Results in Engineering 25 (2025) 104159

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Integrating piecewise and symbolic regression with remote sensing data for spatiotemporal analysis of surface water total dissolved solids in the Karun River, Iran

Javad Zahiri [a,*], Mohammad Reza Nikoo [b,*], Adell Moradi-Sabzkouhi [a], Mitra Cheraghi [c], Nazmi Mat Nawi [d]

[a] Department of Water Engineering, Agricultural Sciences and Natural Resources University of Khuzestan, Iran
[b] Department of Civil and Architectural Engineering, Sultan Qaboos University, Muscat, Oman
[c] Department of Nature Engineering, Agricultural Sciences and Natural Resources University of Khuzestan, Iran
[d] Institute of Plantation Studies, Universiti Putra Malaysia, Malaysia

ARTICLE INFO

Keywords:
Total dissolved solids
Symbolic regression model
M5 model tree
MARS model
Landsat imagery

ABSTRACT

Monitoring water quality, including total dissolved solids (TDS), across various spatial and temporal scales is essential for comprehending water health level. The Karun River's water is a vital resource for drinking water supply and agricultural irrigation, making accurate monitoring of its quality crucial for ensuring public health and sustainable resource management. This study presents a novel hybrid modeling approach that combines piecewise and symbolic regression (SR) models with Landsat imagery to predict TDS levels in river systems. Specific spectral bands, including the red and near-infrared reflectance from Landsat-9, were used as input variables. Field sampling of TDS, as the output variable, was conducted along the Karun River on three dates, synchronized with Landsat satellite images. The innovative use of a fuzzy-based uncertainty analysis, coupled with AHP weighting, allowed for a comprehensive assessment of TDS estimation accuracy and uncertainty. The Composite Uncertainty Index (CUI) approach revealed that the Multivariate Adaptive Regression Splines (MARS) and M5 models performed better than other models, with CUI values of 0.83 and 0.72, respectively. MARS demonstrated higher accuracy under low uncertainty conditions, while M5P excelled in scenarios of elevated uncertainty due to its reduced sensitivity and strong Nash-Sutcliffe coefficient. The hybrid modeling approach presented in this study offers a unique contribution to remote sensing-based water quality monitoring. By leveraging advanced regression techniques and uncertainty quantification, the findings enable more reliable predictions of TDS levels, which are crucial for sustainable river management.

1. Introduction

Growing populations and expanding human activities are increasing the pollution discharged into rivers and other surface water sources [1], leading to a global concern of water quality degradation due to significant anthropogenic pollution loading [2]. Studies have highlighted the impacts of urbanization, industrial discharge, agricultural runoff, and wastewater mismanagement on water bodies worldwide [3,4]. These activities contribute to the escalation of pollutants, including nutrients, heavy metals, and dissolved solids, which adversely affect water quality and ecosystem health. These pollutants not only threaten aquatic ecosystems but also pose risks to human health, particularly in regions where water is used for drinking and irrigation purposes [5]. In recent decades, intensified human activities have exacerbated water quality issues, as reflected in the rising levels of contaminants in rivers and reservoirs [6], making monitoring at finer time intervals crucial for effective water management at the basin scale [7]. This is specifically important in arid and semi-arid regions like Iran, where water scarcity is a major concern [8,9].

Regular monitoring is essential for understanding the health of water bodies and taking timely actions to prevent pollution. Total dissolved solids (TDS) is a key indicator of river water quality, influencing factors such as taste, corrosivity, and suitability for various applications, with high levels indicating pollution or saline intrusion [10]. Therefore,

* Corresponding authors.
E-mail addresses: j.zahiri@asnrukh.ac.ir (J. Zahiri), m.reza@squ.edu.om (M.R. Nikoo).

https://doi.org/10.1016/j.rineng.2025.104159
Received 14 December 2024; Received in revised form 18 January 2025; Accepted 23 January 2025
Available online 23 January 2025

measuring TDS is an essential part of overall water quality assessment. The Karun River, the largest river in Iran, has undergone a serious environmental problem due to the excessive loading of industrial, urban, and agricultural effluents. Sewage discharge raises a river's salinity and increases the risk of severe pollution. The cost of treating this river is minimized at $48.99 million, with an optimal TDS concentration averaging 17.85 % at the loading points [11]. Similar water pollution scenarios happen in many rivers worldwide. While traditional methods for monitoring river water quality are time-consuming and costly, obtaining accurate and relevant data across spatial and temporal scales remains essential for effective river health assessment [12].

In recent years, efforts have been made to simulate TDS levels in surface water using numerical models and data mining techniques. Kadkhodazadeh and Farzin [13] combined the least-squares support vector machine with the arithmetic optimization algorithm to develop a hybrid machine learning model. This model was thoroughly tested on six important benchmark datasets and then used to estimate TDS at six locations in the Karun River Basin. In a different study, Pourhosseini et al. [14] developed several hybrid models to estimate TDS in the Babolrood River in Iran. They utilized monthly measured data for Na, Ca, Cl, Mg, SO4, HCO3, pH, and TDS. They determined the most significant inputs in five distinct scenarios by applying Shannon's entropy and a correlation matrix. Jamei et al. [15] developed a method combining wavelet preprocessing with multigene genetic programming to predict monthly TDS levels in the Sefid Rud River, Northern Iran. They decomposed time series data into sub-series using three different wavelets to identify optimal combinations and lag times for prediction.

Remote sensing technology analyzes and interprets measurements of electromagnetic radiation emitted or reflected by a surface without direct contact using an appropriate viewing point or recording device [16]. Spectral remote sensing reflectance, a primary product of aquatic color remote sensing, provides data on the optical characteristics of surface water components. Satellite reflectance is obtained by removing atmospheric influences from top-of-atmosphere (TOA) reflectance or radiance measurements captured by remote sensors [17]. Using satellite imagery to assess water quality allows for synoptic and cost-effective assessments. Consequently, satellite remote sensing offers a quick and economical method for assessing quality indicators in oceans, seas, and rivers. To estimate these indicators from the emission and reflection of radiation from the water surface, it is crucial to theoretically consider the relationship between radiation transmission, the optical properties of water, and the radiation detected by the sensor. This transmission relationship is modeled using statistical correlations between radiometric data and field measurements after removing atmospheric effects [18]. In recent years, several remote sensing (RS)-based models have been proposed for retrieving total suspended solids (TSS) data from inland waters, including empirical models and bio-optical or analytical approaches [19]. Both empirical and bio-optical models use RS reflectance obtained by satellites at various wavelengths, either in the form of a single band or a combination of bands. Numerous studies have demonstrated a significant relationship between satellite imagery radiance and TSS [20,21,22,23]. However, limited research has been conducted on estimating TDS based on remote sensing data.

In addition to statistical indicators, the uncertainty degree is significant for qualifying the performance of different simulation methods [24]. Uncertainty refers to the error in measuring or estimating a parameter, influenced by factors such as human error, inaccuracies in measuring devices, or the methods used, and is defined as the occurrence of phenomena beyond human control [25]. In other words, uncertainty, attributed to the lack of accurate information about a phenomenon, process, or data involved in problem analysis [26], affects system responses and can introduce variability. The uncertainty associated with the independent input parameters of each TDS estimation model contributes to uncertainty in the final TDS estimates. Significant differences often arise between the results of various studies, even when only a few parameters are considered as performance criteria. Evaluating the efficiency of different methods becomes more complex when multiple criteria are used. In such cases, multi-criteria decision--making methods can be employed to assess the effectiveness of the various methods. Shi et al. [27] utilized the Analytic Hierarchy Process (AHP) to assign weights to different uncertainty indicators and to compute the Composite Uncertainty Index (CUI), which was used to compare the levels of uncertainty in hydrological models.

This study aims to develop and apply two piecewise models, M5P and MARS, alongside a symbolic regression (SR) model for predicting TDS levels in river systems using Landsat imagery. The primary novelty of this approach lies in the integration of these advanced regression models with remote sensing data, enabling accurate and reliable TDS estimations. Furthermore, the study introduces a novel methodology for quantifying uncertainty in TDS predictions by incorporating a fuzzy-based interval analysis. This dual approach not only enhances the predictive accuracy but also provides a framework for assessing model uncertainty. The study addresses knowledge gaps in satellite-based TDS prediction models, as well as the uncertainty associated with these models in accurately estimating TDS in river systems.

2. Materials and methods

2.1. Study area

Karun River, with an annual flow of about 22 billion m^3 and an average discharge rate of 736 m^3/s, was selected as the study area. The river basin covers an area of 66,352 square kilometers, with an average elevation of 1537 m and an average slope of 0.3 %. In the mountainous regions, the river's width ranges from 25 to 40 m, while in the upstream plains, it widens to between 250 and 400 m. The Karun River is formed by merging the Dez, Shoteyt, and Gargar Rivers in a region known as Band-e Qir, located upstream of Molasani city [28]. This river has poor water quality due to sewage discharge and agricultural runoff. The further the river flows downstream, the more water quality deteriorates. According to the Khuzestan Water and Power Authority (KWPA), the annual average of TDS is about 1300 mg/L at the Molasani hydrometric station, located upstream of the study area, and approximately 1500 mg/L at the Ahvaz hydrometric station, located downstream of the study area. The high TDS level of the river between Molasani and Ahvaz is attributed to the numerous agricultural lands in this area and the discharge of saline effluents into the Karun River. TDS sampling in this study was conducted along a 4-kilometer stretch of the Karun River downstream of the Molasani hydrometric station on three dates that coincided with Landsat satellite imagery: April 6, 2024, May 8, 2024, and May 24, 2024. The Karun River and the sampling sites are illustrated in Fig. 1.

2.2. Data processing

In this study, all necessary Landsat images were obtained from the LANDSAT/LC09/C02/T1_L2 Image Collection available in Google Earth Engine. These images provide TOA reflectance data, with calibration coefficients included in the metadata [29]. Landsat-OLI (Operational Land Imager) or Landsat 9 products (Level 2, Collection 2, Tier 1) with a spatial resolution of 30 m and a temporal resolution of 16 days, corresponding to path and row numbers 165 and 38, were utilized. This data set provides information derived from the data produced by the Landsat 9 OLI/TIRS sensors about Earth's surface, corrected for atmospheric interference. Level-2 data is derived from Level-1 observations provided by the USGS, which are corrected for atmospheric effects such as aerosols, ozone, and water vapor [30].

The red band (band 4, wavelength = 0.636–0.673 μm), the near-infrared band (band 5, wavelength = 0.851–0.879 μm), and the ratio of the red to near-infrared bands from Landsat-9 OLI-2/TIRS-2 Collection 2 were utilized to train and validate data-driven models. Band ratios are widely used to emphasize differences that may not be apparent in

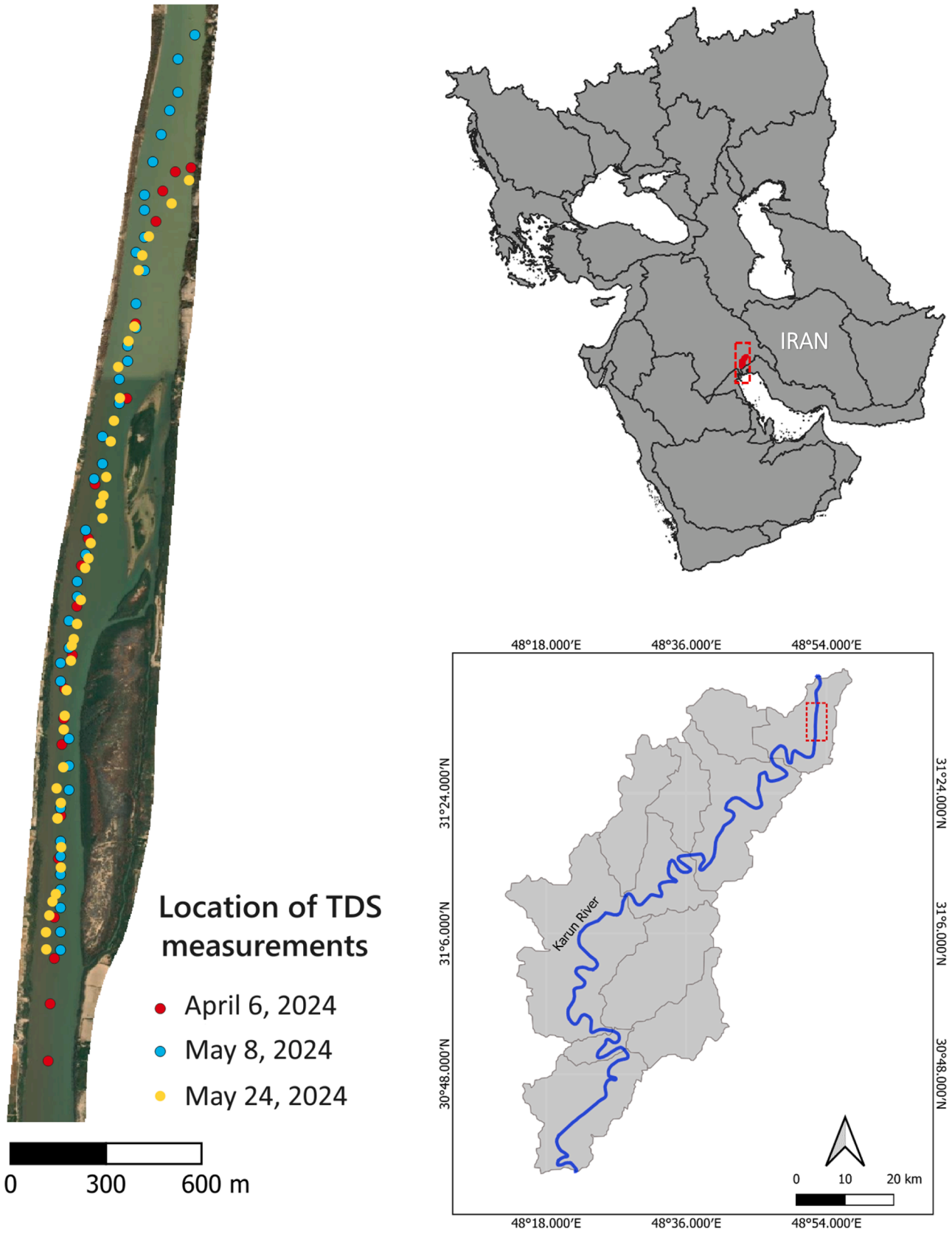


Fig. 1. Location of study area and sampling points in Karun River in Iran created by QGIS.

