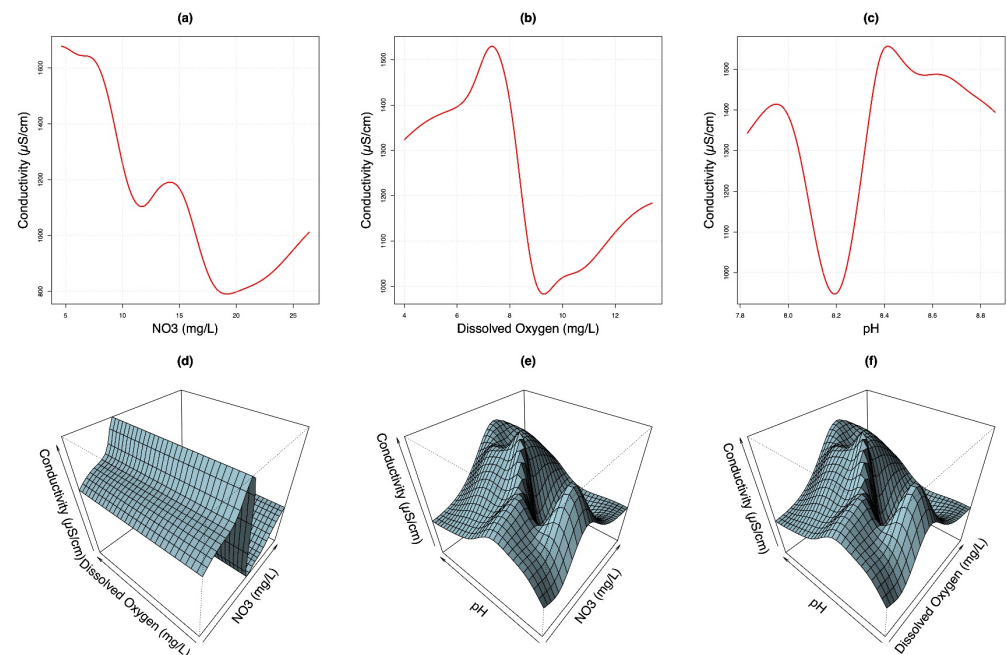


Figure 7. Visualization of the NM/GPR-based model's first- and second-order partial relationships using a Matern32 kernel for the conductivity indication, regarding its three most important independent variables: (a) conductivity vs. NO_3^- ion concentration; (b) conductivity vs. dissolved oxygen; (c) conductivity vs. pH values; (d) conductivity vs. NO_3^- ion concentration and dissolved oxygen; (e) conductivity vs. NO_3^- ion concentration and pH values; and (f) conductivity vs. dissolved oxygen and pH values.

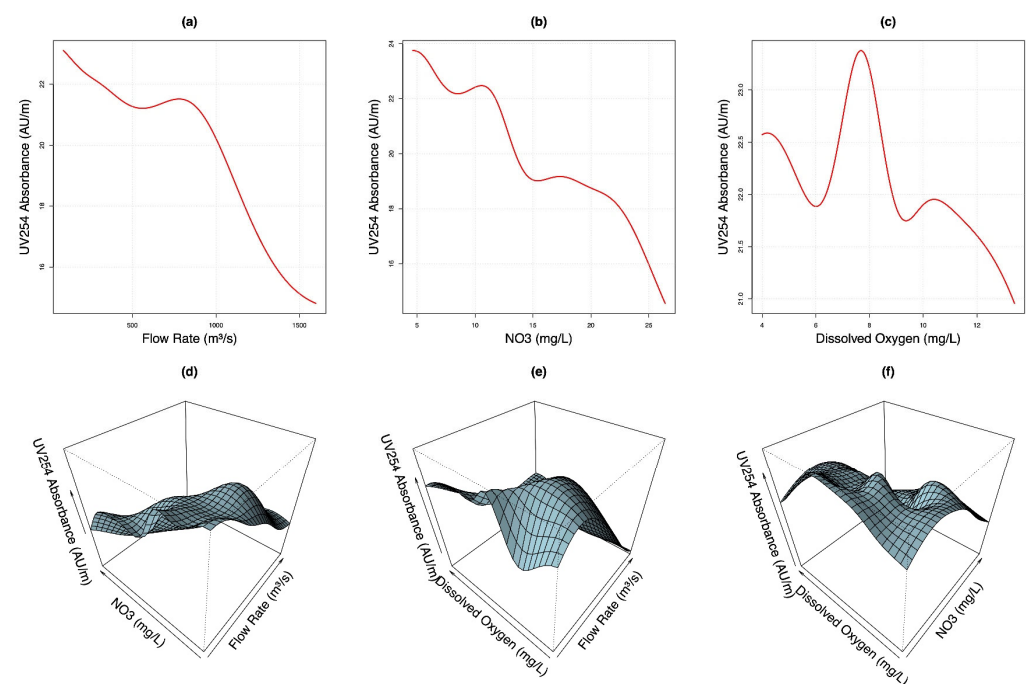


Figure 8. Visualization of the NM/GPR-based model's first- and second-order partial relationships using a Matern32 kernel for the absorbance indication, regarding its three most important independent variables: (a) absorbance vs. flow rate; (b) absorbance vs. NO_3^- ion concentration; (c) absorbance vs. dissolved oxygen; (d) absorbance vs. flow rate and NO_3^- ion concentration; (e) absorbance vs. flow rate and dissolved oxygen; and (f) absorbance vs. NO_3^- ion concentration and dissolved oxygen.

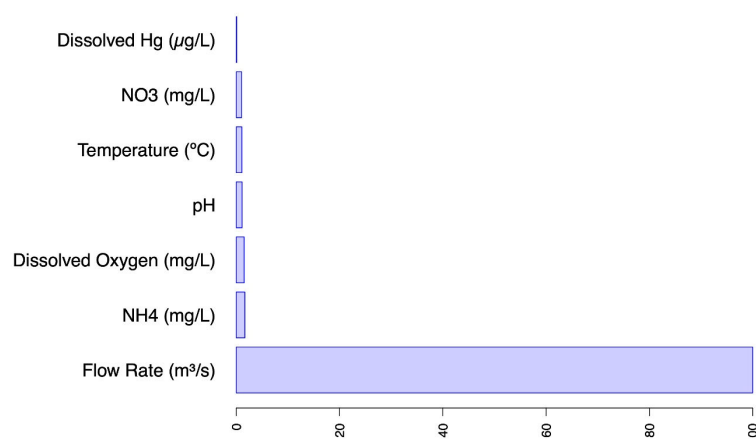


Figure 9. The input variables' relative importance for the turbidity prediction utilizing the posterior latent mean's variance for the model NM/GPR-based approximation with the Matern32 kernel.

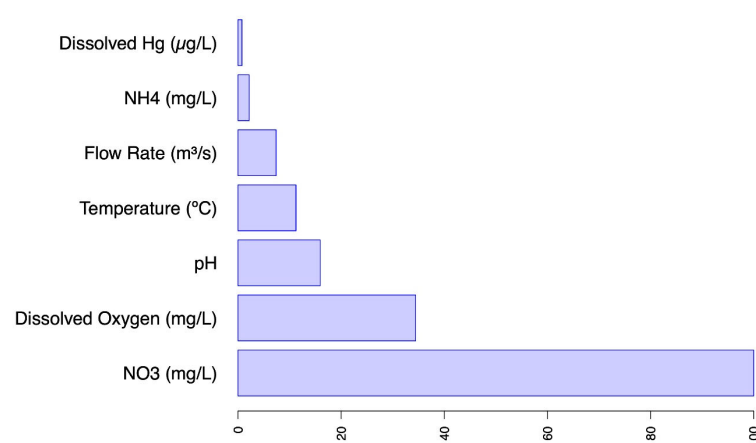


Figure 10. The input variables' relative importance for the conductivity prediction utilizing the posterior latent mean's variance for the model NM/GPR-based approximation with the Matern32 kernel.

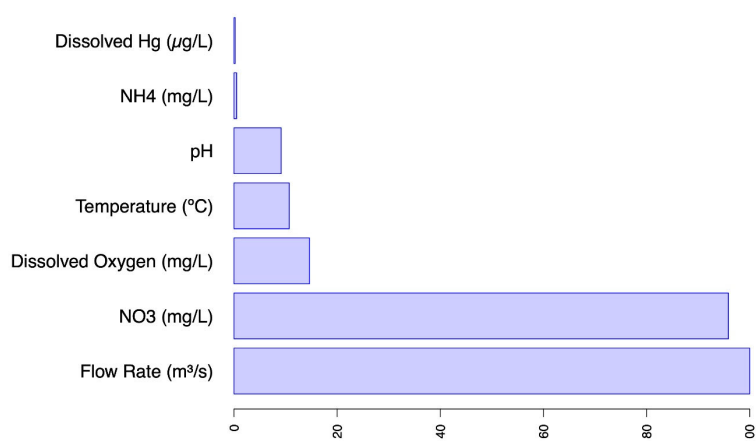


Figure 11. The input variables' relative importance for the absorbance prediction utilizing the posterior latent mean's variance for the model NM/GPR-based approximation with the Matern32 kernel.

