Science of Remote Sensing 13 (2026) 100355

Contents lists available at ScienceDirect

Science of Remote Sensing

journal homepage: www.sciencedirect.com/journal/science-of-remote-sensing

Beyond spectral signals: Geographic features drive bathymetric accuracy in the turbid Sancha Lake using machine learning

Xiaojuan Li [a,b], Wei Zhang [c], Hongrui Zheng [d], Zhongqiang Wu [e,*], Hongliang Lu [a]

[a] School of Geographical Sciences, China West Normal University, China
[b] Sichuan Provincial Engineering Laboratory of Monitoring and Control for Soil Erosion in Dry Valleys, China West Normal University, China
[c] National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, China
[d] School of Geoscience and Technology, Southwest Petroleum University, China
[e] School of Artificial Intellicence, Hainan Normal University, China

ARTICLE INFO

Keywords:
Satellite-derived bathymetry
XGBoost
Spatial-spectral integration
Inland waters
Machine learning

ABSTRACT

Accurate bathymetric mapping in inland water bodies presents significant challenges for conventional optical remote sensing due to complex water quality conditions and variable bottom types. This study introduces a novel Spectral-Geospatial XGBoost Regression (SG-XGBoost) model that revolutionizes depth estimation by integrating comprehensive spectral transformations with explicit geographic coordinates through gradient boosting methodology. Applied to Sancha Lake, a morphologically complex reservoir in China's Upper Yangtze watershed, the model achieved exceptional performance with $R^2 = 0.91$ and RMSE = 1.66m, representing 70 % improvement over traditional empirical methods (Stumpf, Log-Linear) and 21 % advancement beyond Random Forest. The iterative error correction and sophisticated regularization of the gradient boosting methodology not only enable the effective exploitation of spatial-spectral interactions but also ensure better accuracy is maintained across all depth ranges (2–31m). The feature importance analysis revealed an unexpected finding, the geographic coordinates dominated predictive power (85 % contribution), while spectral features contributed minimally, challenging fundamental assumptions about optical bathymetry. The iterative error correction and sophisticated regularization of the gradient boosting methodology not only enable the effective exploitation of spatial-spectral interactions but also ensure better accuracy is maintained across all depth ranges (2–31m). Bathymetric maps generated by SG-XGBoost successfully captured fine-scale morphological features invisible to conventional approaches, including channels <30m wide and subtle depth variations of 1–2m. Despite limitations in extreme turbidity and site-specificity requiring readjustment for new water bodies, this research establishes gradient boosting with spatial-spectral integration as a transformative approach for inland water bathymetry, with broader implications for aquatic remote sensing applications including water quality monitoring and habitat mapping.

1. Introduction

Accurate bathymetric information is fundamental for a wide range of applications including water resource management, navigation safety, aquatic ecosystem monitoring, flood modeling, and coastal engineering (Garcia et al., 2020; Kerr et al., 2018). Traditional in-situ bathymetric surveys using ship-based echo sounders, while providing high accuracy, are labor-intensive, time-consuming, and costly, particularly when mapping extensive water bodies such as large lakes and reservoirs (Ashphaq et al., 2021; Ma et al., 2020). Satellite-derived bathymetry (SDB) has emerged as a compelling alternative, offering synoptic coverage, cost-effectiveness, and the ability to access remote or hazardous areas where conventional surveying is challenging (Agrafiotis et al., 2024; Al et al., 2023; Alevizos, 2020; Li et al., 2025; AlHossainy et al., 2025; Lv et al., 2025). The increasing availability of high-resolution multispectral satellite imagery has revolutionized shallow water bathymetric mapping, enabling frequent monitoring of dynamic aquatic environments and supporting sustainable water resource management (Barnes et al., 2018; Liu et al., 2021).