individual spectral bands [28]. The red and near-infrared reflectance band images from Landsat-9 Collection 2 for the sampling date of May 24, 2024, are displayed in Fig. 2, along with a Google Earth Pro image showing the sampling locations. Table 1 presents a statistical summary of TDS and reflectance band data, including minimum, maximum, and mean values. Additionally, descriptive statistics such as standard deviation and coefficient of variation are provided to assess data variability.

Table 1
Statistics of in-situ TDS measurements and corresponding red and near-infrared reflectance bands with standard deviation (SD) and coefficient of variation (CV).

Parameter	Units	Min	Max	Mean	SD	CV (%)
TDS	mg/L	1332	1974	1688	177	10.46
Red reflectance band	–	0.07	0.17	0.14	0.03	23.27
Near-infrared reflectance band	–	0.02	0.22	0.07	0.03	43.03

2.2.1. Google Earth Engine platform

Google Earth Engine (GEE) is a cloud-based platform for analyzing geospatial data. It provides a massive collection of satellite imagery already processed, including data from Landsat, MODIS, and Sentinel satellites [31]. The geo-referenced nature of this data is significant as it allows for easy inclusion into analysis. In addition to its extensive data repository, GEE allows users to upload and process their own data alongside the available datasets. The GEE computing engine offers APIs for Python and JavaScript, enabling the creation of algorithms that run concurrently on Google's data server infrastructure. The programming model is object-oriented based on the MapReduce paradigm [32]. The GEE is accessible through a web-based integrated development environment (IDE) that utilizes the JavaScript API. This IDE enables users to view images, results, tables, and charts, all of which can be conveniently exported. Alternatively, the Python API provides similar methods for requesting the Engine and accessing the catalog but does not include the visualization features available in the web-based IDE [33]. In this study, the web-based IDE was used to process the Landsat images.

The Level 2 Pixel Quality Assessment band (QA_PIXEL) is generated using data from the Level 1 Quality Assessment band, which includes Cloud Confidence, Cloud Shadow, and Snow/Ice flags derived from the CFMask algorithm. Water values are recalculated and high-confidence cloud pixels are expanded to allow Level 2 data to be utilized for scientific products, making QA_PIXEL similar to the older CFMask bands [34]. In this study, QA_PIXEL was used along with the saturation band (QA_RADSAT) to mask clouds and cloud shadows, generating a median cloud-free composite. Analysis of the Landsat images revealed that the QA_PIXEL band was 2192 at all sampling points, indicating a

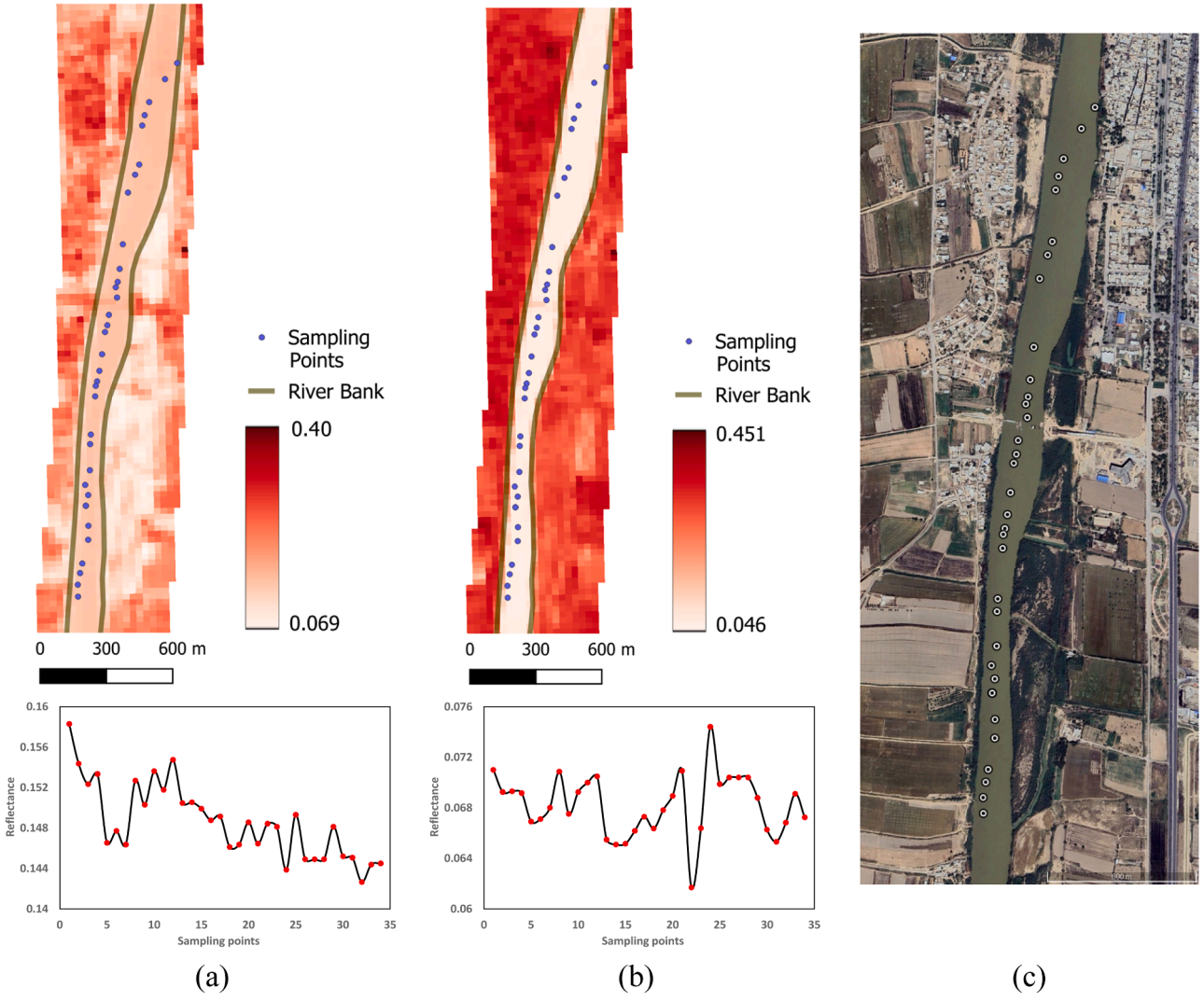


Fig. 2. **a)** The red and **b)** near-infrared reflectance bands of Landsat-9 Collection 2 for the May 24, 2024, sampling date by QGIS (The blue points represent sampling points), and **c)** a Google Earth Pro image accessed on May 2024, with sampling locations.

high-confidence pixel, likely representing clear or open water. However, only the initial sampling points near the riverbank had a QA_PIXEL of 2184, suggesting a pixel with high confidence of clear sky. This difference is because part of the Landsat satellite image pixel overlaps with land.

2.3. Calibrating and validating models

In situ water samples for TDS measurements were collected periodically along the river (Fig. 2) to calibrate and validate the developed RS-based algorithms. The sampling dates were synchronized with the local satellite overpass, matching the Landsat revisit period. Sampling was conducted near the middle of the river using a boat to avoid interference from the reflectance of the riparian habitat. Two hundred fifty mL of river water were collected in polypropylene bottles from each grid location and brought to the laboratory for TDS concentration estimation using the gravimetric method. The standard method 2540 C from the American Public Health Association was used for estimating TDS in the samples [35]. The TDS value in water is attributed to dissolved salts and minerals, typically in the form of ions such as sodium, potassium, carbonates, and sulfates. Environmental laboratories determine gravimetric TDS by filtering a water sample (1–2 μm pore size), evaporating the filtrate at 180 °C, and weighing the remaining residue. This includes ionic, molecular, colloidal substances and some fine clay particles [35]. It is important to note that the TDS concentrations measured in the river during satellite passage dates represent instantaneous values reflecting daily conditions rather than daily averages. Similarly, river engineers typically measure water quality parameters at specific moments during the day, which may not accurately reflect the daily average [12].

This study compared the performance of four TDS prediction models: two piecewise models (M5P and MARS), a symbolic regression (SR) model, and a nonlinear regression (NLR) model.

2.3.1. M5P model

Introduced by Quinlan [36], the M5 model tree divides the problem space into smaller subspaces, constructing an inverted decision tree and generating a linear regression equation for each subspace [37]. A key benefit of the M5P tree algorithm lies in its linear structure, which results in a slower rate of error growth compared to nonlinear models. This model is also known for its ease of development, computational efficiency, and strong prediction reliability. Additionally, it establishes clear, deterministic relationships between dataset parameters, while offering the flexibility to adjust and adapt to changes in the input data. These features make the M5P model particularly advantageous in various applications [38]. Ensemble M5P models, using methods like Bagging, Dagging, Additive Regression, Voting, Iterative Classifier Optimization, Random Subspace, and Rotation Forest, consistently outperformed a stand-alone M5P model in simulating river flow, as demonstrated by statistical analysis of accuracy and efficiency [39]. M5P is optimized for continuous data but may face challenges with categorical variables and risks overfitting or excessive tree growth, which can reduce interpretability if not carefully pruned or regularized [40]. The M5P model involves three phases: building, pruning, and smoothing. To split the problem space, M5P uses the standard deviation reduction (SDR) index:

$$SDR = sd(T) - \sum \left|\frac{T_i}{T}\right| sd(T_i) \quad (1)$$

where T represents a set of data points. T_i is a data vector obtained by dividing the space based on the selected factor, and sd denotes the standard deviation [41]. This branching process continues at each node until the standard deviation at the leaf node reaches zero. Fig. 3 demonstrates how the M5P algorithm splits the input space and derives knowledge from the process.

2.3.2. Multivariate adaptive regression splines (MARS)

Friedman [43] developed the multivariate adaptive regression splines (MARS) model to identify complex relationships between various predictor and dependent variables without requiring prior assumptions about these relationships. The MARS model constructs flexible models by fitting piecewise linear regressions to data, effectively capturing complex, nonlinear relationships without requiring predefined assumptions about the functional form between input variables and the output. The end points of the segments are called knots. A knot marks the end of one region of data and the beginning of another. The resulting piecewise curves (known as basis functions), give greater flexibility to the model, allowing for bends, thresholds, and other departures from linear functions. MARS generates basis functions by searching in a stepwise manner. An adaptive regression algorithm is used to select the knot locations. MARS models are constructed using a two-phase procedure. The forward phase adds functions and finds potential knots to improve the performance, resulting in an overfit model. The backward phase involves pruning the least effective terms. Let y represent the target variable and $X = (X_1, \ldots, X_P)$ be a matrix of P input variables. For a continuous response, this can be expressed as:

$$y = f(X_1, \ldots, X_P) + e = f(X) + e \quad (2)$$

Where e denotes the error distribution. The MARS method approximates f by using basis functions (BFs). These BFs are constructed as splines, which are smooth polynomial functions. MARS can use either piecewise linear or cubic splines. Piecewise linear functions are defined as $max\ (0, x - t)$, where t represents a knot. The function $max\ (.)$ ensures that only the positive component is retained, with all other values assigned zero.