Tables 5, 6, and 7 indeed show the variances of the posterior latent mean when one independent variable is perturbed at a time across its observed range for the normalized dependent variables turbidity, conductivity, and absorbance, respectively.

Table 5. Variance of the posterior latent mean perturbing one independent variable at a time across its observed range for the dependent normalized variable, turbidity.

Perturbed Variable	Variance
Flow rate	0.9387
NH ₄ ⁺	0.0156
Dissolved oxygen	0.0143
pH	0.0105
Water temperature	0.0101
NO ₃ [−]	0.0098
Dissolved Hg	0.0008

Table 6. Variance of the posterior latent mean perturbing one independent variable at a time across its observed range for the dependent normalized variable, conductivity.

Perturbed Variable	Variance
NO ₃ [−]	0.5813
Dissolved oxygen	0.2003
pH	0.0927
Temperature	0.0654
Flow rate	0.0430
NH ₄ ⁺	0.0125
Dissolved Hg	0.0044

Table 7. Variance of the posterior latent mean perturbing one independent variable at a time across its observed range for the dependent normalized variable, absorbance.

Perturbed Variable	Variance
Flow rate	0.4326
NO ₃ [−]	0.4147
Dissolved oxygen	0.0633
Temperature	0.0463
pH	0.0395
NH ₄ ⁺	0.0022
Dissolved Hg	0.0011

The sensitivity analysis revealed that discharge had a clear and dominant influence on turbidity dynamics. Discharge accounted for a variance contribution of 0.9387, far exceeding that of any other predictor. These results are consistent with previous findings associating increased flow with sediment resuspension and enhanced particulate transport (e.g., [76,77]). This indicates that hydrodynamic conditions are the primary driver of particulate mobilization in the study system.

All remaining predictors exhibited notably lower contributions. Ammonium (0.0156), dissolved oxygen (0.0143), pH (0.0105), water temperature (0.0101), and nitrates (0.0098) played minor yet significant roles. These variables may act as proxies for biogeochemical or seasonal processes that indirectly affect turbidity, such as biological productivity or temperature-driven stratification effects [13]. Dissolved mercury contributed only 0.0008, indicating its minimal relevance to particulate dynamics in this system. This aligns with mercury's predominantly dissolved speciation of mercury under typical freshwater physicochemical conditions [78].

For the conductivity model, nitrate concentration was the strongest predictor, contributing 0.5813 to the variance. This result is consistent with the well-established ionic contribution of nitrate anions to bulk conductivity in freshwater systems (e.g., [6]). The

subsequent importance of dissolved oxygen (0.2003) and pH (0.0927) suggests that redox and acid–base conditions also modulate ionic equilibria that influence conductivity, likely through their control of carbonate and metal speciation [79].

Water temperature (0.0654) also contributed meaningfully, as expected, since conductivity increases with temperature due to enhanced ionic mobility [80]. Discharge, ammonium, and dissolved mercury contributed progressively less (0.0430, 0.0125, and 0.0044, respectively), which reinforces the idea that conductivity in this system is primarily modulated by nutrient-driven ionic loads rather than by hydrodynamic or trace metal factors.

The model for UV254 absorbance, an optical proxy for dissolved organic carbon (DOC), revealed a more balanced distribution of variable influence. Both discharge (0.4325) and nitrates (0.4147) exerted substantial effects. The importance of discharge is consistent with the mobilization of DOC from catchment soils during high-flow events [81], while the strong nitrate signal may reflect shared biogeochemical sources or parallel responses to seasonal or land use drivers [82].

Other variables made moderate yet meaningful contributions: dissolved oxygen (0.0633), water temperature (0.0463), and pH (0.0395). These factors influence DOC composition, microbial activity, and chromophoric dissolved organic matter (CDOM) reactivity [83]. Ammonium (0.0022) and dissolved mercury (0.0011) contributed minimally, suggesting limited direct coupling with aromatic organic matter signatures at 254 nm.

Taken together, the results highlight that discharge is a primary driver of turbidity and optical parameters, reflecting hydrology's overarching role in mobilizing particulate and dissolved material. In contrast, conductivity is governed primarily by nitrate concentrations, reflecting the ionic nature of this variable. The relatively low influence of dissolved mercury across all models is consistent with its trace-level occurrence and the dominance of geochemical speciation over direct physical or optical effects in natural waters.

The divergence in influential predictors across the three dependent variables highlights the necessity of using multi-parameter models to capture the heterogeneous processes that govern water quality dynamics. These findings contribute to a broader understanding of hydrochemical controls in freshwater systems and can inform the design of monitoring programs and predictive management tools.

Overall, our study clarifies the relative importance of several environmental parameters and emphasizes the complexity of turbidity, conductivity, and absorbance regulations in aquatic ecosystems. By calculating the relative importance of each variable and evaluating its individual contribution to model accuracy, this study provides valuable insights into the river water quality dynamics. These findings have implications for ecosystem monitoring, water quality management, and creating focused mitigation plans for high chlorophyll concentrations and their ecological effects. Future research that considers temporal and regional variability will improve our understanding of these complicated linkages and facilitate the implementation of more efficient methods for maintaining ecosystem health and water quality.

Figure 12 presents a comparison of the predicted and observed turbidity values using the M5 model tree (Figure 12a), MLR model (Figure 12b), MLP model (Figure 12c), and the NM/GPR-based approach with a Matern32 kernel (Figure 12d). The findings underscore the importance of using the GPR methodology in order to find the most precise solution to the regression model. These results unequivocally show that the NM/GPR-based model with a Matern32 kernel provides the best fit and meets the critical statistical goodness-of-fit criterion (R^2) requirement. As a result, solving this issue successfully requires integrating the GPR framework with the NM optimizer.

Similarly, Figure 13 contrasts the predicted and experimental conductivity values obtained using the M5 model tree (Figure 13a), the MLR model (Figure 13b), the MLP model (Figure 13c), and the NM/GPR-based approximation using the Matern32 kernel (Figure 13d). The results suggest that adopting a GPR framework is essential to solving the regression problem as accurately as possible. The findings clearly show that the NM/GPR approach using the Matern32 kernel satisfies the primary statistical goodness-of-fit criterion and provides the most accurate overall fit (R^2). For this reason, using the NM optimizer in conjunction with the GPR approach (Matern32 kernel) is crucial to overcoming this issue.