The evolution of bathymetric remote sensing has progressed through several paradigmatic shifts since Lyzenga's pioneering work in the late 1970s (Lyzenga, 1978, 1985). Traditional empirical methods, including

* Corresponding author.
E-mail addresses: lixiaojuan_2009@163.com (X. Li), zhangwrs@163.com (W. Zhang), myzhenghr@163.com (H. Zheng), wuzhongqiang2008@qq.com (Z. Wu).

https://doi.org/10.1016/j.srs.2025.100355
Received 20 September 2025; Received in revised form 7 November 2025; Accepted 17 December 2025
Available online 23 December 2025

the single-band linear model and band-ratio algorithms, established the foundation for optical bathymetry by exploiting the exponential attenuation of light in water (Chen et al., 2019). Stumpf et al. (2003) developed a log-transformed band ratio method that effectively addresses variable bottom albedo, becoming one of the most widely applied approaches for satellite-derived bathymetry. The log-linear model proposed by Lyzenga, 1978, 1985 provided a physics-based framework linking water-leaving radiance to depth through Beer's law, offering robust performance in clear to moderately turbid waters.

Semi-analytical and physics-based models have advanced the field by incorporating radiative transfer theory to account for water column optical properties and bottom reflectance (Lee et al., 2012; Gao, 2009). Lee et al. (2012) developed an optimization-based inversion approach that simultaneously retrieves water depth and inherent optical properties, while Hedley et al. (2016) introduced methods to handle environmental noise and sun glint effects. These physically-grounded approaches provide theoretical rigor but often require extensive parameterization and may struggle in optically complex waters where assumptions about water optical properties are violated (Legleiter et al., 2011; Huang et al., 2017).

The advent of machine learning has marked a transformative phase in bathymetric remote sensing, offering unprecedented capability to capture complex, non-linear relationships between spectral signatures and water depth (Agrafiotis et al., 2024, 2025; Li et al., 2023a; Misra et al., 2018; Liu et al., 2025a; Wu et al., 2024). Random Forest algorithms have demonstrated superior performance compared to traditional methods, particularly in turbid waters where conventional approaches fail to deliver results (Sagawa et al., 2019). Sagawa et al. (2019) successfully applied Random Forest to multi-temporal satellite imagery, achieving improved bathymetric accuracy through ensemble learning. Deep learning approaches, including convolutional neural networks and transformer architectures, have further pushed the boundaries of bathymetric mapping accuracy (Lv et al., 2025; Wan et al., 2021). Recent studies by Benshila et al. (2020) and Al Najar et al. (Al et al., 2023) have shown that deep neural networks can extract hierarchical features from satellite imagery, enabling robust depth estimation even in challenging conditions. Advanced architectures such as transformers for hyperspectral fusion (e.g., CasFormer) and specialized hyperspectral imaging techniques demonstrate the potential for enhanced spectral-spatial feature extraction, though these approaches require data availability beyond standard multispectral platforms. While such hyperspectral and deep learning methods represent important research frontiers, our study focuses on gradient boosting with widely accessible Sentinel-2 multispectral data to ensure practical applicability in resource-constrained settings.

Recent advances in 2024–2025 have further expanded the methodological toolkit, with BiGRU (Bidirectional Gated Recurrent Units) models for large-area mapping (Xi et al., 2025), virtual coastal band optimization approaches (Vinayaraj et al., 2016; Liu et al., 2025b), multi-temporal fusion methods combining active and passive remote sensing (Li et al., 2025; Wu et al., 2024). and transformer-based architectures like BathyFormer demonstrating state-of-the-art performance in complex coastal environments (Lv et al., 2025; Wan et al., 2021). Notable progress includes deep learning methods that operate without in-situ depth measurements using structure-from-motion photogrammetry (Agrafiotis et al., 2025; Liu et al., 2025b). And multimodal datasets enabling comprehensive method benchmarking (Agrafiotis et al., 2024). The integration of geographic features with gradient boosting machines has shown particular promise for inland rivers and transitional waters (Li et al., 2025; Wu et al., 2024)., supporting our proposed spatial-spectral fusion framework.