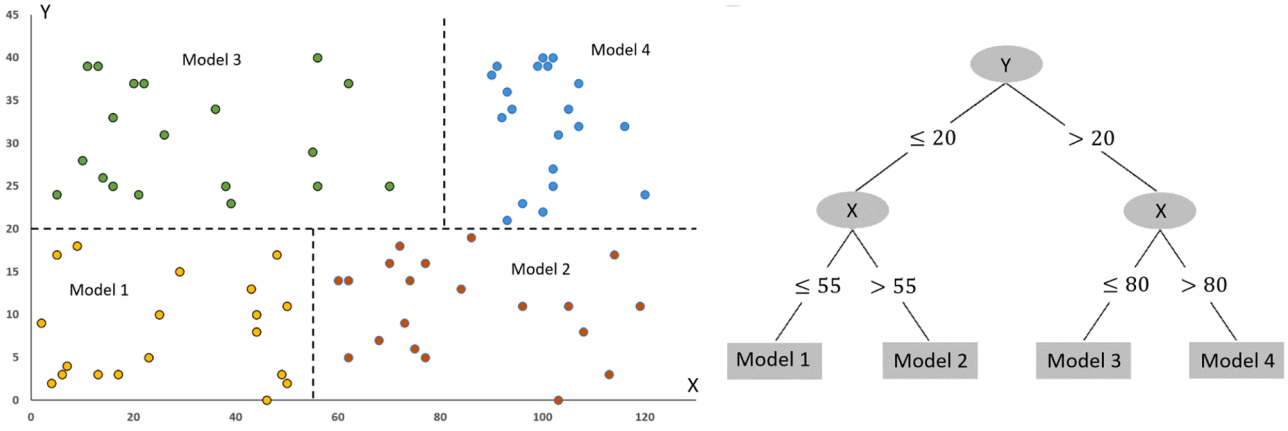


Fig. 3. Partitioning the input domain and tree structure of M5P algorithm [42].

$$\max(0,\ x-t) = \left\{ \begin{array}{l} x-t \text{ if } x \geq t \\ 0 \text{ otherwise} \end{array} \right\} \quad (3)$$

The MARS model, $f(X)$, is represented as a linear combination of BFs and their interactions:

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(X) \quad (4)$$

Here, $\lambda_m(X)$ refers to the m-th basis function, which can either be a single spline function or a product of multiple spline functions already present in the model. Interactions of higher orders can only be included if supported by sufficient data. Typically, MARS restricts interactions to second-order for simplicity and predictive accuracy. The coefficient β_0 is a constant, while β_m represents the coefficient for the m-th basis function, estimated using the least-squares method [44].

MARS model offers a powerful alternative to linear regression by modeling complex, non-linear relationships without predefined assumptions. Automatic variable selection simplifies model building and reduces overfitting, particularly with high-dimensional data. Smoothing splines further aid interpretation of variable interactions [45]. While MARS efficiently handles numerous predictors and automatically detects interactions, its complexity can lead to overfitting and hinder interpretability compared to simpler methods [46].

2.3.3. Symbolic regression

Symbolic Regression (SR) is a field of machine learning that seeks to discover a mathematical equation, represented by a formula, that best fits a given dataset [47]. Parameter optimization and selection for the SR model is crucial for finding expressions that accurately model the underlying data while avoiding overfitting. Unlike traditional machine learning models with fixed functional forms, SR searches a vast space of possible mathematical expressions, making the optimization process more complex. In traditional regression, the task involves identifying a set of numerical coefficients for a combination of independent variable (s) that minimizes a chosen error metric, such as the square root of the sum of squared differences, between observed and predicted values of the dependent variable(s). The type of regression function such as linear, quadratic, or higher-order polynomial regression, or to model the data with non-polynomial functions, such as trigonometric functions like sines and cosines, must be selected. However, the challenge often lies in selecting the most suitable type of function to represent the data, not just calculating the coefficients after choosing a function type. Identifying both the appropriate functional form that best fits the data and the corresponding numerical coefficients is known as symbolic regression [48].

As the most popular method for Symbolic Regression, Genetic Programming (GP) generally outperforms other approaches, especially with real-world problems lacking known models or causal relationships [49]. However, challenges persist, including "bloat," or excessive growth of solution complexity. Standard GP operators, like crossover and mutation, offer limited localized search compared to other numerical optimization heuristics, sometimes disrupting progress. Moreover, GP's stochastic nature prevents a guarantee of finding the absolute optimal solution [50]. Like other evolutionary algorithms, GP begins with a population of randomly generated solutions, commonly created using the Ramped half-and-half method to ensure structural diversity. It iterates through reproduction, mutation, and selection processes, with solutions represented as expression trees. The reproduction process combines beneficial components of two or more solutions to produce a new, potentially improved solution. This method works effectively when parts of a solution correspond to subsets of the original problem. The mutation operator introduces small changes to a solution, helping to avoid local optima. This operator is particularly useful when minor adjustments to a solution do not significantly alter its fitness [51]. GP can create various symbolic expressions by combining these functions and terminals in different ways. The final structure resembles a tree with a root node and branches extending from it. For instance, Fig. 4 shows a tree representing specific expressions (S_1, S_2, S_3 and S_4). The numbers on the nodes are for reference only and don't affect the outcome. Nodes labeled 1 and 2 in expression S_1 represent functions (division and subtraction), while nodes 3, 4, and 5 contain constants and variables [52]. These tree structures are not only expressions but also individuals within a population used by the GP algorithm. The population contains many such individuals, allowing GP to explore different solutions through its evolutionary process.

2.4. Proposed approach framework

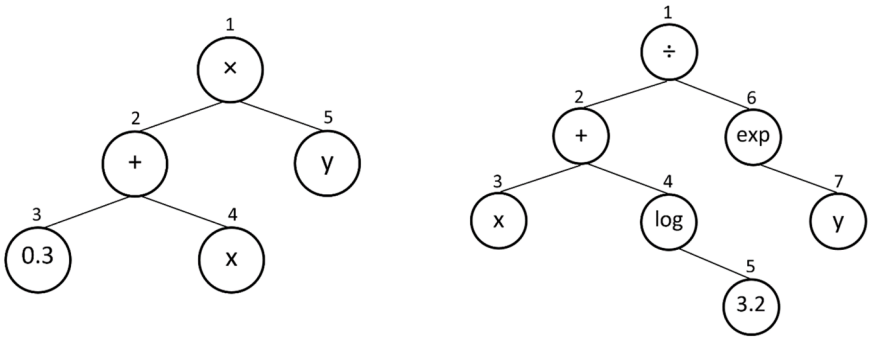
Errors in remote sensing reflectance and its derived products are primarily attributed to uncertainties in atmospheric correction. These uncertainties arise from factors such as highly turbid or eutrophic waters, complex aerosol compositions, proximity to land or clouds, cloud shadows, and the presence of thin clouds [53].

The fuzzy method to apply uncertainty includes two major steps: In the first step, called Input Parameters' Fuzzification (detailed in Section 2.4.1), the fuzzy uncertainty of the input parameters was simplified by defining and applying a variation interval to their values. In the second step, called Output Parameters' Fuzzification, after developing TDS estimation models using remote sensing input data, the lower and upper limits of TDS variation for the estimation methods were determined using the first-order Taylor series expansion method (detailed in Section 2.4.2). By applying these two steps, the uncertainty and accuracy indices of the different estimation methods were calculated based on the boundary limits (detailed in Section 2.4.3). According to the diversity of accuracy and uncertainty indicators used, the AHP hierarchical weighting method was applied to determine the weight of each criterion. Finally, the efficiency of TDS estimation methods was prioritized using a multi-criteria decision-making approach. The study's process is presented in Fig. 5.

2.4.1. Input parameters' fuzzification: applying uncertainty on input parameters

Fuzzy set theory provides a flexible tool for describing vague systems and has been extensively used for uncertainty analysis of hydrosystems [54]. Fuzzy set theory considers a multidegree membership in the range of [0,1] for fuzzy set elements. In other words, for an arbitrary fuzzy variable R, a range of variation is assumed in which the degree of membership $\mu\ (R)$ changes over the interval [0,1]. According to Fig. 6, consider a vague system with an input parameter R and an output response TDS. Corresponding to the deterministic value of the input parameter (i.e., R_c in the traditional approach for TDS estimation; see Fig. 6a), in the fuzzy analysis approach, the input parameter becomes a fuzzy number R whose degree of membership varies from 0 to 1. Let us imagine that the fuzzy membership function of R has the simplest triangular shape with three characteristic values, R_c (the crisp, the most likely value of the parameter when the uncertainty associated with the system is of its lowest degree, i.e., $\mu = 1$), and $(R_c - \Delta R_1)$ and $(R_c + \Delta R_2)$ are called the supports of fuzzy numbers R. In simple terms, $(R_c - \Delta R_1)$ and $(R_c + \Delta R_2)$ represent the most uncertain values expected for R when the system's uncertainty is at its maximum, i.e., when $\mu = 0$. The membership degree (i.e., μ) of a fuzzy number is an effective criterion for expressing the uncertainty associated with vague parameters.

According to Fig. 6b, the fuzzy alpha-cut operator (α) is applied to the membership function, discretizing it into different levels. At the uncertainty level $\mu = \alpha$, the input parameter R exhibits the highest uncertainty, potentially varying over the interval $[R^{a,\alpha}, R^{b,\alpha}]$. Considering the varying interval $[R^{a,\alpha}, R^{b,\alpha}]$ instead of the crisp value R_c for the input parameter R, as shown in Fig. 6c, the output response TDS will also be fuzzy. In this case, TDS is expected to vary over the interval $[TSS^{a,\alpha}, TSS^{b,\alpha}]$. This implies that in a fuzzy system, the uncertainty associated



$$R_1 = (0.3 + x) \times y \qquad R_2 = (\log(3.2) + x)/\, exp(y)$$

Fig. 4. Typical structure of symbolic expression [52].

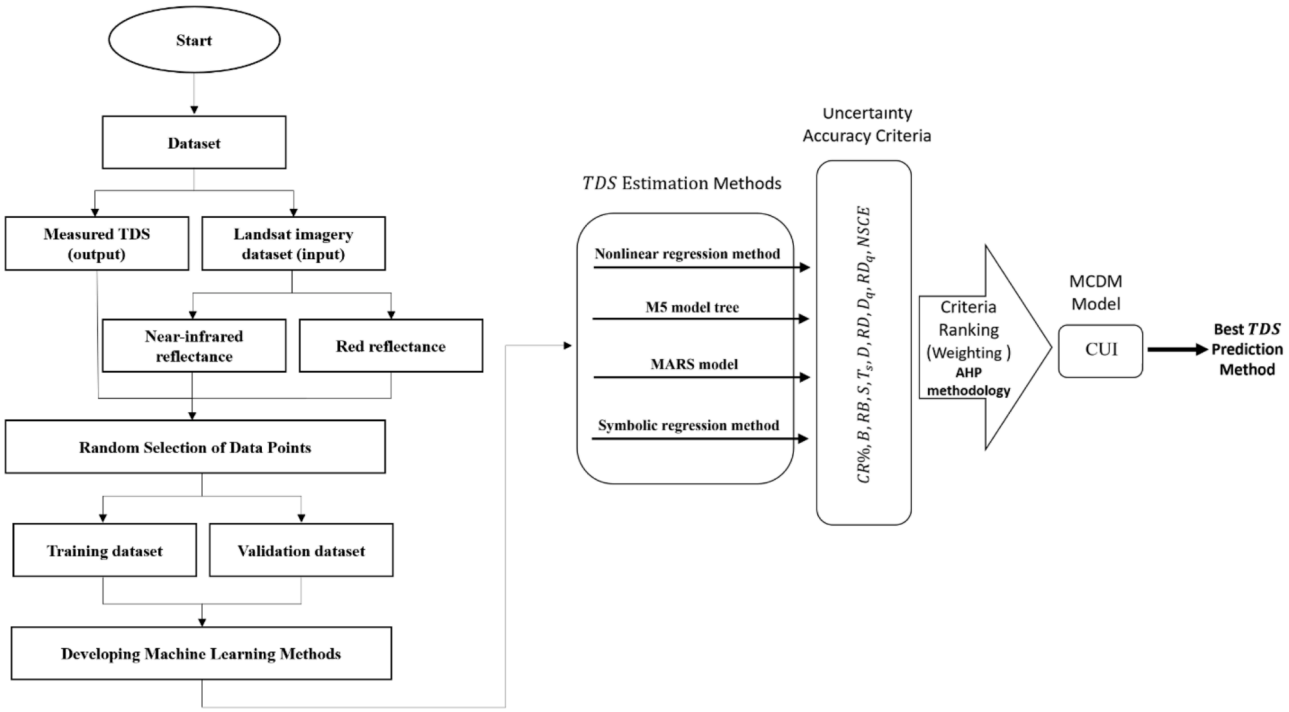


Fig. 5. Comprehensive study workflow and methodology flowchart.