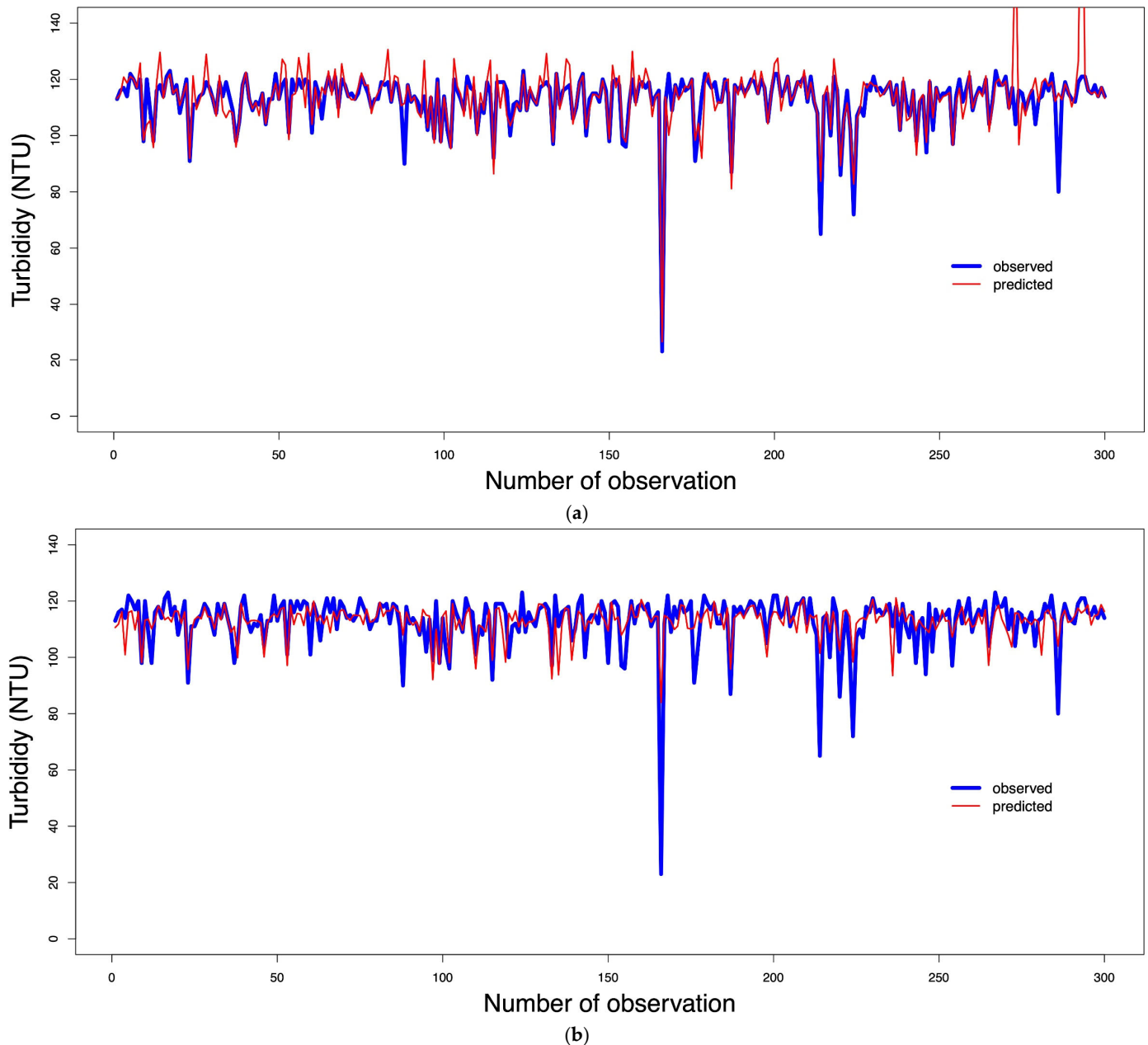


Figure 12. Cont.

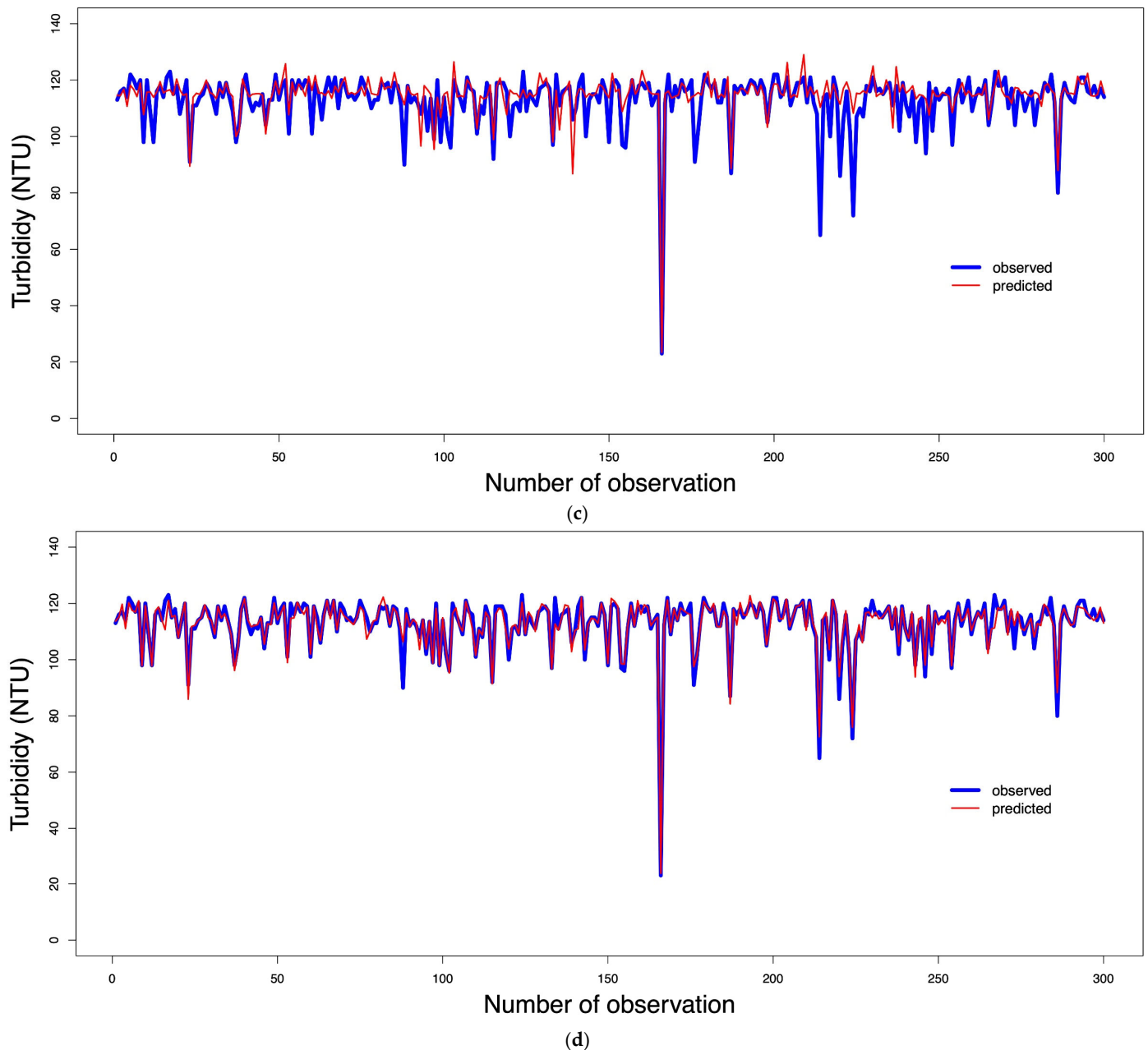


Figure 12. Comparison of observed versus predicted turbidity values for 300 randomly selected samples from the testing dataset using: (a) M5 model tree ($R^2 = 0.2842$); (b) MLR model ($R^2 = 0.3179$); (c) MLP model ($R^2 = 0.5292$); and (d) NM/GPR-based model with a Matern32 kernel ($R^2 = 0.8956$).

Moreover, Figure 14 compares the predicted and experimental absorbance values obtained using the M5 model tree (Figure 14a), the MLR model (Figure 14b), the MLP model (Figure 14c), and the NM/GPR-based approximation employing the Matern32 kernel (Figure 14d). The results suggest that using a GPR framework is crucial for finding the most accurate solution to the regression problem. These results clearly demonstrate that the NM/GPR-based approximation with the Matern32 kernel provides the best overall fit and satisfies the key statistical goodness-of-fit criterion (R^2). To properly address this issue, the GPR approach (Matern32 kernel) must be integrated with the NM optimizer.