The emergence of gradient boosting algorithms represents the latest advancement in machine learning for bathymetry. XGBoost (eXtreme Gradient Boosting), with its sophisticated regularization framework and efficient handling of feature interactions, has shown exceptional promise for environmental remote sensing applications (AlHossainy et al., 2025; Chen et al., 2016). Unlike Random Forest's parallel tree construction, XGBoost's sequential boosting mechanism allows each tree to learn from the residual errors of its predecessors, resulting in superior predictive performance.Sancha Lake, a significant reservoir in the Upper Yangtze River watershed, exemplifies such environments where complex interactions between natural processes and anthropogenic influences create spatially heterogeneous depth patterns. The lake's ecosystem experiences multiple stressors from agricultural nutrient loading and urban expansion, necessitating sophisticated monitoring approaches that can capture both spectral variations related to water quality and spatial patterns reflecting bathymetric structure (Su et al., 2023).

Despite these advances, existing machine learning approaches for bathymetry have not fully exploited the synergistic potential of spectral-spatial features within gradient boosting frameworks. Most studies either focus solely on spectral band combinations or treat spatial information as supporting data rather than core features (Agrafiotis et al., 2025; Liu et al., 2025b). This represents a significant gap, as the gradient boosting mechanism of XGBoost is particularly well-suited to discover and exploit complex interactions between heterogeneous feature types (Chen et al., 2016). This study introduces a novel Spectral-Geospatial XGBoost Regression (SG-XGBoost) model specifically designed for bathymetric mapping in inland water bodies. Our approach systematically integrates comprehensive spectral transformations including band ratios, spectral indices, and non-linear transforms with explicit geospatial features such as geographic coordinates and their interactions. By utilizing XGBoost's gradient boosting system, which iteratively refines predictions through sequential tree construction, we aim to capture the complex, non-linear relationships between observable surface features and water depth that traditional methods and even standard machine learning approaches may fail to capture.

This study introduces a novel Spectral-Geospatial XGBoost Regression (SG-XGBoost) model specifically designed for bathymetric mapping in inland water bodies. Our approach systematically integrates comprehensive spectral transformations including band ratios, spectral indices, and non-linear transforms with explicit geospatial features such as geographic coordinates and their interactions. By utilizing XGBoost's gradient boosting system, which iteratively refines predictions through sequential tree construction, we aim to capture the complex, non-linear relationships between observable surface features and water depth that traditional methods and even standard machine learning approaches may fail to capture.

The primary objectives of this research are to: (1) develop a robust feature extraction and integration framework that effectively combines spectral and spatial information for bathymetric estimation; (2) evaluate the performance of the SG-XGBoost model against established methods including Stumpf, Log-linear, and Random Forest approaches; (3) assess the relative importance of spectral in comparison to spatial features in determining bathymetric accuracy; and (4) demonstrate the real-world applicability of the proposed method for bathymetric mapping in Sancha Lake.

This paper is organized as follows: Section 2 describes the study area, data acquisition, and preprocessing procedures. Section 3 details the SG-XGBoost methodology, including feature engineering, model architecture, and hyperparameter optimization. Section 4 presents the comparative results and bathymetric mapping outputs. Section 5 discusses the implications of our findings, model uncertainties, and future research directions. Finally, Section 6 concludes with key insights and recommendations for advancing satellite-derived bathymetry in inland waters.

2. Materials and methods

2.1. Study area

Sancha Lake (30°18′N, 104°26′E), located in the Upper Yangtze River

watershed in Sichuan Province, China, represents a critical water resource for the region's socio-economic development and ecological sustainability (Li et al., 2022). This artificial reservoir, constructed in 1977, spans approximately 27 km^2 with an average depth of 8.3 m and maximum depth exceeding 30 m (Li et al., 2023b). The lake's complex morphology, characterized by multiple tree-like tributaries and varying bathymetric gradients, presents unique challenges for remote sensing-based depth estimation (see Fig. 1).