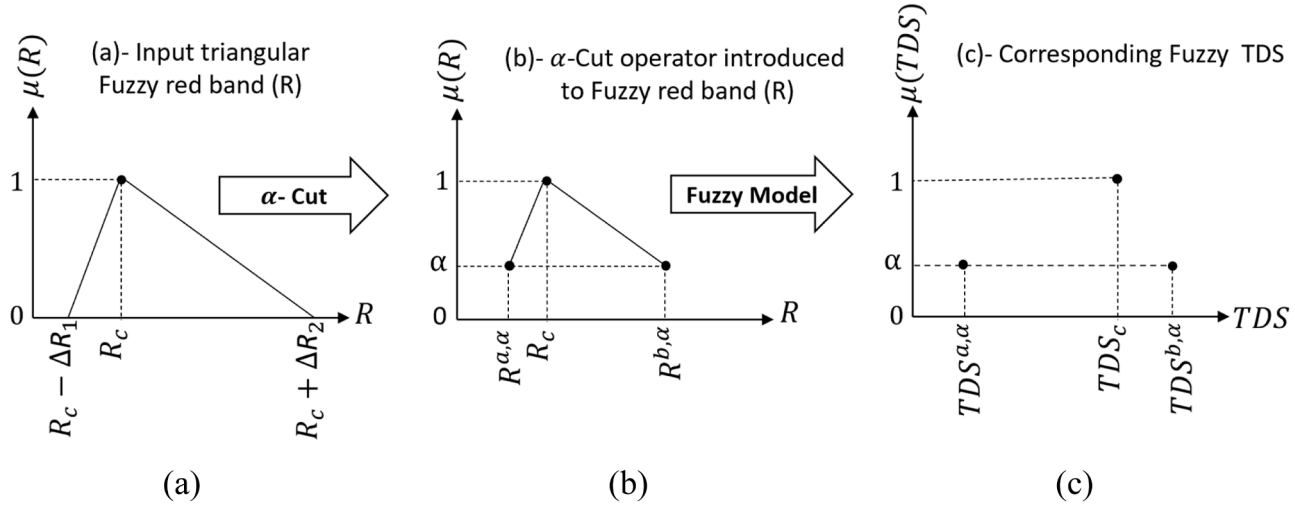


Fig. 6. Conceptual model of fuzzy input-fuzzy output.

with input parameters propagates through the system, causing the output responses to also be uncertain. The greater the input uncertainty, the greater the expected output uncertainty. By introducing different levels of α from 0 to 1, the membership function for the output parameter TDS is obtained, and uncertainty analysis is performed. In fuzzy uncertainty analysis, each level of alpha-cut defines an interval X, making each alpha-cut level equivalent to an interval analysis (IA). Various methods exist to solve an IA problem, such as mathematical optimization using the generalized reduction gradient method [55], metaheuristic optimization [56], and first-order Taylor series expansion [57]. In the present study, the first-order Taylor series expansion was chosen to solve the IA problem resulting from the uncertainty analysis.

2.4.2. Output parameters' fuzzification: interval analysis using first-order Taylor series expansion

If the variable TDS is a multivariate function of the variables $a = (a_1, a_2, \ldots, a_i, \ldots a_{np})$ such that $TDS = \varphi(a)$, then based on the first-order approximation of the Taylor series expansion, the approximation of TDS in the vicinity of a known point a^c where $a_i, a_i^c \in [\underline{a_i}, \overline{a_i}]$ would be [57]:

$$TDS = \varphi(a^c) + (a - a^c)\nabla\varphi(a) \tag{5}$$

$$\Rightarrow TDS = \varphi(a^c) + \sum_{i=1}^{np} \left(a_i - a_i^c\right) \left.\frac{\partial\varphi}{\partial a_i}\right|_{a_i^c} \tag{6}$$

Accordingly, for the estimation of the upper and lower bounds of TDS, we infer:

$$\underline{TDS} = \varphi(a^c) - \sum_{i=1}^{np} \left|\Delta a_i \left(\frac{\partial\varphi}{\partial a_i}\right)_{a_i = a^c}\right| \tag{7}$$

$$\overline{TDS} = \varphi(a^c) + \sum_{i=1}^{np} \left|\Delta a_i \left(\frac{\partial\varphi}{\partial a_i}\right)_{a_i = a^c}\right| \tag{8}$$

where $\Delta a_i = \overline{a}_i - a_i^c = a_i^c - \underline{a}_i$.

By coding in the MATLAB environment and implementing the first-order Taylor series expansion, the lower and upper bounds of variations in the TDS (i.e., the interval of uncertainty TDS) are obtained for each method based on the presumptive uncertainty in the input parameters.

2.4.3. Uncertainty-accuracy indices

In this study, the probability bands of computational TDS were obtained with $a \pm 10$ % uncertainty in the independent parameters used for estimating TDS. This was done to evaluate the performance of different methods using a combination of uncertainty and accuracy indices. An interval analysis approach was employed to assess the performance of various methods and determine the uncertainty associated with different TDS estimation methods (NLR, M5P, MARS, and SR). Since it is necessary to assume the same degree of uncertainty in the input parameters for all methods, an estimator method approximating a relatively smaller bandwidth (i.e., less uncertainty) for TDS is considered more suitable for uncertainty propagation. To compare the efficiency of these estimator methods, both uncertainty indices and accuracy indices must be taken into account. For this purpose, a set of statistical criteria representing accuracy and uncertainty indices was used to compare, evaluate, and select the most suitable methods. The accuracy and uncertainty indices used in this study are presented in Table 2.

2.5. The efficiency of TDS estimation methods

This study employed ten different indices related to uncertainty and accuracy to evaluate the effectiveness of the estimator methods (Table 2). Assessing methods based on ten criteria can complicate their prioritization for practical application. To address this, a multi-criteria decision-making approach, specifically the CUI, was used to evaluate the effectiveness of the estimation methods.

First, the AHP approach was applied to determine the weight of each accuracy and uncertainty index based on the importance of each criterion. The impact of uncertainty parameters on estimator efficiency varies; for example, higher Nash–Sutcliffe coefficient values indicate greater model accuracy, while model efficiency tends to decrease with increasing bandwidth. To balance the effects of different accuracy and uncertainty criteria, a normalization strategy was applied, following the methods proposed by Ibáñez-Forés et al. [58], Ma et al. [59], Shi et al. [27], and Zou et al. [60]. Eqs. (9) and (10) were used to normalize parameters with positive and negative impacts on method performance, respectively. After normalization, all parameters had an impact factor ranging from 0 to 1.

$$r_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \quad i = 1,\ 2,\ \cdots,\ n;\ j = 1,\ 2,\ \cdots,\ m \tag{9}$$

$$r_{ij} = \frac{x_{ij}^{max} - x_{ij}}{x_j^{max} - x_j^{min}} \quad i = 1,\ 2,\ \cdots,\ n;\ j = 1,\ 2,\ \cdots,\ m \tag{10}$$

x_{ij}^{max} و x_{ij}^{min} were estimated based on the following equations:

Table 2
Different accuracy and uncertainty criteria used in this study [27].

Properties	Indices	Formula
Coverage	Containing ratio (CR)	$CR = n_c/N$
Band-width	Average band width (B)	$B = 1/N\sum_{i=1}^{N}\left(k_i^u - k_i^l\right)$
	Average relative bandwidth (RB)	$RB = 1/N\sum_{i=1}^{N}\left[\left(k_i^u - k_i^l\right)/K_i\right]$
Symmetry	Average asymmetry degree 1 (S)	$S = 1/N\sum_{i=1}^{N}\left\|\left(k_i^u - K_i\right)/\left(k_i^u - k_i^l\right) - 0.5\right\|$
	Average asymmetry degree 2 (Ts)	$Ts = 1/N\sum_{i=1}^{N}\left[\left\|\left(k_i^u - K_i\right)^3 + \left(k_i^l - K_i\right)^3\right\|/\left(k_i^u - k_i^l\right)\right]^{1/3}$
Deviation amplitude	Average deviation amplitude (D)	$D = 1/N\sum_{i=1}^{N}\left\|\left(k_i^u - k\right)/2 - K_i\right\|$
	Average relative deviation amplitude (RD)	$RD = 1/N\sum_{i=1}^{N}\left\|\left(k_i^u - k_i^l\right)/(2K_i) - 1\right\|$
Expectation	Average deviation of expectation (Dq)	$Dq = 1/N\sum_{i=1}^{N}\left\|Eq_i - K_i\right\|$
	Average relative deviation of expectation (RDq)	$RDq = 1/N\sum_{i=1}^{N}\left\|Eq_i/K_i - 1\right\|$
	Nash–Sutcliffe efficiency index (NSCE)	$NSCE = 1 - \sum_{i=1}^{N}(Eq_i - K)^2/\sum_{i=1}^{N}(K_i - \overline{K})^2$

n_c is a number of observation data surrounded in band distances, N is a number of observation data, k_i^u and k_i^l are upper and lower limit of ith data sample, K_i is a measured TDS of ith data sample, Eq_i is a computed TDS of ith data sample and $\overline{K}$ is an average value of measured TDS.

$$x_j^{max} = \max\left(x_{1j},\ x_{2j},\ \cdots,\ x_{nj}\right) \tag{11}$$

$$x_j^{min} = \min\left(x_{1j},\ x_{2j},\ \cdots,\ x_{nj}\right) \tag{12}$$

The weight of each uncertainty-accuracy index used in the CUI is typically determined based on expert opinions. In this study, the AHP method was used, which is one of the most widely adopted weighting techniques across various fields. AHP is a decision-making method designed for scenarios involving multiple criteria. It relies on pairwise comparisons, which simplify the judgment process, and it also calculates the degree of inconsistency within the pairwise comparison matrices. For the pairwise comparisons in this study, an absolute scale of 1–9 was used. This scale consists of five main levels for prioritizing two criteria: 1 (equally importance), 3 (moderately importance of one over another), 5 (essential or strong importance), 7 (very strong importance), and 9 (extremely importance). There are also four intermediate values—2, 4, 6, and 8—that can be used when a compromise is needed [61].

Near-infrared surface reflectance played a crucial role in estimating TDS in the M5P model, serving as a logical criterion and being utilized in both proposed equations. The MARS model (Eq. (16)) utilized all $R(NIR)$, $R(red)$, and *ratio* to develop the TDS estimation model. Near-infrared surface reflectance was used twice in the MARS model, and its coefficients are greater than the others, highlighting the importance of this parameter in TDS estimation.

$$\begin{aligned} TDS = {} & 7.57 - 0.725 * max(0,\ R(NIR) + 2.71) + 0.86 \\ & * max(0,\ R(NIR) + 2.36) - 0.27 * max(0,\ Ratio - 0.72) \\ & - 1.35 * max(0,\ R(red) + 1.86) \end{aligned} \tag{16}$$

Genetic programming used only $R(NIR)$ and *ratio* to estimate TDS. Similar to the M5P and MARS algorithms, $R(NIR)$ played a key role in the final symbolic regression model, as shown in Eq. (17).

$$TDS = \left(log\left(\frac{exp((c_0 R(NIR) + c_1) c_2 R(NIR))}{\frac{(c_3 + c_4)}{log(c_5 ratio)}} + \frac{(c_6 + c_7 R(NIR))}{\left(\frac{c_8 R(NIR)}{c_9} + c_{10} ratio\right)} \frac{(c_{11} + c_{12}) c_{13} R(NIR)}{(c_{14} R(NIR) + c_{15})} \right) c_{16} + c_{17} \right) \tag{17}$$

2.5.1. Composite uncertainty index (CUI)

The composite uncertainty index (CUI) was estimated by combining the values of the uncertainty-accuracy indices and their corresponding weights from the AHP method [27].

$$CUI = w_1 \times I_1 + w_2 \times I_2 + \cdots + w_m \times I_m \ i = 1, 2, \cdots, m \tag{13}$$

where I_i is the value of the ith uncertainty-accuracy index, w_i is the corresponding weight of that index, and m represents the number of uncertainty-accuracy indices. The CUI used normalized values and varied from 0 to 1, with larger values indicating better model performance.

3. Results

To develop data-driven models for TDS estimation based on remote sensing data, the dataset was divided into two parts: 80 % for the training set and the remaining 20 % for the testing set. The main concept was to use training data to develop data-driven models, and then use the test data to validate the developed models. In this study, all input and output data were transformed using natural logarithms, enabling linear models in the logarithmic scale to represent nonlinear relationships in the original data.