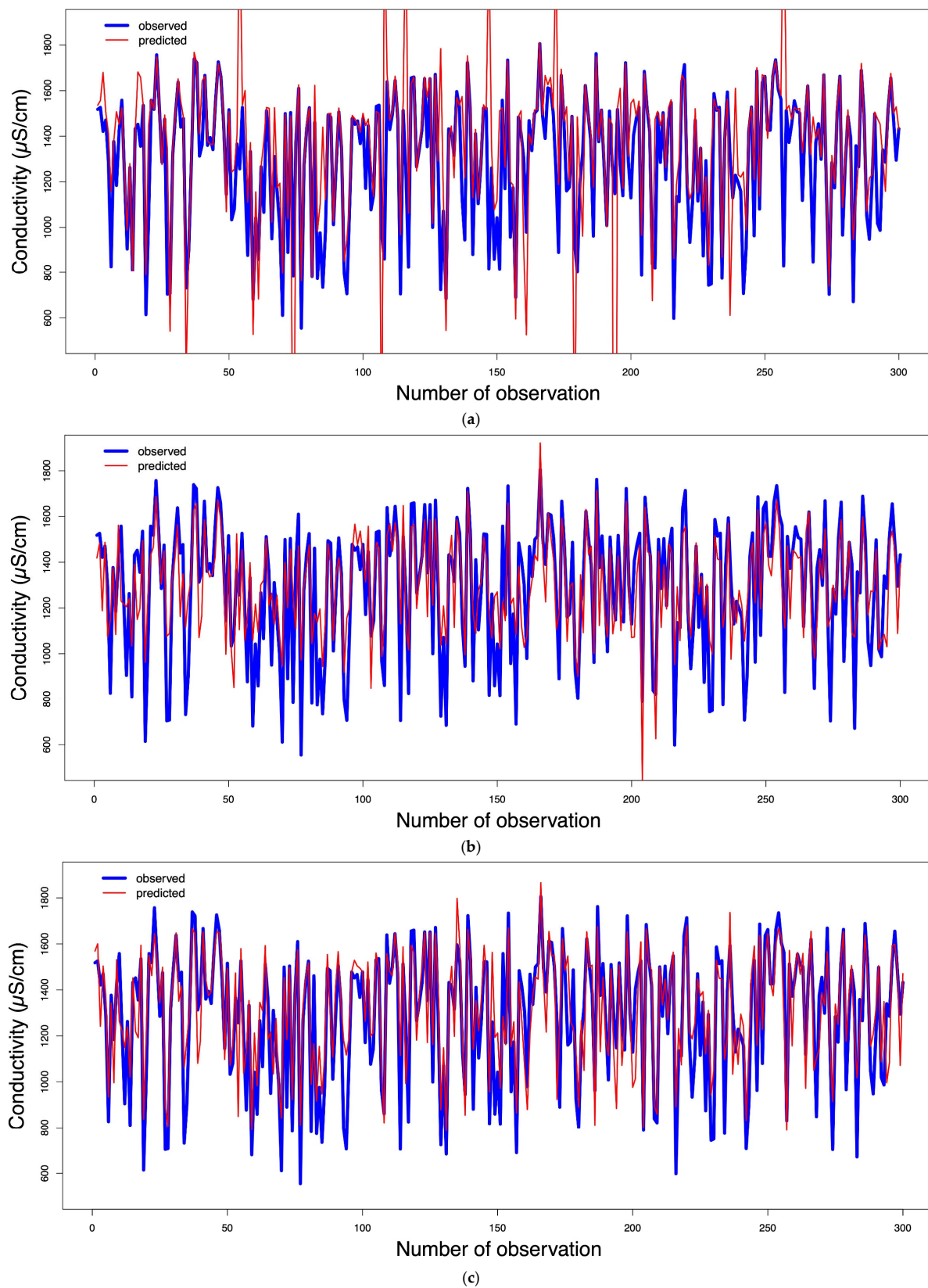


Figure 13. Cont.

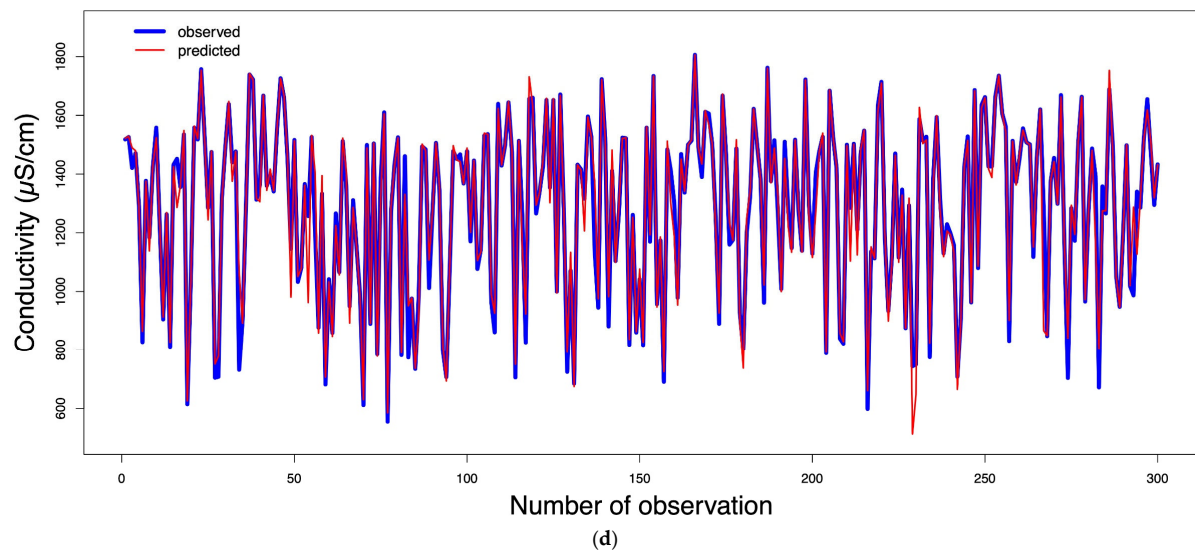


Figure 13. Illustration of the observed vs. predicted conductivity values for 300 random samples of the testing dataset utilizing: (a) M5 model tree ($R^2 = 0.3693$); (b) MLR model ($R^2 = 0.5696$); (c) MLP model ($R^2 = 0.7167$); and (d) NM/GPR-based model with a Matern32 kernel ($R^2 = 0.9503$).

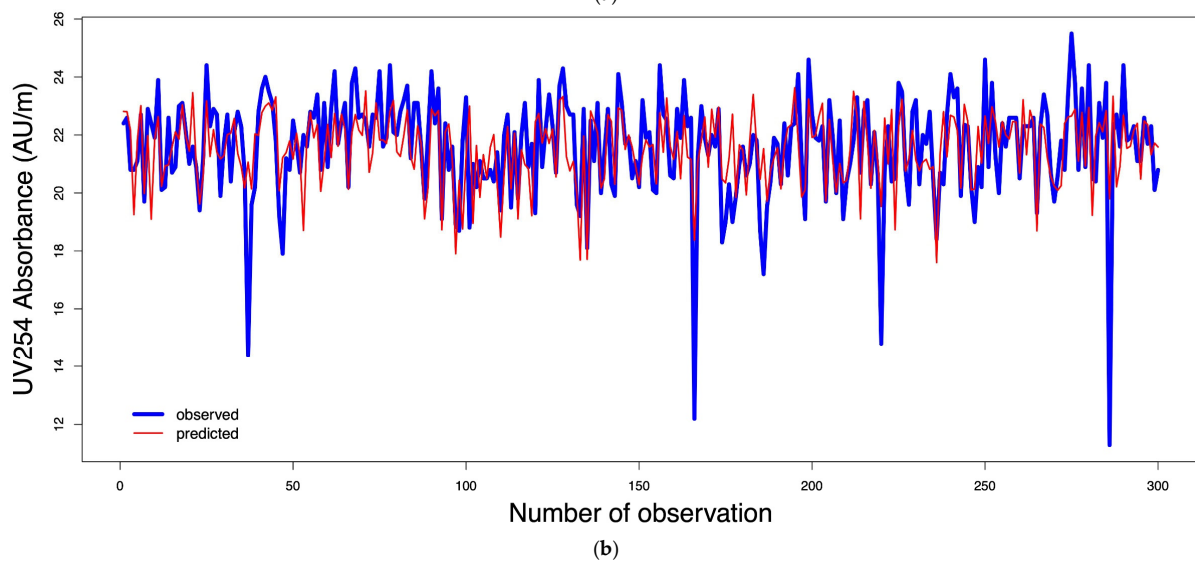
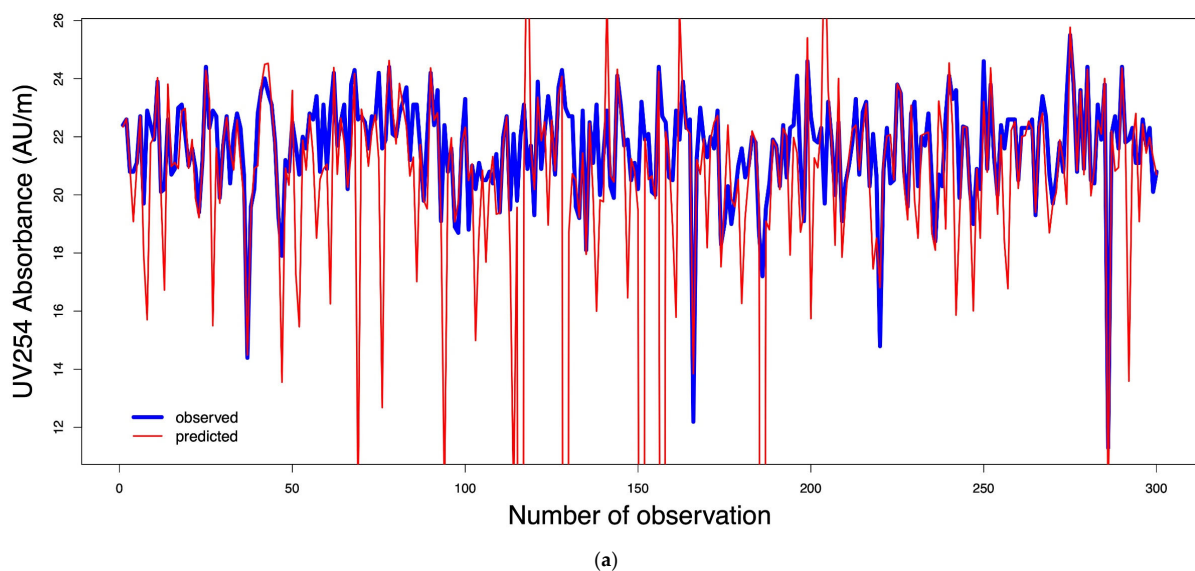


Figure 14. Cont.