The lake ecosystem experiences significant anthropogenic pressures from agricultural runoff, aquaculture activities, and tourism development, resulting in spatially heterogeneous water quality conditions (Li et al., 2023b). These varying optical properties, combined with seasonal algal blooms and suspended sediment dynamics, create a challenging environment for bathymetric mapping that necessitates advanced machine learning approaches. The subtropical monsoon climate of the region contributes to distinct seasonal variations in water level and turbidity, with peak rainfall occurring from May to September, affecting both water clarity and spectral signatures (Li et al., 2023b).

2.2. Data acquisition and processing

2.2.1. Satellite imagery acquisition

In this study, we utilized Sentinel-2 Level-1C multispectral imagery acquired on December 18, 2017 (S2B_MSIL1C_20171218T034139_N0500_R061_T48RVU_20230915T070233.SAFE),coinciding with the winter low-water period when water clarity typically reaches optimal conditions (Kutser et al., 2020). The Sentinel-2 MultiSpectral Instrument (MSI) provides 13 spectral bands ranging from visible to shortwave infrared wavelengths, with spatial resolutions of 10m, 20m, and 60m (Drusch et al., 2012). For bathymetric applications, we focused on bands 1–4 (coastal aerosol, blue, green, red) at 10m resolution, and bands 5–8 (red edge and near-infrared) resampled to 10m using cubic convolution interpolation (Main-et al., 2017).

2.2.2. Atmospheric correction and preprocessing

The atmospheric correction was performed using ACOLITE software, specifically designed for aquatic applications (Vanhellemont, 2019). The Dark Spectrum Fitting (DSF) algorithm was applied to derive water-leaving reflectance (Rrs) values, accounting for atmospheric path radiance and aerosol contributions (Vanhellemont et al., 2021). Sun glint contamination, a persistent challenge in optical bathymetry, was mitigated using the ACOLITE's integrated sun glint correction module based on Cox and Munk's wave slope statistics.

The quality control procedures implemented included: (1) cloud masking using the Sentinel-2 Scene Classification Layer with pixels flagged as cloud or cloud shadow excluded from analysis; (2) land-water boundary delineation using the Normalized Difference Water Index (NDWI >0.3) to isolate water pixels (McFeeters, 1996); (3) outlier removal based on interquartile range filtering (values beyond Q1-1.5 × IQR or Q3+1.5 × IQR excluded) to eliminate anomalous reflectance values.

2.2.3. In-situ bathymetric data collection

In-situ bathymetric measurements were collected on December 10, 2017, eight days before the Sentinel-2 satellite overpass (December 18, 2017). While ideally satellite and field data should be acquired simultaneously, this 8-day temporal gap was considered acceptable given several mitigating factors: (1) Both dates occurred during the winter dry season when Sancha Lake experiences its most stable hydrological conditions with minimal rainfall and reduced runoff; (2) Local meteorological records confirm no significant precipitation events (>10 mm) during December 10–18, 2017; (3) Nearby hydrological station records indicate minimal water level variation (<0.15m) during this period, within the vertical accuracy of our survey equipment; (4) Winter water temperatures (<10 °C) suppress algal activity and reduce suspended sediment dynamics compared to summer conditions, contributing to stable optical properties. Nevertheless, we acknowledge that the temporal mismatch introduces potential uncertainties including minor water level fluctuations affecting shallow-water measurements, possible