3.1. Data-driven output models

The nonlinear regression model was developed using the training data, as shown in Eq. (14), where $R(red)$ represents the red band surface reflectance, and *ratio* is defind as $R(red)/R(NIR)$, with $R(NIR)$ denoting the near-infrared surface reflectance.

$$TDS = 7.27 - 0.03R(red) + 0.13ratio \tag{14}$$

M5P model divided the dataset into two subsets and proposed nonlinear regression for each subset as Eq. (15). In this model, $R(NIR) = 0.08$ was estimated by the tree model and used as a logical criterion.

$$TDS = 7.56 - 0.02\mathrm{R(NIR)} - 0.16\mathrm{ratio}\ for\ R(NIR) \le 0.08 \tag{15}$$

$$TDS = 6.95 - 0.17\mathrm{R(red)} - 0.03\mathrm{R(NIR)}\ for\ R(NIR) > 0.08$$

3.2. A comparison of data-driven algorithm

Fig. 7 compares the predicted and measured TDS values using the NLR, M5P, MARS, and SR models. The results indicate that the NLR approach has the highest residuals among all methods. The findings demonstrate that the MARS and SR algorithms provide the best fit with the smallest residuals. Compared to the MARS and SR approaches, the M5P model exhibits narrower deviation ranges, similar to the NLR approach, based on the two simple equations proposed for the final model tree. Therefore, determining the best model for TDS estimation requires further statistical and uncertainty analysis.

The results of statistical analysis such as the coefficients of determination (R^2), root mean squared error (RMSE) and Nash-Sutcliffe Efficiency (NSE) of the TDS estimation using machine learning models, are presented in Table 3. According to Table 3, NLR exhibits the highest RMSE and the lowest R^2 and NSE values across all datasets, indicating poor predictive performance. M5 shows significant improvement over NLR, with lower RMSE and higher R^2 and NSE values. The performance is consistent across datasets, though slightly better on the training set, suggesting minimal overfitting. MARS outperforms both NLR and M5, achieving the lowest RMSE and highest R^2 and NSE values, particularly on the training set. The slight decrease in performance on the test set indicates a minor overfitting tendency. SR demonstrates competitive performance, similar to MARS, on the training set. However, the increased RMSE and decreased R^2 and NSE on the test set suggest overfitting, as the model does not generalize as well to unseen data.

3.2.1. Taylor diagrams

How closely a pattern (or group of patterns) matches observations can be graphically represented using Taylor diagrams [62]. These diagrams measure the correlation, amplitude of variations (indicated by standard deviations), and the centered root-mean-square difference between two patterns to assess their similarity. They are particularly useful for evaluating multiple aspects of different models [63]. Fig. 8 shows the Taylor diagrams for the training, test, and total datasets. For the training and total datasets, the MARS and SR models align well with the measured values. Fig. 8(b) shows that the M5P model outperforms the others in the test dataset, with a standard deviation of 114 mg/L and

a correlation of approximately 0.84. The correlations for the MARS and SR models are about 0.85 and 0.81, respectively.

3.3. Uncertainty analysis

Assuming a 10 % uncertainty in the input parameters for TDS estimation using the IA approach, the 90 % confidence interval diagrams for various estimator methods are presented in Fig. 9. The NLR model exhibits the narrowest bandwidth, with most measured data points falling outside its confidence intervals. The M5 model captures the general trend of the observed data but shows occasional deviations, particularly in regions with higher TDS variability. This model tree demonstrates a significantly better containment ratio compared to the NLR model. The MARS model has slightly broader confidence intervals than the NLR and M5 models, indicating higher uncertainty in predictions. However, it shows strong predictive capability, closely aligning with observed data and effectively capturing nonlinear trends. The 90 % confidence interval suggests that the model's predictions are reliable, with most measured values falling within this range. The SR model displays the widest confidence intervals, indicating greater uncertainty, yet its predictions still capture the general pattern of the observed data.

Table 4 shows the normalized uncertainty indices for each estimator method. According to Table 4, the containing ratio (CR) is directly related to the average bandwidth (B) and the relative average bandwidth (RB). For instance, the NLR model tree has the lowest CR, B, and RB among all methods, while the SR method has the highest CR, B, and RB. The MARS and M5P models also have the highest Nash-Sutcliffe coefficient values, while the NLR method has the lowest RDq. Due to the variety of uncertainty-accuracy indices and the different impacts of each parameter on the efficiency of estimator models, choosing the best method is a challenging problem. Therefore, the AHP weighting method, along with ICU approach, was used to compare the efficiency of different models in this study. In this study to estimate AHP weights, it was assumed that expectation was more important than bandwidth; bandwidth was more important than coverage; coverage was more important than deviation amplitude, and deviation amplitude was more important than symmetry. In addition, relative indices were considered more important than common ones [27]. Based on the normalized matrix of uncertainty-accuracy indices and weights obtained from the AHP method, the CUI approach shows that MARS and M5P models with CUI

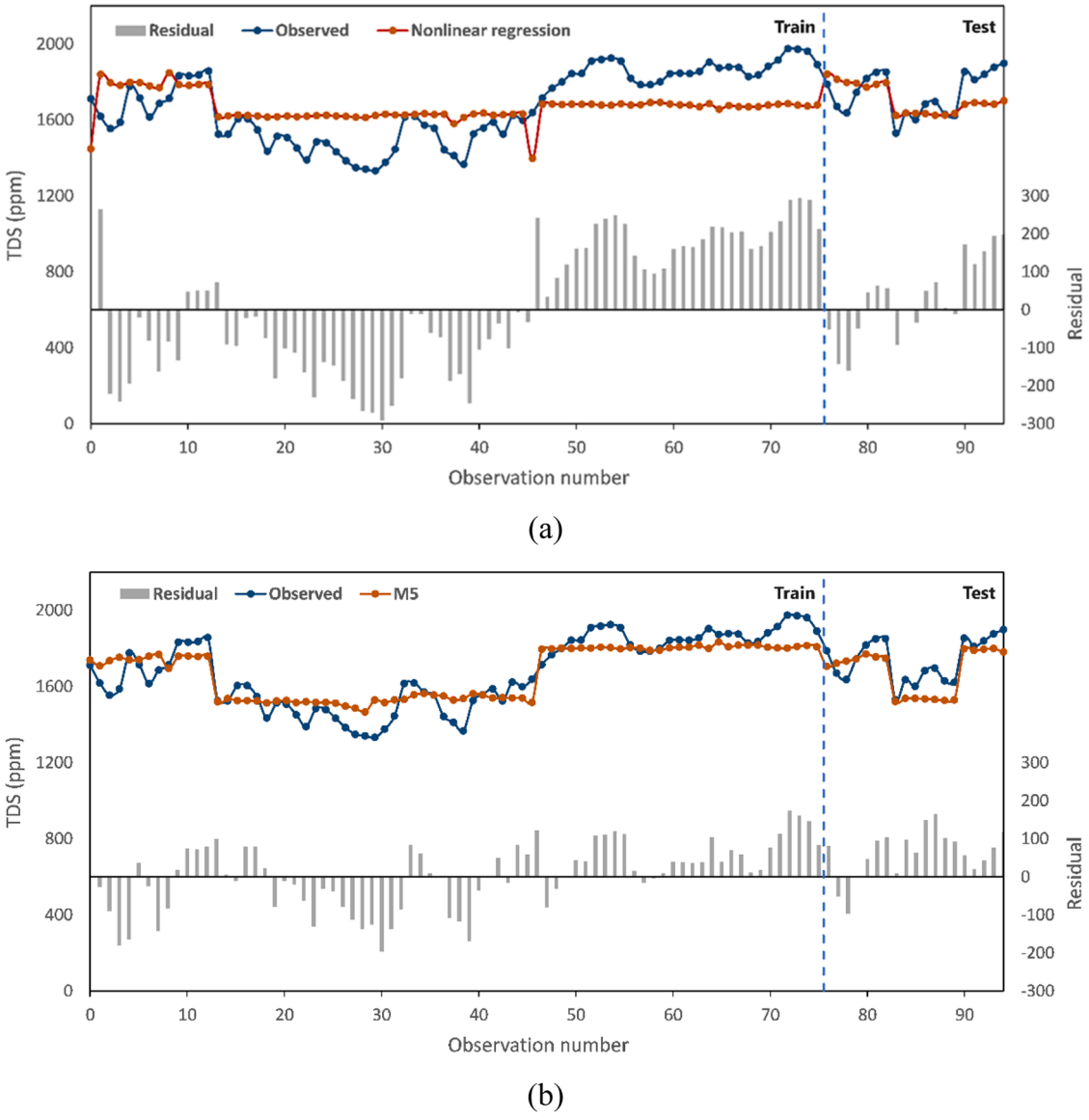
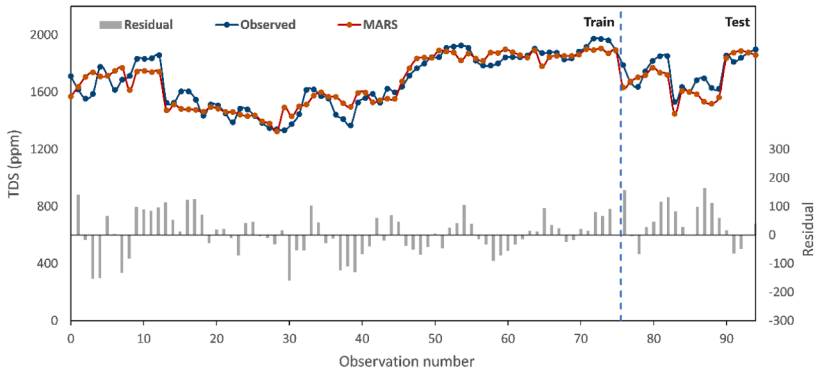
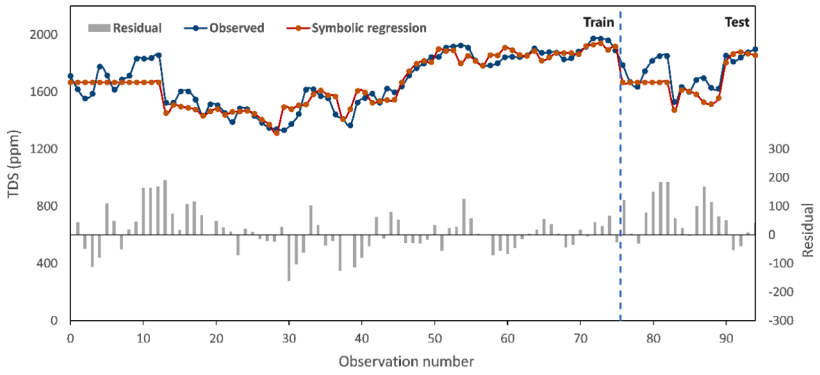


Fig. 7. Representation of the observed vs. predicted TDS values with residuals for both the training and testing dataset employing: **(a)** Nonlinear regression model; **(b)** M5P model tree; **(c)** MARS model; and **(d)** Symbolic regression.



(c)



(d)

Fig. 7. (*continued*).

Table 3
Statistical indices of machine learning models for estimating TDS.

Model	Train			Test			All data		
	NSE	R^2	RMSE	NSE	R^2	RMSE	NSE	R^2	RMSE
NLR	0.16	0.15	171.85	0.11	0.17	108.27	0.17	0.17	161.03
M5	0.82	0.78	87.54	0.55	0.71	88.09	0.78	0.78	87.65
MARS	0.86	0.86	71.00	0.62	0.73	82.93	0.83	0.83	73.57
SR	0.86	0.86	70.42	0.40	0.65	97.42	0.82	0.82	76.64

of 0.83 and 0.72, respectively, perform better than others. Due to the diverse uncertainty-accuracy indices and the varying impacts of each parameter on the performance of estimator models, selecting the best method is challenging. Therefore, the AHP weighting method was used in conjunction with the ICU approach to compare the efficiency of different models in this study. In prioritizing AHP weights, expectation holds the highest importance, followed by bandwidth, coverage, deviation amplitude, and lastly, symmetry. Additionally, relative indices were considered more significant than common ones [27]. The CUI approach, utilizing the normalized uncertainty-accuracy index matrix and AHP weights, identifies the MARS and M5P models as superior performers. With CUI values of 0.83 and 0.72, respectively, these models outperform others in the evaluation.

3.4. Spatiotemporal distribution of TDS

To analyze the spatiotemporal distribution of TDS in the study area, interpolation was carried out using the kriging method. Fig. 10 shows relative error maps of TDS along the Karun River, generated with the NLR, M5P, MARS, and SR algorithms. Based on its accuracy and uncertainty parameters, the NLR model is less effective at simulating TDS levels from remote sensing data, as its maps show the highest relative error in the study area, with large regions exhibiting error values approaching 20 %. This suggests that the NLR model has difficulties to capture the complex dynamics - TDS in the Karun River. In contrast, the M5P, MARS, and SR models generally exhibit lower relative errors for all three dates compared with NLR method. The error distributions of the MARS and M5P models are highly similar on all three sampling dates, although the MARS model consistently demonstrates lower error rates than the M5P model. The SR model's relative error distribution, especially on the second and third dates, is very similar to the MARS model. However, on the first date, particularly in the downstream section of the river, the SR model exhibits a significantly higher relative error.