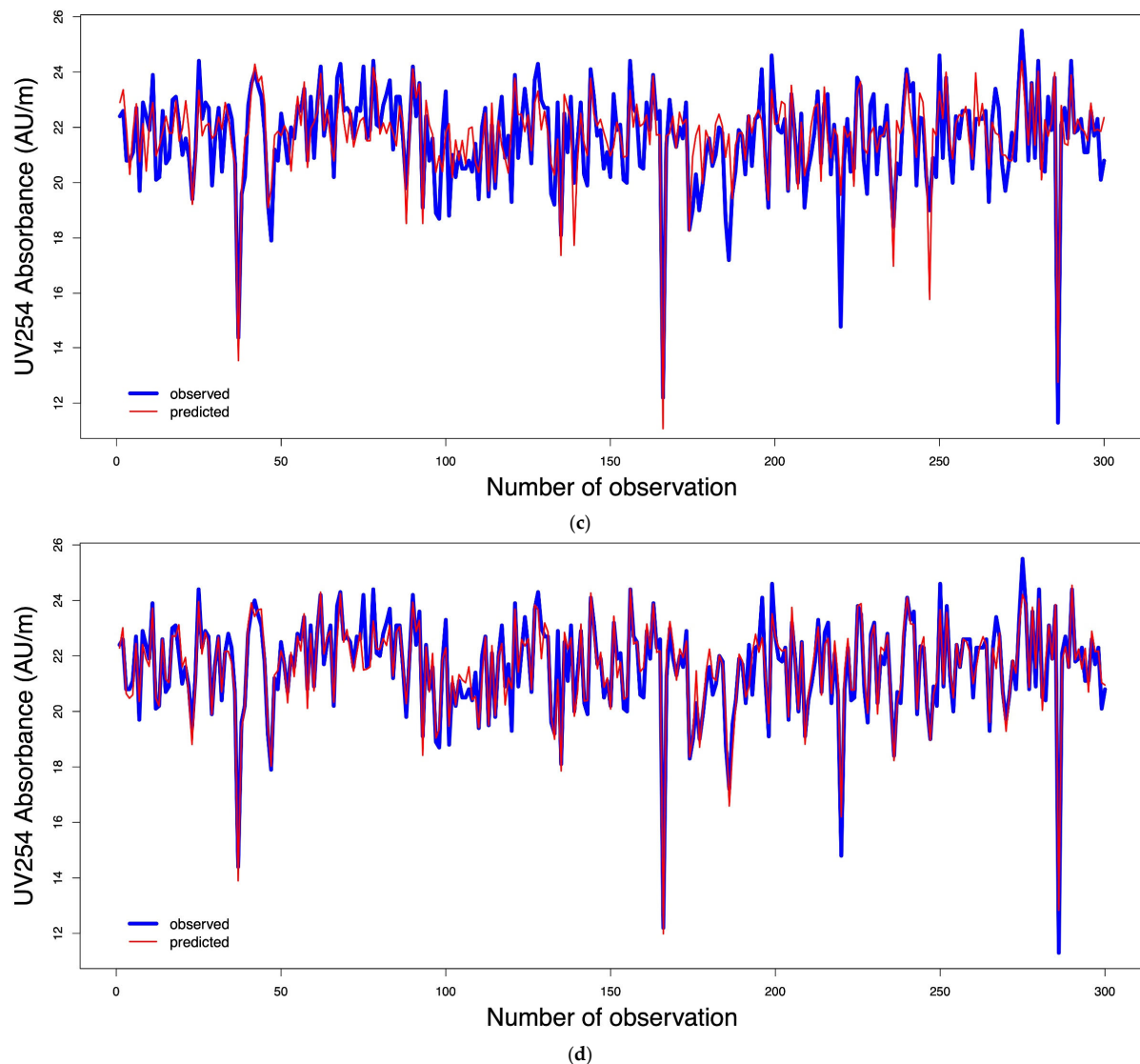


Figure 14. Illustration of the observed vs. predicted absorbance values for 300 random samples of the testing dataset utilizing: (a) M5 model tree ($R^2 = 0.0712$); (b) MLR model ($R^2 = 0.3581$); (c) MLP model ($R^2 = 0.6878$); and (d) NM/GPR-based model with a Matern32 kernel ($R^2 = 0.9180$).

Additionally, Figures 15–17 show the comparison of the experimental and predicted turbidity, conductivity, and absorbance values, respectively, using the best-performing model—the NM/GPR model with the Matern32 kernel across the entire testing dataset. This model performs best. Addressing this nonlinear regression problem required integrating the NM/GPR model with the Matern32 kernel, creating a unique hybrid approach that outperformed the other three regression methods (MLP, MLR, and the M5 model tree) in terms of robustness and efficiency.

Finally, Figure 18 compares the observed and predicted values of turbidity, conductivity, and absorbance indicators for 50 random samples of the testing dataset using the best-fitting NM/GPR model with a Matern32 kernel. The figure shows a 95% confidence interval. These intervals highlight the uncertainty associated with each prediction, a unique feature of GPR that provides valuable insights for water quality management. For example, wider intervals for turbidity at high flow rates (e.g., during storm events) suggest greater uncertainty due to the increased variability in sediment transport. Conversely, narrower intervals for conductivity suggest more stable and predictable ionic contributions from dissolved nutrients, such as nitrates. These uncertainty estimates can be used to

inform decision-making processes, such as setting conservative pollution alert thresholds or prioritizing monitoring efforts during periods of high prediction uncertainty.

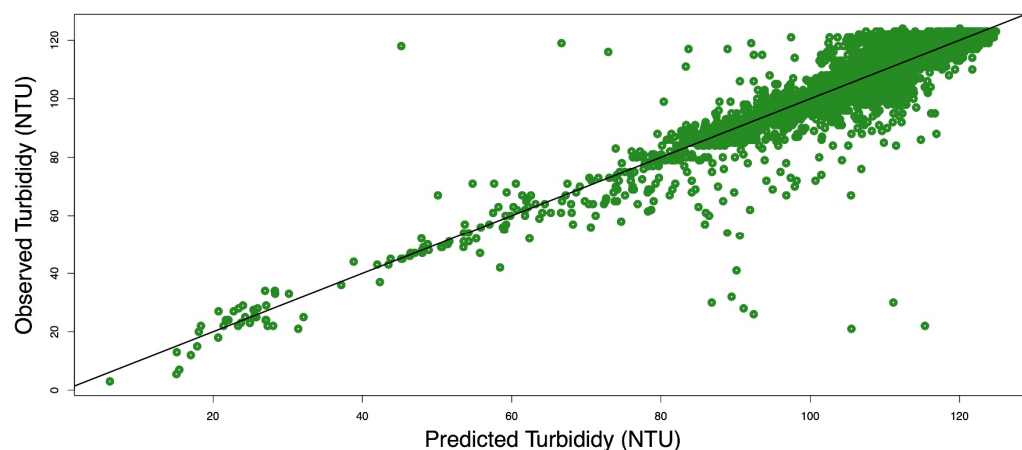


Figure 15. Observed vs. predicted turbidity value scatterplot from the entire testing dataset for the best-fitted model, the NM/GPR model, with the Matern32 kernel.

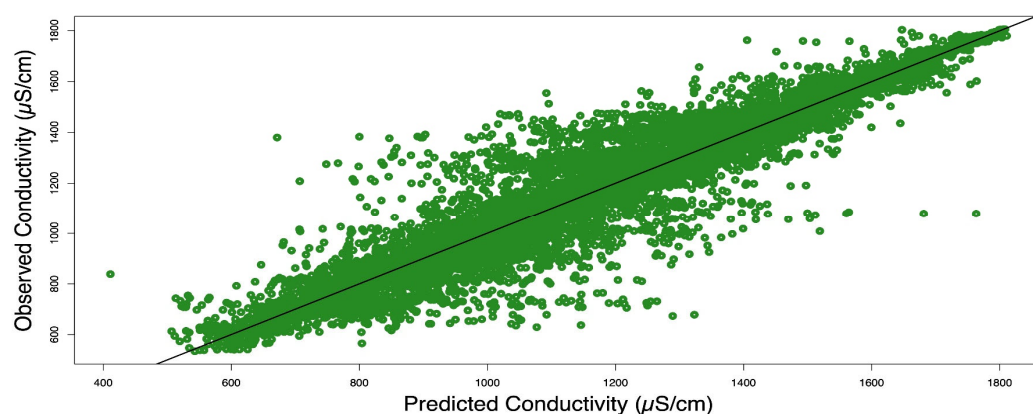


Figure 16. Observed vs. predicted conductivity value scatterplot from the entire testing dataset for the best-fitted model, the NM/GPR model, with the Matern32 kernel.