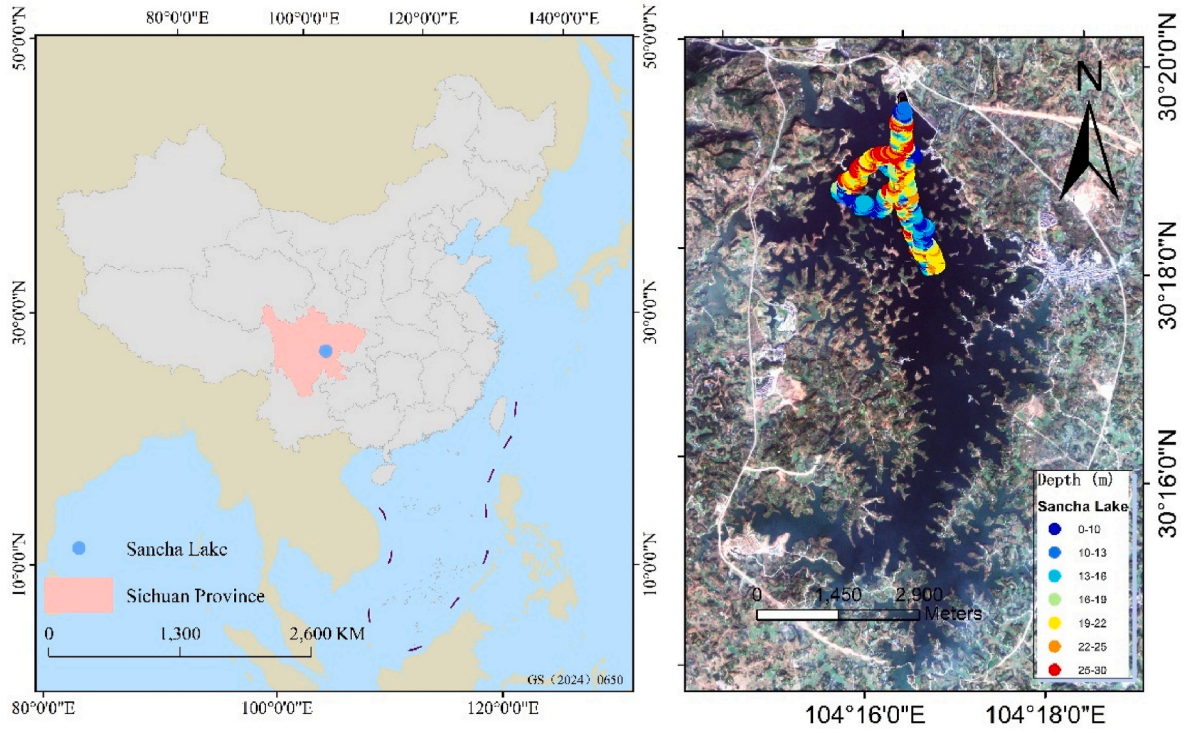

Fig. 1. (a) The geographical location of the study area, showing Sancha Lake's position in Sichuan Province within China's three-step topographic gradient descending from west to east; (b) Bathymetric sampling distribution with color gradient indicating depth variations, illustrating the dominant northwest-southeast orientation of the main deep channel aligned with regional topographic controls.

localized changes in suspended sediment distribution, and variations in atmospheric conditions. Post-processing included tide correction using nearby gauge station data and datum transformation to the 1985 National Height Datum of China, though we recognize this correction may not fully account for spatially heterogeneous water level variations across the lake's complex morphology. The final dataset was randomly split into training (70 %, n = 1400) and testing (30 %, n = 600) subsets using stratified sampling to maintain representative depth distributions.

A total of 2000 depth points were collected following a systematic grid pattern with 50m line spacing, ensuring comprehensive spatial coverage. Post-processing included tide correction using nearby gauge station data and datum transformation to the 1985 National Height Datum of China. The final dataset was randomly split into training (70 %, n = 1400) and testing (30 %, n = 600) subsets using stratified sampling to maintain representative depth distributions.

2.3. Spectral-geospatial XGBoost regression (SG-XGBoost) model development

2.3.1. Feature engineering framework

The SG-XGBoost model integrates comprehensive spectral transformations with explicit geospatial parameters through systematic feature engineering (Reichstein et al., 2019). The feature set encompasses two primary categories, namely spectral features and geospatial features. Spectral features include original reflectance bands (B1-B4, 10m resolution bands); band ratios (B1/B2, B2/B3, B3/B4) designed to address variable bottom albedo; spectral indices, specifically the Normalized Difference Water Index (NDWI), calculated as (B3-B4)/(B3+B4), and the Normalized Difference Vegetation Index (NDVI), calculated as (B4-B3)/(B4+B3); statistical measures across bands, including mean, standard deviation, and coefficient of variation; and non-linear transformations applied to each band, which are log (1+Bi), Bi^2, and $\sqrt{Bi}$. Geospatial features, by contrast, comprise geographic coordinates (longitude λ and latitude φ); spatial interactions ($\lambda \times \varphi$, λ^2, φ^2); distance metrics, namely the Euclidean distance from the lake center and the distance from the shoreline; and topographic derivatives, such as local spatial autocorrelation indices (Moran's I within 3 × 3 pixel windows).