4. Discussion

The primary strategic policy for most nations globally should focus on evaluating potentially accessible surface water resources using advanced technologies to address the increasing demands driven by rapid population growth and urbanization [64]. Identifying and understanding the primary drivers of water quality is essential for effective remediation of contaminated water resources [65]. The findings of this study have significant implications for regional water management, particularly in areas where monitoring water quality through

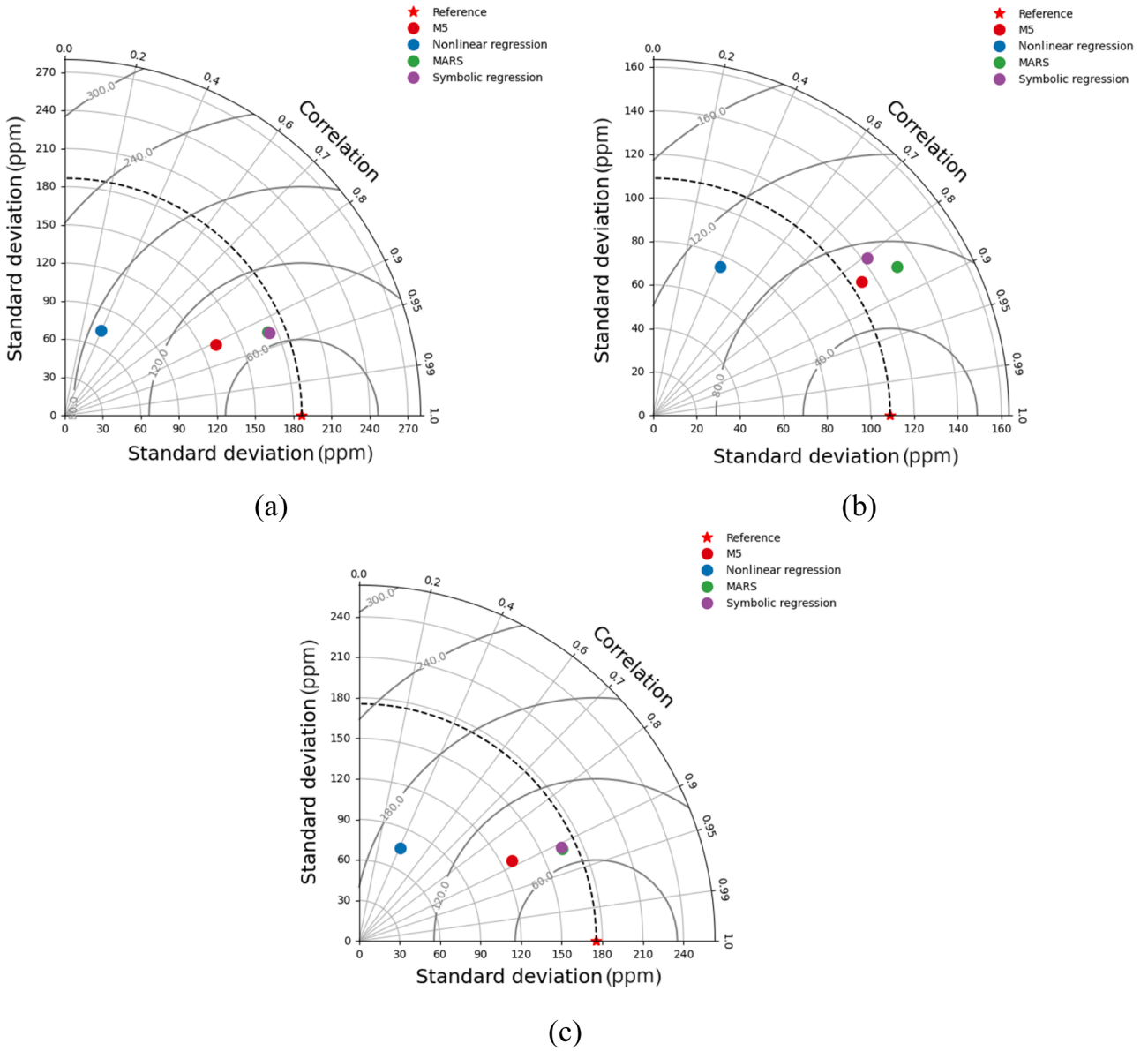


Fig. 8. Taylor diagram for all TDS estimation methods—NLR, M5P, MARS, and SR—for **(a)** the training dataset, **(b)** the validation dataset, and **(c)** the entire dataset.

conventional methods is logistically challenging. The capability of remote sensing, along with machine learning methods to monitor TDS across large spatial extents, enables water managers to identify critical areas at risk of water quality degradation. Since numerous agricultural fields surround the studied river, and their drainage flows directly into it, the proposed model can identify pollution hotspots. This allows for targeted mitigation strategies to reduce the inflow of agricultural runoff and other contaminants. Such an approach supports integrated watershed management by ensuring that resources are allocated efficiently to areas of greatest need. From a policy standpoint, the modeling framework introduced in this study offers a powerful tool for supporting evidence-based decision-making. Policymakers can use TDS maps to define region-specific thresholds for permissible TDS levels in rivers, aligning them with local ecological and socioeconomic conditions. This is especially important in areas where current guidelines may fall short in addressing localized water quality challenges [66]. Additionally, the high temporal resolution of remote sensing data enables real-time monitoring, ensuring that industries and agricultural practices comply with established water quality regulations.

The integration of TDS modeling into water management practices is also critical for building climate resilience. Changes in land use, urbanization, and increasing incidences of extreme weather events are expected to exacerbate water quality challenges [67]. By leveraging remote sensing technologies, managers can proactively monitor and predict the impacts of these changes on TDS levels, enabling timely interventions. Water quality simulations frequently illustrate the in-stream impacts of changes in contaminant sources under various future pollution control strategies. By connecting water quality conditions to pollution reduction efforts, these simulations play a vital role in evaluating the costs and benefits of implementing pollution control measures [68].

Monitoring and data sharing on water quality is one of the key actions for more effective adaptation and more resilience to extremes in the region. Engaging stakeholders in water management processes is essential for the success of any water quality effort. The accessible and scalable nature of remote sensing-based outputs makes them ideal for stakeholder communication. Local governments, industries, and communities can use these visualizations to understand the extent of water quality issues and collaborate on solutions. By fostering participatory decision-making, the insights gained from water quality modeling can

strengthen public trust and cooperation, leading to more sustainable water management outcomes [69].

Comparing the results with other studies, including those of Noori et al. [70] and earlier works, highlights the strong predictive performance of the MARS model. Comparatively, the work of Noori et al. [70] focused on advanced machine learning algorithms, such as Bayesian regularization artificial neural network (BRANN) and Exponential Gaussian process regression (GPR). Their BRANN model outperformed other approaches, achieving an RMSE of 75.03 mg/L and R^2 of 0.89 during the training phase, while the GPR model recorded slightly lower performance with an RMSE of 85.02 mg/L. These results indicate that ANN-based models, particularly BRANN, are effective in capturing complex patterns in TDS modeling. However, SR and MARS models compare favorably, particularly in terms of RMSE during training, where MARS achieved a lower RMSE (73.57 mg/L) than GPR. While the BRANN model suggests slightly better overall predictive power. In the broader context, other studies on TDS modeling show varying levels of performance depending on the algorithm and dataset used. For instance, Peterson et al. [71] used an extreme learning machine regression (ELR) model with RMSE of 41.53 mg/L, while Bourouhou and Salmoun [72] applied a cubic regression equation that achieved RMSE of 114 mg/L.

Noori et al. [70] achieved the best performance among these studies with their BRANN model, supported by 63 in-situ data points. Despite this, the MARS model's performance surpasses most other models in the literature, reinforcing its robustness and suitability for TDS modeling.

The findings of this study highlight the potential of hybrid modeling approaches that integrate symbolic and piecewise regression, leveraging Landsat imagery to achieve a balance between accuracy and interpretability as a viable alternative to complex machine learning models. While the BRANN model [70] achieved the highest predictive power, the difference between the MARS model and BRANN is insignificant, especially considering the computational simplicity and transparency of the proposed approach. Additionally, the MARS model's superior RMSE (73.57 mg/L) compared to GPR (85.02 mg/L) further underscores the practical advantages of hybrid techniques. An important point regarding Noori et al. [70] is the use of various models combined with remote sensing to estimate TDS levels in the Chah-Nimeh Reservoir. Considering the differences in the rate of pollutant variations between the reservoir and the river, the difference in RMSE compared to the present study is entirely justifiable. On the other hand, in Peterson et al. [71], the maximum measured TDS value was 384 mg/L, significantly lower than the 1974 mg/L observed in the present study, and the RMSE reported by Peterson et al. [71] is relatively high given the observational values.

5. Conclusions

The innovation of this study lies in integrating advanced regression

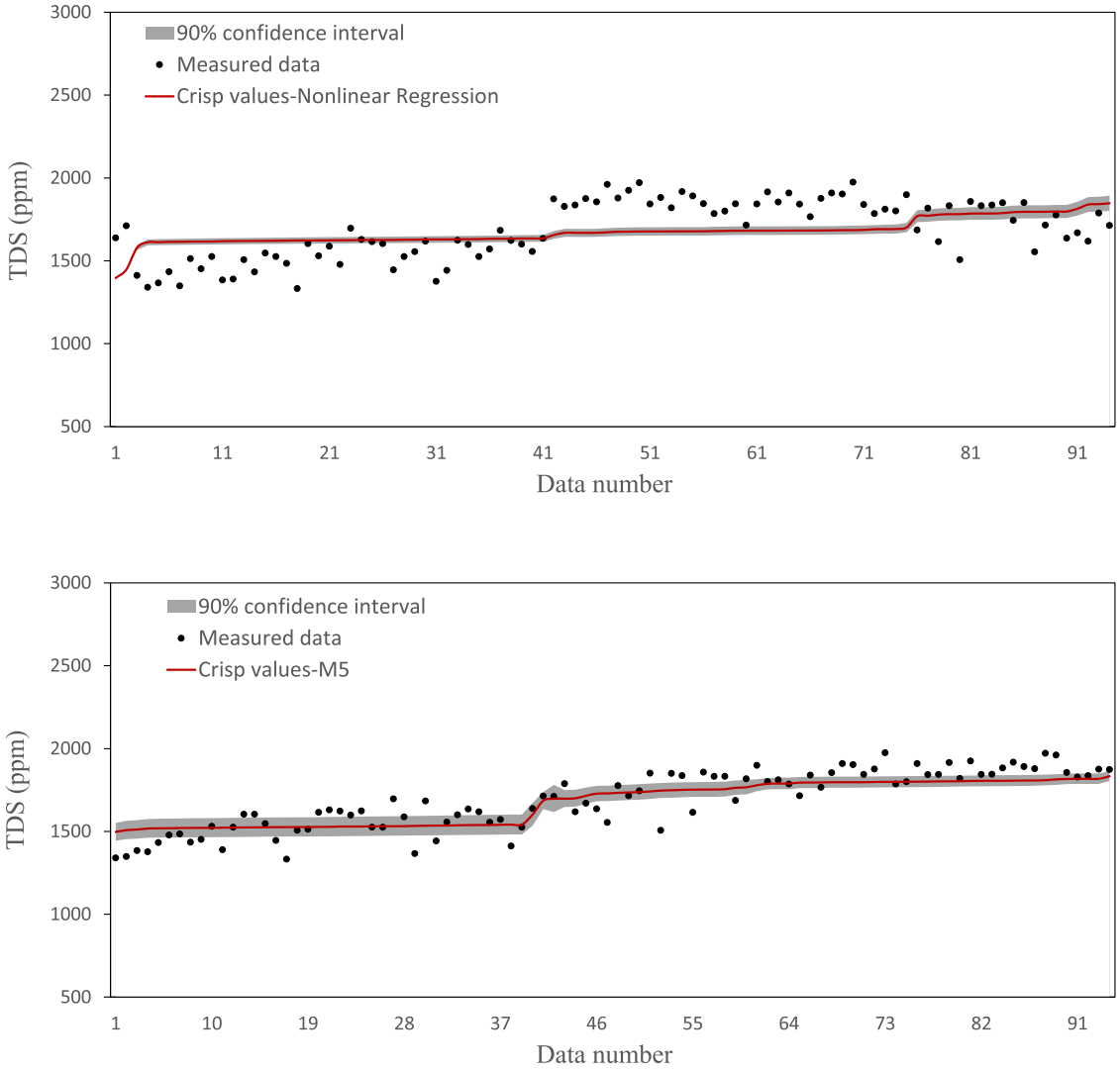


Fig. 9. Crisp values (red line) and 90 % confidence intervals (gray shaded area) for various TDS estimation methods.