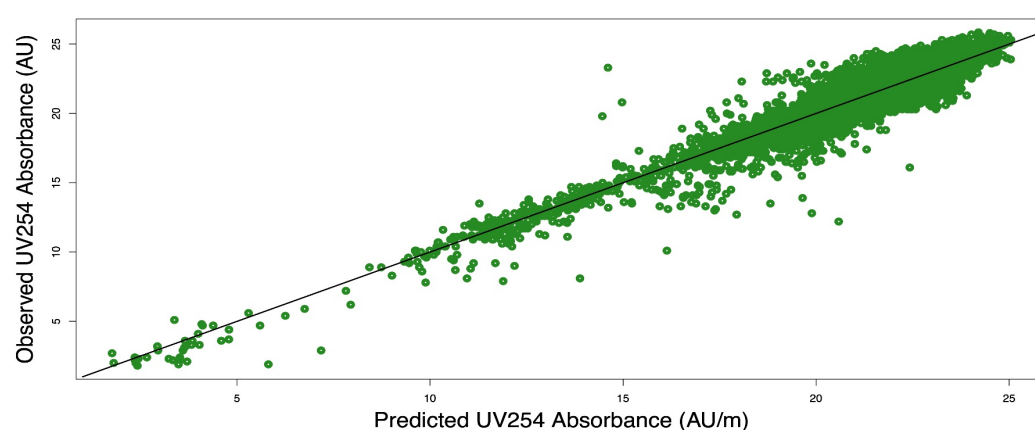


Figure 17. Scatterplot of observed versus predicted absorbance values for the entire testing dataset, using the best-fitting NM/GPR model with the Matern32 kernel.

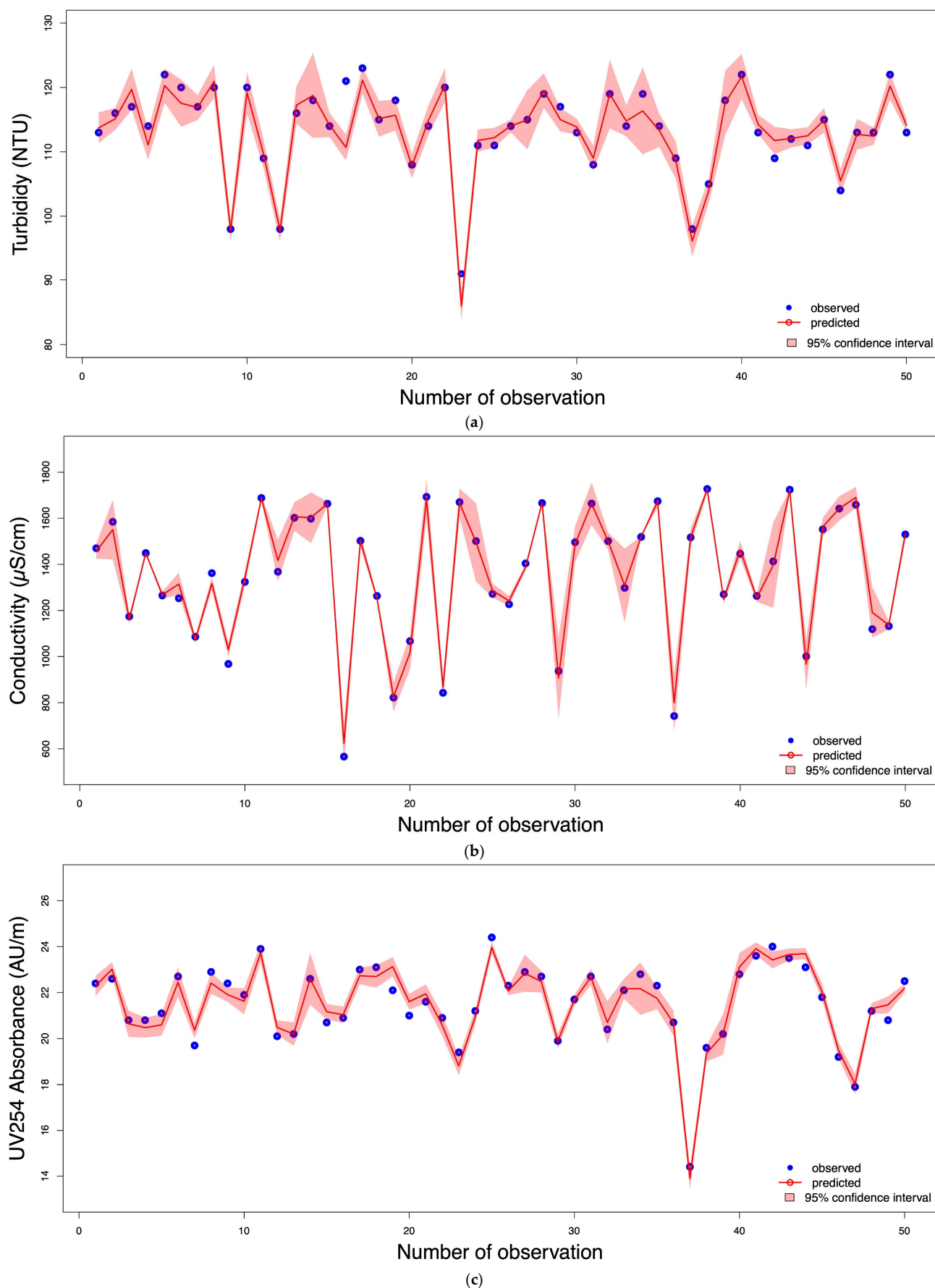


Figure 18. Observed and predicted values with a confidence interval of 95% for 50 random samples in the testing dataset with: (a) turbidity; (b) conductivity; and (c) absorbance.

4. Conclusions

This study shows that Gaussian Process Regression (GPR), when combined with Nelder–Mead optimization and a Matern32 kernel, is an effective method for predicting turbidity, conductivity, and UV254 absorbance in the Ebro River.

The feature importance analysis provides actionable insights for water quality management, and the confidence intervals offer a measure of uncertainty for decision-making. The feature importance analysis revealed the primary drivers of river hydrochemistry. Turbidity is primarily controlled by discharge, which reinforces the importance of hydrological forcing as the main mechanism for suspended sediment transport. Conductivity is predominantly governed by nitrate concentrations, which highlights the significant contribution of dissolved nutrients to the river's ionic strength. UV254 absorbance reflects the coupled influence of discharge and nitrates, suggesting that organic matter dynamics are regulated by both hydrological mobilization and nutrient-associated inputs. Notably, dissolved mercury exhibited negligible explanatory power across all models, indicating its limited role as a proxy for the studied parameters.

The model achieved high predictive accuracy. The experimental datasets for turbidity, conductivity, and absorbance yielded strong NSE values of 0.8956, 0.9503, and 0.9180, respectively, underscoring the model's reliability for high-frequency environmental monitoring.

The GPR+NM framework can be integrated into real-time monitoring systems to enable early detection of contamination events and the adaptive management of water resources. The methodological framework is potentially transferable to other river systems; however, direct application would require site-specific recalibration and validation using local data, as the identified predictor–response relationships may vary depending on hydrological, climatic, and anthropogenic conditions specific to each basin. These findings advance the fields of water quality modeling and environmental monitoring and have potential applications in real-time systems and other river basins.

This study demonstrates that Gaussian Process Regression (GPR) is a highly effective method for modeling high-frequency, nonlinear environmental datasets. The use of the Matern32 kernel is particularly advantageous, as it captures the non-smooth, heterogeneous variability inherent in such data. Additionally, the study emphasizes the critical role of hyperparameter optimization, achieved through the NM algorithm, in ensuring robust and accurate predictions.

Author Contributions: Conceptualization, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; methodology, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; software, P.J.G.-N., E.G.-G., and L.A.M.-G.; validation, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; formal analysis, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; data curation, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; writing—original draft preparation, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; writing—review and editing, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; visualization, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M.; supervision, P.J.G.-N., E.G.-G., L.A.M.-G., J.R.A.F., and C.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The Department of Mathematics at the University of Oviedo provided computational support, for which the authors are grateful.

Conflicts of Interest: The authors declare no competing interests.