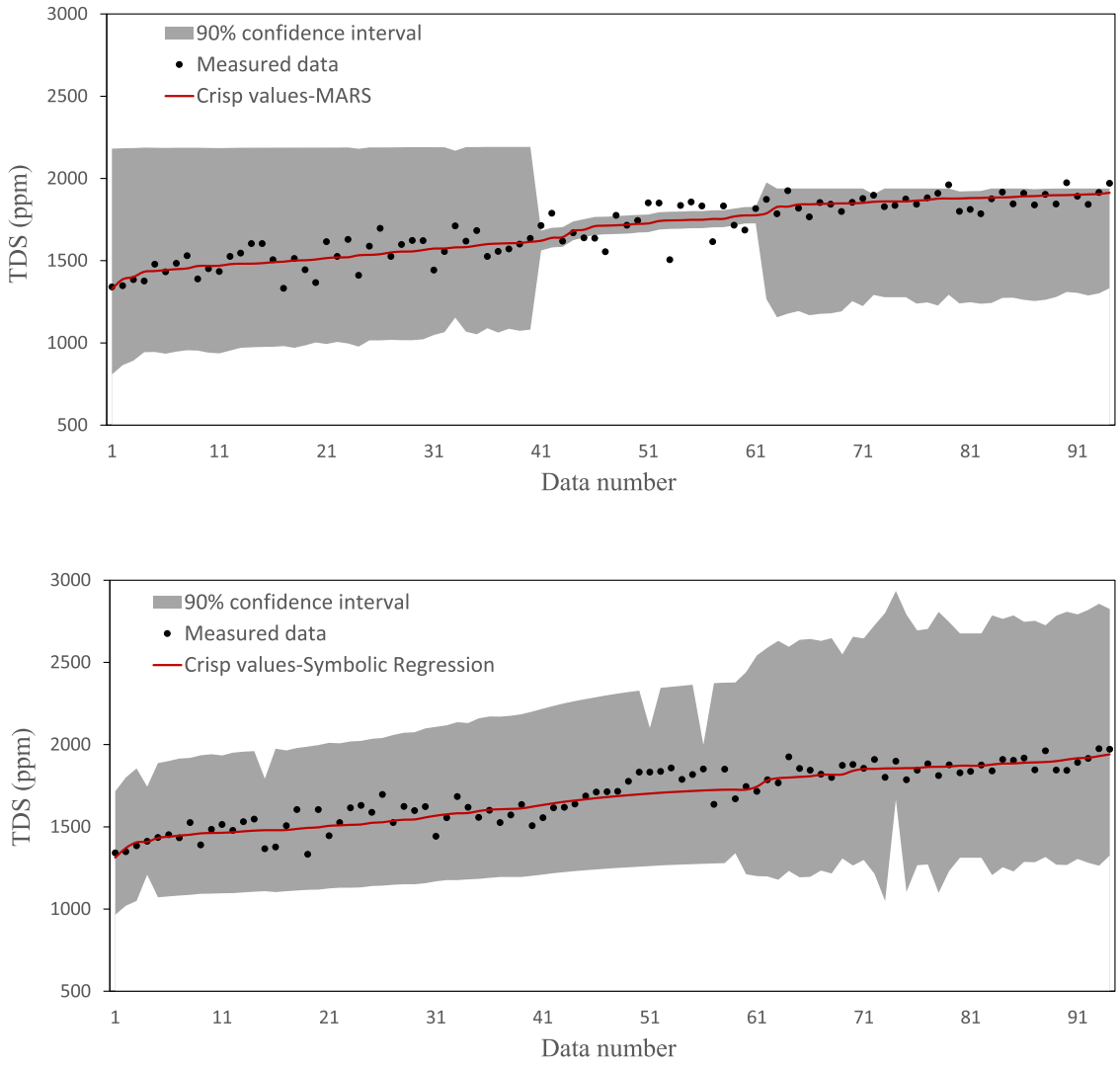


Fig. 9. (*continued*).

Table 4
Uncertainty-accuracy indices for different TDS estimation methods.

TDS estimation method	NSCE	RDq	Dq	RB	B	CR%	RD	D	S	Ts	CUI
NLR	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.19
M5P	0.90	0.46	0.88	0.96	0.96	0.27	0.04	0.04	0.87	0.85	0.72
MARS	1.00	1.00	1.00	0.22	0.30	0.80	0.66	0.76	0.95	0.94	0.83
SR	0.83	0.60	0.98	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.68

models—M5P, MARS, and SR—with remote sensing data and introducing a novel approach to measuring uncertainty in TDS predictions using fuzzy-based interval analysis. The results suggest that nonlinear regression alone cannot accurately simulate TDS due to the complex relationships between the reflectance band and observed TDS values, indicating the need for more sophisticated models. Among the data-driven models used, the MARS and SR models demonstrated better agreement with the measured values. The Nash-Sutcliffe coefficient for the SR method was 0.8, indicating high simulation efficiency. However, interval analysis accounting for $a \pm 10$ % uncertainty in the independent parameters revealed that the SR method had the widest bandwidth. Additionally, the model's efficiency decreased for red band reflectance values below 0.09 and near-infrared reflectance values below 0.03. This study demonstrates that both the MARS and M5P models are well-suited for simulating TDS levels in the Karun River using satellite data. The MARS model shows higher accuracy when input parameters have low uncertainty, whereas the M5P model performs better under high input uncertainty due to its lower sensitivity.

The integration of symbolic and piecewise regression into TDS modeling is particularly significant, where model interpretability is critical for decision-making in water management and policy. As TDS levels are a key indicator of water quality, these findings could have far-reaching implications for monitoring and managing freshwater resources in a sustainable and efficient manner. In conclusion, the hybrid modeling approach-based TDS modeling framework developed in this study offers practical applications for regional water management and policy development. The ability to estimate TDS with high spatial and temporal resolution offers a transformative approach to monitoring water quality.

The reliance on a 4-kilometer stretch of the Karun River for sampling,

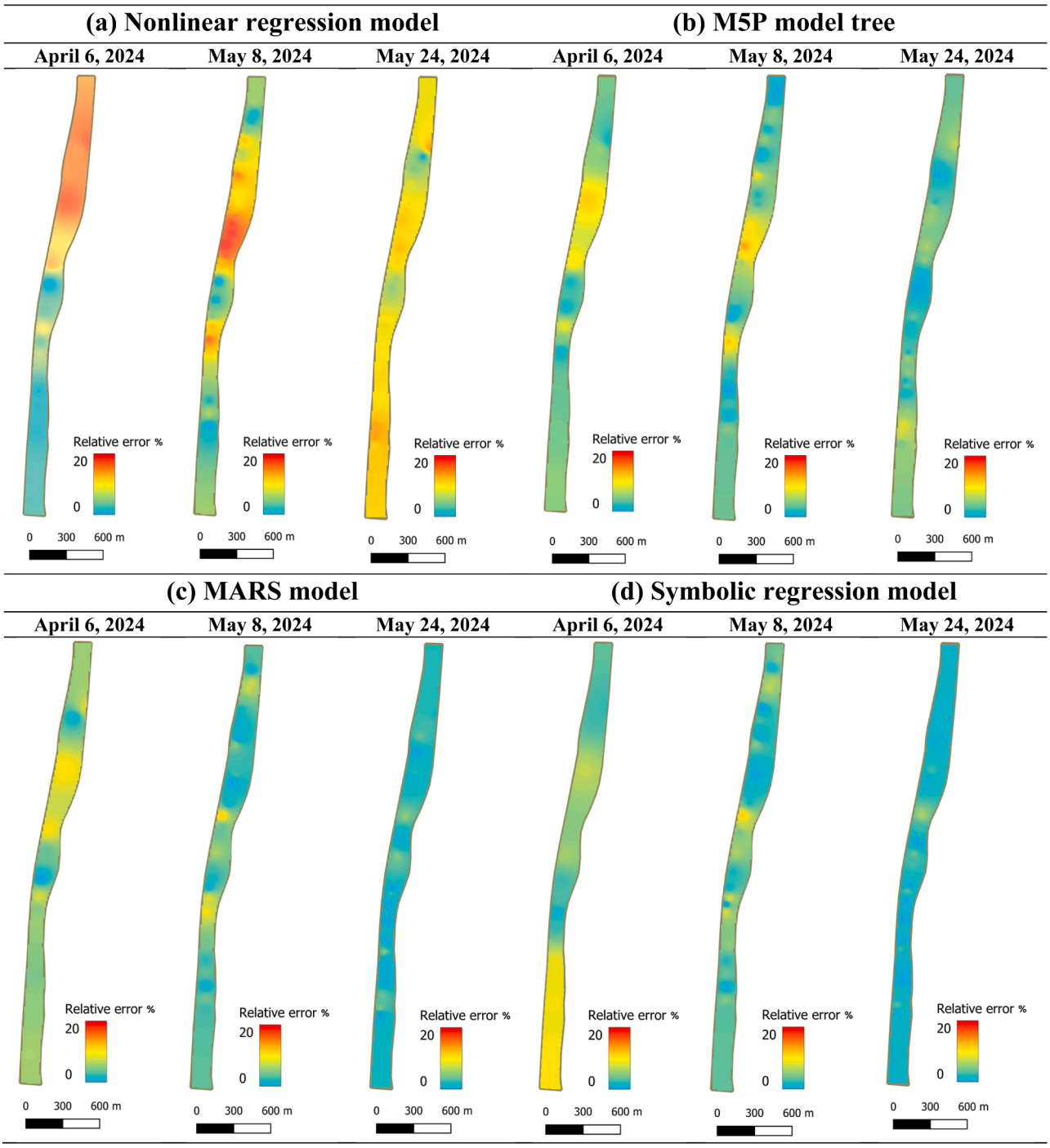


Fig. 10. Relative error maps of TDS spatiotemporal variability along the Karun River, for three different dates (April 6, May 8, and May 24, 2024) using four different machine learning models: (a) Nonlinear Regression, (b) M5P Model Tree, (c) MARS, and (d) Symbolic Regression. The relative error is expressed as a percentage, ranging from 0 % (perfect prediction) to 20 %.

while useful for localized analysis, may limit the broader applicability of the findings to other regions with different hydrological or environmental conditions. Additionally, the temporal resolution of the satellite imagery used (Landsat) may not capture rapid fluctuations in TDS levels, potentially leading to discrepancies between observed and modeled data. To address these limitations, future research could expand the study area to include diverse river systems across different climatic regions, enabling a more comprehensive understanding of TDS dynamics under varying environmental and hydrological conditions. Integrating additional environmental variables, such as water temperature, pH, and turbidity, could further refine the models by capturing more complex interactions affecting TDS levels. Additionally, leveraging data from other satellite sensors, such as Sentinel-2 or hyperspectral imagery, could improve spatial and spectral resolution, enhancing the precision of TDS estimations. These advancements would not only strengthen the applicability of the models but also provide valuable insights for water resource management in regions facing increasing pollution and water scarcity challenges.

Funding

This work was supported by Agricultural Sciences and Natural

Resources University of Khuzestan, Iran.

Ethical approval

Not required as no animal/human was implicated in the study.

Consent to participate

Authors agreeing to participate any survey or feedback tasks.

Consent to publish

Authors providing consent publication of the manuscript to the journal publisher.

CRediT authorship contribution statement

Javad Zahiri: Writing – original draft, Software, Methodology, Formal analysis. **Mohammad Reza Nikoo:** Writing – review & editing, Methodology. **Adell Moradi-Sabzkouhi:** Writing – review & editing, Software. **Mitra Cheraghi:** Conceptualization. **Nazmi Mat Nawi:** Writing – review & editing.

Declaration of competing interest

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rineng.2025.104159.

Data availability

Supplementary data are available with this manuscript.

References

[1] X. Zhou, Y. Leng, M. Salarijazi, I. Ahmadianfar, A.A. Farooque, Development of forecasting of monthly SAR time series in river systems: a multivariate data decomposition-based hybrid approach, Process Saf. Environ. Prot. 188 (2024) 1355–1375.

[2] B. Stride, S. Abolfathi, M.G.N. Odara, G.D. Bending, J. Pearson, Modeling microplastic and solute transport in vegetated flows, Water Resour. Res. 59 (5) (2023) e2023WR034653.

[3] M. Guo, R. Noori, S. Abolfathi, Microplastics in freshwater systems: dynamic behaviour and transport processes, Resour. Conserv. Recycl. 205 (2024) 107578.

[4] H. Tian, L. Wang, X. Zhu, M. Zhang, L. Li, Z. Liu, S. Abolfathi, Biodegradation of microplastics derived from controlled release fertilizer coating: selective microbial colonization and metabolism in plastisphere, Sci. Total Environ. 920 (2024) 170978.

[5] A. Mohammadpour, E. Gharehchahi, M.A. Gharaghani, E. Shahsavani, M. Golaki, R. Berndtsson, S. Abolfathi, Assessment of drinking water quality and identifying pollution sources in a chromite mining region, J. Hazard. Mater. 480 (2024) 136050.

[6] R. Noori, B. Ghiasi, S. Salehi, M. Esmaeili Bidhendi, A. Raeisi, S. Partani, S. Abolfathi, An efficient data driven-based model for prediction of the total sediment load in rivers, Hydrology 9 (2) (2022) 36.

[7] J. Wu, J. Lu, Spatial scale effects of landscape metrics on stream water quality and their seasonal changes, Water Res. 191 (2021) 116811.

[8] R. Noori, M. Maghrebi, S. Jessen, S.M. Bateni, E. Heggy, S. Javadi, A. AghaKouchak, Decline in Iran's groundwater recharge, Nat. Commun. 14 (1) (2023) 6674.

[9] R. Noori, M. Noury, M.K. Poshtegal, M. Sadrinasab, M. Mahdian, R. Bhattarai, S. Abolfathi, Thermal stratification and mixing of dam reservoirs in Iran, Watershed Ecology and the Environment 6 (2024) 138–145.

[10] G.E. Adjovu, H. Stephen, D. James, 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) 15 (14) (2023) 3534.

[11] B. Fakouri, J. Mohammad Vali Samani, H. Mohammad Vali Samani, M. Mazaheri, Optimal waste load model in Karoon River with the pollution loading loss analysis, Iran-Water Resour. Res. 17 (3) (2021) 330–344.

[12] D.P. Sahoo, B. Sahoo, M.K. Tiwari, MODIS-Landsat fusion-based single-band algorithms for TSS and turbidity estimation in an urban-waste-dominated river reach, Water Res. 224 (2022) 119082.

[13] M. Kadkhodazadeh, S. Farzin, Introducing a novel hybrid machine learning model and developing its performance in estimating water quality parameters, Water Resour. Manage. 36 (10) (2022) 3901–3927.

[14] F.A. Pourhosseini, K. Ebrahimi, M.H. Omid, Prediction of total dissolved solids, based on optimization of new hybrid SVM models, Eng. Appl. Artif. Intell. 126 (2023) 106780.

[15] M. Jamei, I. Ahmadianfar, X. Chu, Z.M. Yaseen, Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: new approach, J. Hydrol. 589 (2020) 125335.

[16] P.M. Mather, Computer Processing of Remotely-Sensed images: An introduction, 3rd ed., Wiley, New York, 2009.

[17] H.R. Gordon, M. Wang, Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm, Appl. Opt. 33 (1994) 443–452.

[18] Kazemzadeh M., Ayoubzadeh S., Moridnejad A. (2013) Estimation of suspended sediment concentration in surface water with high concentrations using remote sensing techniques. 6th National Congress and Exhibition of Environmental Engineering, Tehran, Iran (In Persian).

[19] N. Bernardo, A. do Carmo, E. Park, E. Alcˊantara, Retrieval of suspended particulate matter in inland waters with widely differing optical properties using a semi-analytical scheme, Remote Sens. (Basel) 11 (2019) 2283.

[20] Y. Du, K. Song, Q. Wang, S. Li, Z. Wen, G. Liu, H. Tao, Y. Shang, J. Hou, L. Lyu, B. Zhang, Total suspended solids characterization and management implications for lakes in East China, Sci. Total Environ. 806 (2022) 151374.

[21] S. Ghosh, S. Saha, B. Bera, Dynamics of total suspended solid concentrations in the lower Raidak river (Himalayan foreland Basin), India, Adv. Space Res. (2022).

[22] J. Li, L. Tian, Y. Wang, S. Jin, T. Li, X. Hou, Optimal sampling strategy of water quality monitoring at high dynamic lakes: a remote sensing and spatial simulated annealing integrated approach, Sci. Total Environ. 777 (2021) 146113.

[23] H. Xu, G. Xu, X. Wen, X. Hu, Y. Wang, Lockdown effects on total suspended solids concentrations in the Lower Min River (China) during COVID-19 using time-series remote sensing images, Int. J. Appl. Earth Obs. Geoinf. 98 (2021) 102301.

[24] P.M. Davis, T.C. Atkinson, T.M.L. Wigley, Longitudinal dispersion in natural channels: 2. The roles of shear flow dispersion and dead zones in the River Severn, U.K. Hydrol. Earth Syst. Sci. Discussions 4 (3) (2000) 355–371.

[25] Mays L. 1992. Water demand forecasting. Hydrosystem Engineering and Management, 24–32.

[26] Y.K. Tung, Hydrosystems Engineering Uncertainty Analysis, McGraw-Hill, 2005.

[27] P. Shi, T. Yang, B. Yong, Z. Li, C.Y. Xu, Q. Shao, Y. Qin, A new uncertainty measure for assessing the uncertainty existing in hydrological simulation, Water, 11 (4) (2019) 812.

[28] J. Zahiri, Z. Mollaee, M.R. Ansari, Estimation of suspended sediment concentration by M5 model tree based on hydrological and moderate resolution imaging spectroradiometer (MODIS) data, Water Resour. Manage. 34 (12) (2020) 3725–3737.

[29] G. Chander, B.L. Markham, D.L. Helder, Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors, Remote Sens Environ 113 (5) (2009) 893–903.

[30] C. Teixeira Pinto, X. Jing, L Leigh, Evaluation analysis of Landsat level-1 and level-2 data products using in situ measurements, Remote Sens (Basel) 12 (16) (2020) 2597.

[31] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google Earth Engine: planetary-scale geospatial analysis for everyone, Remote Sens. Environ. 202 (2017) 18–27.

[32] Dean, J., & Ghemawat, S. (2004). MapReduce: simplified data processing on large clusters. In Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI'04), San Francisco, CA, USA, 6–8 December 2004; pp. 137–150.

[33] G. Mateo-García, L. Gómez-Chova, J. Amorós-López, J. Muñoz-Marí, G. Camps-Valls, Multitemporal cloud masking in the Google Earth Engine, Remote Sens (Basel) 10 (7) (2018) 1079.

[34] Sayler, K. (2024). Landsat 8-9 collection 2 (C2) level 2 science Product (L2SP) Guide, Department of the Interior, U.S. Geological Survey, LSDS-1619, version 6.

[35] Rice, E.W., Bridgewater, L., & American Public Health Association (Eds.). (2012). *Standard methods for the examination of water and wastewater* (Vol. 10). Washington, DC: american public health association.

[36] Quinlan, J.R. (1992). Learning with continuous classes. Proceedings of Australian Joint Conference on Artificial Intelligence.

[37] J. Zahiri, H. Nezaratian, Estimation of transverse mixing coefficient in streams using M5, MARS, GA, and PSO approaches, Environ. Sci. Pollution Res. 27 (13) (2020) 14553–14566.

[38] A. Yeganeh-Bakhtiary, H. EyvazOghli, N. Shabakhty, S. Abolfathi, Machine learning prediction of wave characteristics: comparison between semi-empirical approaches and DT model, Ocean Eng. 286 (2023) 115583.

[39] K. Khosravi, N. Attar, S.M. Bateni, C. Jun, D. Kim, M.J.S. Safari, S. Abolfathi, Daily river flow simulation using ensemble disjoint aggregating M5-Prime model, Heliyon 10 (20) (2024).

[40] H. Song, I. Gi, J. Ryu, Y. Kwon, J. Jeong, Production planning forecasting system based on M5P algorithms and master data in manufacturing processes, Applied Sciences 13 (13) (2023) 7829.

[41] A. Etemad-Shahidi, M. Taghipour, Predicting longitudinal dispersion coefficient in natural streams using M5 model tree, J. Hydraul. Eng. 138 (6) (2012) 542–554.
[42] N.C. Jung, I. Popescu, P. Kelderman, D.P. Solomatine, R.K. Price, Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea, J. Hydroinf. 12 (3) (2010) 262–274.
[43] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1) (1991) 1–67.
[44] W. Zhang, A.T. Goh, Y. Zhang, Multivariate adaptive regression splines application for multivariate geotechnical problems with big data, Geotech. Geol. Eng. 34 (2016) 193–204.
[45] M.B. Adiguzel, M.A. Cengiz, Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria, Heliyon 9 (9) (2023).
[46] N. Murat, Outlier detection in statistical modeling via multivariate adaptive regression splines, Communic. Statist.-Simulat. Comput. 52 (7) (2023) 3379–3390.
[47] K. Takaki, T. Miyao, Symbolic regression for the interpretation of quantitative structure-property relationships, Artificial Intellig. Life Sci. 2 (2022) 100046.
[48] J.R. Koza, Genetic programming as a means for programming computers by natural selection, Statist. Comput. 4 (1994) 87–112.
[49] Makke, Nour, and Sanjay Chawla. "Symbolic regression: a pathway to interpretability towards automated scientific discovery." In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6588-6596. 2024.
[50] J. Kubalík, E. Derner, J. Žegklitz, R. Babuška, Symbolic regression methods for reinforcement learning, IEEE Access 9 (2021) 139697–139711.
[51] F.O. de França, A greedy search tree heuristic for symbolic regression, Inf Sci (Ny) 442 (2018) 18–32.
[52] D. Angelis, F. Sofos, T.E. Karakasidis, Artificial intelligence in physical sciences: symbolic regression trends and perspectives, Arch. Comput. Meth. Eng. 30 (6) (2023) 3845–3865.
[53] S.V. Balasubramanian, N. Pahlevan, B. Smith, C. Binding, J. Schalles, H. Loisel, E. Boss, Robust algorithm for estimating total suspended solids (TSS) in inland and nearshore coastal waters, Remote Sens. Environ. 246 (2020) 111768.
[54] F. Nasiri, I. Maqsood, G. Huang, N. Fuller, Water quality index: a fuzzy river-pollution decision support expert system, J. Water Resour. Plann. Manage. 133 (2) (2007) 95–105.
[55] L.S. Lasdon, A.D. Waren, A. Jain, M. Ratner, Design and testing of a generalized reduced gradient code for nonlinear programming, ACM Transactions on Mathematical Software (TOMS) 4 (1) (1978) 34–50.
[56] X.J. Xie, Research on material selection with multi-attribute decision method and G1 method, Adv. Mat. Res. 952 (2014) 20–24.
[57] L.J. Herrera, H. Pomares, I. Rojas, O. Valenzuela, A. Prieto, TaSe, a Taylor series-based fuzzy system model that combines interpretability and accuracy, Fuzzy Sets Syst. 153 (3) (2005) 403–427.
[58] V. Ibáñez-Forés, M.D. Bovea, A. Azapagic, Assessing the sustainability of best available techniques (BAT): methodology and application in the ceramic tiles industry, J. Clean. Prod. 51 (2013) 162–176.
[59] J. Ma, Z.P. Fan, L.H. Huang, A subjective and objective integrated approach to determine attribute weights, Eur. J. Oper. Res. 112 (2) (1999) 397–404.
[60] Z.H. Zou, Y. Yi, J.N. Sun, Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment, J. Environ. Sci. 18 (5) (2006) 1020–1023.
[61] T.L. Saaty, How to make a decision: the analytic hierarchy process, Eur. J. Oper. Res. 48 (1) (1990) 9–26.
[62] K.E. Taylor, Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.: Atmos. 106 (D7) (2001) 7183–7192.
[63] C. Toffanin, F. Di Palma, F. Iacono, L. Magni, LSTM Network for the oxygen concentration modeling of a wastewater treatment plant, Applied Sciences 13 (13) (2023) 7461.
[64] A. Das, Surface water quality evaluation, apportionment of pollution sources and aptness testing for drinking using water quality indices and multivariate modelling in Baitarani River basin, Odisha, Hydro Res. 8 (2025) 244–264.
[65] M. Salarijazi, I. Ahmadianfar, Z.M. Yaseen, Prediction enhancement for surface water sodium adsorption ratio using limited inputs: implementation of hybridized stacked ensemble model with feature selection algorithm, Phys. Chem. Earth, Parts A/B/C 134 (2024) 103561.
[66] Carr, G.M., & Neary, J.P. (2008). Water quality for ecosystem and human health. UNEP/Earthprint.
[67] Parmesan, C., Morecroft, M.D., & Trisurat, Y. (2022). Climate change 2022: impacts, adaptation and vulnerability (Doctoral dissertation, GIEC).
[68] Schwarz, G.E., Hoos, A.B., Alexander, R.B., & Smith, R.A. (2006). The SPARROW surface water-quality model: theory, application and user documentation.
[69] Pltonykova, H., Koeppel, S., Bernardini, F., Tiefenauer-Linardon, S., de Strasser, L., & Connor, R. (2020). The United Nations World Water Development Report 2020: water and Climate Change.
[70] A. Noori, S.H. Mohajeri, M. Delnavaz, R. Rezazadeh, A spatiotemporal monitoring model of TSM and TDS in arid region lakes utilizing Sentinel-2 imagery, J. Arid Environ. 216 (2023) 105024.
[71] K.T. Peterson, V. Sagan, P. Sidike, E.A. Hasenmueller, J.J. Sloan, J.H. Knouft, Machine learning-based ensemble prediction of water-quality variables using feature-level and decision-level fusion with proximal remote sensing, Photogrammetric Eng. Remote Sens. 85 (4) (2019) 269–280.
[72] I. Bourouhou, F. Salmoun, Sea water quality monitoring using remote sensing techniques: a case study in Tangier-Ksar Sghir coastline, Environ. Monit. Assess. 193 (9) (2021) 557.