References

- Deng, R.; Zhu, T.; Zhou, W.; Liu, F.; Lin, X. Machine Learning Based Water Quality Evolution and Pollution Identification in Reservoir Type Rivers. *Environ. Pollut.* **2025**, *382*, 126668. [[CrossRef](#)] [[PubMed](#)]
- Egbueri, J.C.; Enyigwe, M.T.; Ayejoto, D.A. Modeling the Impact of Potentially Harmful Elements on the Groundwater Quality of a Mining Area (Nigeria) by Integrating NSFQI, HERisk Code, and HCs. *Environ. Monit. Assess.* **2022**, *194*, 150. [[CrossRef](#)] [[PubMed](#)]
- Agbasi, J.C.; Ayejoto, D.A.; Egbueri, J.C.; Khan, N.; Abba, S.I.; Ahmad, V.; Abuzinadah, M.F. HERisk and Statistical Clustering Integrated for Health Risk Modelling of PTEs in Natural Water Resources for Drinking and Sanitary Uses. *Toxin Rev.* **2024**, *43*, 513–539. [[CrossRef](#)]
- Zheng, Y.; Wei, C.; Fu, H.; Li, H.; He, Q.; Yu, D.; Fu, M. Spatial-Temporal Evolution Analysis of Pollutants in Daitou River Watershed Based on Sentinel-2 Satellite Images. *Ecol. Indic.* **2024**, *166*, 112436. [[CrossRef](#)]
- Walling, D.E.; Webb, B.W. The Reliability of Suspended Sediment Load Data. In *Proceedings of the Florence Symposium, Erosion and Sediment Transport Measurement*; IAHA Publ.: Florence, Italy, 1981; pp. 177–194.
- Hem, J.D. *Study and Interpretation of the Chemical Characteristics of Natural Water*; U.S. Geological Survey: Reston, VA, USA, 1985.
- Weishaar, J.L.; Aiken, G.R.; Bergamaschi, B.A.; Fram, M.S.; Fujii, R.; Mopper, K. Evaluation of Specific Ultraviolet Absorbance as an Indicator of the Chemical Composition and Reactivity of Dissolved Organic Carbon. *Environ. Sci. Technol.* **2003**, *37*, 4702–4708. [[CrossRef](#)] [[PubMed](#)]
- De Oliveira, F.G.; Dos Santos, L.D.; Palmeiro, A.S. Assessment of Surface Water Quality Based on Physical and Chemical Parameters in a GIS, for Three Rivers in Southern Brazil. *Environ. Pollut.* **2025**, *375*, 126295. [[CrossRef](#)] [[PubMed](#)]
- Van Vliet, M.T.H.; Jones, E.R.; Flörke, M.; Franssen, W.H.P.; Hanasaki, N.; Wada, Y.; Yearsley, J.R. Global Water Scarcity Including Surface Water Quality and Expansions of Clean Water Technologies. *Environ. Res. Lett.* **2021**, *16*, 024020. [[CrossRef](#)]
- Fida, M.; Li, P.; Wang, Y.; Alam, S.M.K.; Nsabimana, A. Water Contamination and Human Health Risks in Pakistan: A Review. *Expo. Health* **2023**, *15*, 619–639. [[CrossRef](#)]
- Meybeck, M. The Quality of Rivers: From Pristine Stage to Global Pollution. *Glob. Planet. Change* **1989**, *1*, 283–309. [[CrossRef](#)]
- Dodds, W.; Smith, V. Nitrogen, Phosphorus, and Eutrophication in Streams. *Inland Waters* **2016**, *6*, 155–164. [[CrossRef](#)]
- Wetzel, R.G. *Limnology: Lake and River Ecosystems*, 3rd. ed.; Academic Press: San Diego, CA, USA, 2011.
- Anand, V.; Oinam, B.; Wieprecht, S. Machine Learning Approach for Water Quality Predictions Based on Multispectral Satellite Imageries. *Ecol. Inform.* **2024**, *84*, 102868. [[CrossRef](#)]
- Badrudeen, A.T.; Sahoo, D.; Sawyer, C.B.; Pike, J.W.; Harmel, R.D. A Critical Review of Statistical, Signal Processing and Machine Learning Methods for Continuous and High-Frequency Water Quality Data Improvement. *Ecol. Inform.* **2026**, *94*, 103619. [[CrossRef](#)]
- Peterson, J.T.; Freeman, M.C. Integrating Modeling, Monitoring, and Management to Reduce Critical Uncertainties in Water Resource Decision Making. *J. Environ. Manag.* **2016**, *183*, 361–370. [[CrossRef](#)] [[PubMed](#)]
- Hou, Y.; Miao, C.; Zhu, D.; Li, Z.; Du, F.; Wang, W.; Yang, X.; Cao, Z. Flow Mechanism of Grouting Slurry in Rough Fracture Based on CFD-DEM Coupling Method. *Processes* **2026**, *14*, 1307. [[CrossRef](#)]
- Cao, Z.; Miao, C.; Du, F.; Zhu, D.; Teng, T.; Xue, Y. A Bifurcation Dynamical Analysis of a Non-Darcy Seepage System in Post-Failure Rock Based on a Novel Truncated Spectral Method. *Processes* **2026**, *14*, 1468. [[CrossRef](#)]
- Vapnik, V.N. *Statistical Learning Theory*; Adaptive and learning systems for signal processing, communications, and control; Wiley: New York, NY, USA; Weinheim, Germany, 1998.
- Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer series in statistics; Springer: New York, NY, USA, 2003.
- Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; Adaptive computation and machine learning; MIT Press: Cambridge, MA, USA, 2006.
- Schneider, M.; Ertel, W. Robot Learning by Demonstration with Local Gaussian Process Regression. In *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010*; IEEE: New York, NY, USA, 2010; pp. 255–260.
- Shi, J.Q.; Choi, T. *Gaussian Process Regression Analysis for Functional Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2011.
- Ebden, M. Gaussian Processes: A Quick Introduction. *arXiv* **2015**, arXiv:1505.02965.
- Camps-Valls, G.; Verrelst, J.; Munoz-Mari, J.; Laparra, V.; Mateo-Jimenez, F.; Gomez-Dans, J. A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 58–78. [[CrossRef](#)]
- Vidales, A. *Machine Learning with Matlab: Gaussian Process Regression, Analysis of Variance and Bayesian Optimization*; Independently published, 2019.

27. Elkadeem, M.R.; Zayed, M.E.; Kotb, K.M.; Shboul, B.; AlZahrani, A.S.; Rehman, S.; Abido, M.A. Design Optimization and ML-Based Performance Prediction of Microgrid-Hydrogen Refueling Systems Using Gaussian Process Regression. *Int. J. Hydrogen Energy* **2025**, *158*, 150382. [\[CrossRef\]](#)
28. Fogelquist, J.; Lin, X. Combining Electrochemistry and Data-Sparse Gaussian Process Regression for Lithium-Ion Battery Hybrid Modeling. *Appl. Energy* **2025**, *399*, 126458. [\[CrossRef\]](#)
29. Wang, G.; Feng, Y.; Dai, Y.; Chen, Z.; Wu, Y. An Efficient Prediction Method for Wind Loads on Container Ships Based on Numerical Simulation and Gaussian Process Regression. *Ocean Eng.* **2026**, *343*, 123384. [\[CrossRef\]](#)
30. Paul, E.; Kulathunga, G.; Krishnaswamy, H.; Klimchik, A. Springback Estimation in Incremental Sheet Metal Forming Using the Gaussian Process Regression Model. *Manuf. Lett.* **2025**, *44*, 1378–1386. [\[CrossRef\]](#)
31. Zhang, X.; Sun, X.; Lin, Z. Improving Soil Moisture Prediction Using Gaussian Process Regression. *Smart Agr. Technol.* **2025**, *11*, 100905. [\[CrossRef\]](#)
32. Wang, W.; Wang, Q.J.; Nathan, R. Gaussian Process Regression on Multiple Drivers and Attributes for Rapid Prediction of Maximum Flood Inundation Extent and Depth. *J. Hydrol.* **2025**, *649*, 132476. [\[CrossRef\]](#)
33. Agrawal, T. *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*; Apress: Berkeley, CA, USA, 2021.
34. Aggarwal, C.C. *Linear Algebra and Optimization for Machine Learning: A Textbook*; Springer: Cham, Switzerland, 2020.
35. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
36. Chong, E.K.P. *An Introduction to Optimization*, 4th ed.; Wiley: Hoboken, NJ, USA, 2013.
37. Kennedy, J.; Eberhart, R.C.; Shi, Y. *Swarm Intelligence*, 8th ed.; The Morgan Kaufmann series in evolutionary computation; Morgan Kaufmann: San Francisco, CA, USA, 2009.
38. Hassoun, M.H. *Fundamentals of Artificial Neural Networks*; A Bradford book; MIT Press: Cambridge, MA, USA, 1995.
39. Haykin, S.S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice-Hall: Upper Saddle River, NJ, USA, 1999.
40. Fine, T.L. *Feedforward Neural Network Methodology*; Statistics for engineering and information science; Springer: New York, NY, USA, 1999.
41. Shepherd, A.J. *Second-Order Methods for Neural Networks: Fast and Reliable Training Methods for Multi-Layer Perceptrons*; Perspectives in neural computing; Springer: London, UK, 1997.
42. Vang-Mata, R. (Ed.) *Multilayer Perceptrons: Theory and Applications*; Computer science, technology and applications; Nova Science Publishers: Hauppauge, NY, USA, 2020.
43. Weisberg, S. *Applied Linear Regression*, 4th ed.; Wiley series in probability and statistics; Wiley: Hoboken, NJ, USA, 2014.
44. Fox, J.; Fox, J. *Applied Regression Analysis and Generalized Linear Models*, 3rd ed.; SAGE: Los Angeles, CA, USA, 2016.
45. Roback, P.; Legler, J. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models*, 1st ed.; Chapman and Hall Texts in Statistical Science; CRC Press: Boca Raton, FL, USA, 2020.
46. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 6th ed.; Wiley series in probability and statistics; Wiley: Hoboken, NJ, USA, 2021.
47. Stone, J.V. *Linear Regression: A Tutorial Introduction to the Mathematics of Regression Analysis*; Sebtel Press: London, UK, 2022.
48. Quinlan, J.R. Learning with Continuous Classes. In *Proceedings of the Proceedings of 5th Australian Joint Conference on Artificial Intelligence*; World Scientific Press: Singapore, 1992; Volume 92, pp. 343–348.
49. Mitchell, T.M. *Machine Learning*; McGraw-Hill series in Computer Science; McGraw-Hill: New York, NY, USA, 2013.
50. Kisi, O. Pan Evaporation Modeling Using Least Square Support Vector Machine, Multivariate Adaptive Regression Splines and M5 Model Tree. *J. Hydrol.* **2015**, *528*, 312–320. [\[CrossRef\]](#)
51. Solomatine, D.P.; Xue, Y. M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *J. Hydrol. Eng.* **2004**, *9*, 491–501. [\[CrossRef\]](#)
52. Rahimikhoob, A.; Asadi, M.; Mashal, M. A Comparison Between Conventional and M5 Model Tree Methods for Converting Pan Evaporation to Reference Evapotranspiration for Semi-Arid Region. *Water Resour. Manag.* **2013**, *27*, 4815–4826. [\[CrossRef\]](#)
53. Behnood, A.; Olek, J.; Glinicki, M.A. Predicting Modulus Elasticity of Recycled Aggregate Concrete Using M5' Model Tree Algorithm. *Constr. Build. Mater.* **2015**, *94*, 137–147. [\[CrossRef\]](#)
54. Pal, M.; Deswal, S. M5 Model Tree Based Modelling of Reference Evapotranspiration. *Hydrol. Process.* **2009**, *23*, 1437–1443. [\[CrossRef\]](#)
55. Khorrami, R.; Derakhshani, A.; Moayedi, H. New Explicit Formulation for Ultimate Bearing Capacity of Shallow Foundations on Granular Soil Using M5' Model Tree. *Measurement* **2020**, *163*, 108032. [\[CrossRef\]](#)
56. Seghier, M.; Keshtegar, B.; Correia, J.; Jesus, A.D.; Lesiuk, G. Structural Reliability Analysis of Corroded Pipeline Made in X60 Steel Based on M5 Model Tree Algorithm and Monte Carlo Simulation. *Procedia Struct. Integr.* **2018**, *13*, 1670–1675. [\[CrossRef\]](#)
57. Pal, M. M5 Model Tree for Land Cover Classification. *Int. J. Remote Sens.* **2006**, *27*, 825–831. [\[CrossRef\]](#)
58. ISO 7027; Water Quality. Determination of Turbidity. International Organization for Standardization: Geneva, Switzerland, 2016.

59. European Union. *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy*; European Union: Luxembourg, 2000; Volume 327, pp. 1–73.
60. Chen, C.W. Concepts and Utilities of Ecologic Model. *J. Sanit. Eng. Div.* **1970**, *96*, 1085–1097. [\[CrossRef\]](#)
61. Riegl, B.; Glynn, P.W.; Wieters, E.; Purkis, S.; d'Angelo, C.; Wiedenmann, J. Water Column Productivity and Temperature Predict Coral Reef Regeneration across the Indo-Pacific. *Sci. Rep.* **2015**, *5*, 8273. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Smith, V.H. Responses of Estuarine and Coastal Marine Phytoplankton to Nitrogen and Phosphorus Enrichment. *Limnol. Oceanogr.* **2006**, *51*, 377–384. [\[CrossRef\]](#)
63. Marsland, S. *Machine Learning: An Algorithmic Perspective*, 2nd ed.; CRC Press, Taylor and Francis Group: Boca Raton, FL, USA, 2015.
64. García Nieto, P.J.; García-Gonzalo, E.; Paredes-Sánchez, B.M.; Paredes-Sánchez, J.P. Forecast of the Higher Heating Value Based on Proximate Analysis by Using Support Vector Machines and Multilayer Perceptron in Bioenergy Resources. *Fuel* **2022**, *317*, 122824. [\[CrossRef\]](#)
65. Chong, E.K.P.; Lu, W.; Žak, S.H. *An Introduction to Optimization: With Applications to Machine Learning*, 5th ed.; Wiley: Hoboken, NJ, USA, 2024.
66. McClave, J.T.; Sincich, T. *Statistics*, 13th ed.; Pearson: Boston, MA, USA, 2016.
67. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*; Springer Texts in Statistics; Springer: New York, NY, USA, 2004.
68. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, NY, USA, 2006.
69. Freedman, D.; Pisani, R.; Purves, R. *Statistics*, 4th ed.; Norton: New York, NY, USA, 2007.
70. Knalf, G.J.; Ding, K. *Adaptive Regression for Modeling Nonlinear Relationships*; Springer: Berlin, Germany, 2016.
71. Picard, R.R.; Cook, R.D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583. [\[CrossRef\]](#)
72. Efron, B.; Tibshirani, R. Improvements on Cross-Validation: The 632+ Bootstrap Method. *J. Am. Stat. Assoc.* **1997**, *92*, 548. [\[CrossRef\]](#)
73. Piironen, J. *Gplite: General Purpose Gaussian Process Modelling*, version 0.13.0. 2021. Available online: <https://CRAN.R-project.org/package=gplite> (accessed on 25 June 2026).
74. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann series in data management systems; Morgan Kaufmann: Burlington, MA, USA, 2011.
75. Paananen, T.; Piironen, J.; Andersen, M.R.; Vehtari, A. Variable Selection for Gaussian Processes via Sensitivity Analysis of the Posterior Predictive Distribution. *arXiv* **2017**, arXiv:1712.08048. [\[CrossRef\]](#)
76. Jones, J.A.; Grant, G.E. Peak Flow Responses to Clear-Cutting and Roads in Small and Large Basins, Western Cascades, Oregon. *Water Resour. Res.* **1996**, *32*, 959–974. [\[CrossRef\]](#)
77. Wu, W. *Sediment Transport Dynamics*; CRC Press: Boca Raton, FL, USA, 2023.
78. Driscoll, C.T.; Mason, R.P.; Chan, H.M.; Jacob, D.J.; Pirrone, N. Mercury as a Global Pollutant: Sources, Pathways, and Effects. *Environ. Sci. Technol.* **2013**, *47*, 4967–4983. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Stumm, W.; Morgan, J.J. *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*, 3rd ed.; Environmental science and technology; J. Wiley & Sons: New York, NY, USA, 1996.
80. Bridgewater, L.L.; Baird, R.B.; Eaton, A.D.; Rice, E.W.; American Public Health Association; American Water Works Association; Water Environment Federation (Eds.) *Standard Methods for the Examination of Water and Wastewater*, 23rd ed.; American Public Health Association: Washington, DC, USA, 2017.
81. Saraceno, J.F.; Shanley, J.B.; Dewald, J.T.; Chricton, L. Characterization of Dissolved Organic Carbon in Streams Using in Situ Spectrophotometry. *Water Resour. Res.* **2009**, *45*, W11427.
82. Hood, E.; Gooseff, M.N.; Johnson, S.L. Changes in the Character of Stream Water Dissolved Organic Carbon during Flushing in Three Small Watersheds, Oregon. *J. Geophys. Res.* **2006**, *111*, 2005JG000082. [\[CrossRef\]](#)
83. Burns, J.M.; Cooper, W.J.; Ferry, J.L.; King, D.W.; DiMento, B.P.; McNeill, K.; Miller, C.J.; Miller, W.L.; Peake, B.M.; Rusak, S.A.; et al. Methods for Reactive Oxygen Species (ROS) Detection in Aqueous Environments. *Aquat. Sci.* **2012**, *74*, 683–734. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.