2.3.2. XGBoost model and optimization

The XGBoost algorithm employs gradient boosting to sequentially construct an ensemble of regression trees, where each tree corrects residual errors from previous iterations (Chen et al., 2016). The objective function combines a differentiable loss function with regularization terms:

$$\mathrm{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{1}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{2}$$

where l represents the loss function, $\Omega(f)$ is the regularization term controlling model complexity,The SG-XGBoost model integrates comprehensive spectral transformations with explicit geospatial parameters through systematic feature engineering (Friedman, 2001).

To optimize the model's performance, Bayesian optimization via the Optuna framework was adopted for hyperparameter tuning, with a total of 100 iterations. The search space for model hyperparameters covered the following key parameters and their value ranges including n_estimators ranging from 100 to 1500, max_depth from 3 to 15, learning_rate from 0.001 to 0.3, subsample from 0.5 to 1.0, colsample_bytree from 0.5 to 1.0, reg_alpha from 0 to 10, and reg_lambda from 0 to 10. To prevent model overfitting and ensure the robustness of parameter selection, a 5-fold cross-validation approach was implemented.

2.4. Comparative methods

To comprehensively evaluate the differences in accuracy of the water depth estimation method proposed in this study, we conduct a comparative analysis between the model developed in this study and three other existing water depth estimation models. The specific models for comparison are as follows:

Stumpf Model: Employs a log-transformed band ratio approach (Stumpf et al., 2003):

$$z = m_1 \frac{\ln[nR(\lambda_i)]}{\ln[nR(\lambda_j)]} + m_0 \tag{3}$$

where n = 1000, $R(\lambda_i)$ and $R(\lambda_j)$ represent blue and green band reflectances.

Log-Linear Model: Based on Beer-Lambert law for light attenuation (Lyzenga, 1978):

$$z = a_1 \ln[L(\lambda_i) - L_\infty(\lambda_i)] + a_2 \ln[L(\lambda_j) - L_\infty(\lambda_j)] + a_3 \tag{4}$$

where L represents band radiance and L_∞ denotes deep water radiance.

Random Forest: An ensemble of 500 decision trees with maximum depth of 20, using the same feature set as SG-XGBoost for fair comparison.

2.5. Accuracy assessment metrics

Model performance was evaluated using standard bathymetric accuracy metrics:

Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n} |Z_i - Z_i'|}{n} \tag{5}$$

Root Mean Square Error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Z_i - Z_i')^2}{n}} \tag{6}$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Z_i' - Z_i)^2}{\sum_{i=1}^{n} (\bar{Z} - Z_i)^2} \tag{7}$$

Mean Relative Error:

$$MRE = \frac{\sum_{i=1}^{n} |(Z_i - Z_i')/Z_i'|}{n} \tag{9}$$

2.6. Bathymetry mapping workflow

Our bathymetric prediction method integrates remote sensing data with in-situ measurement data to develop an accurate and efficient water depth estimation model. The process first includes acquiring Sentinel-2 remote sensing images, which provide high-resolution spectral data of the study area. Subsequently, the images go through preprocessing including atmospheric correction and glint correction to eliminate atmospheric interference and surface reflections that may affect water depth estimation. After the preprocessing is completed, geometric correction is performed on the images to ensure precise spatial position matching.

The core of this method lies in extracting remote sensing reflectance (Rrs) from the corrected satellite images. These Rrs data are then matched with in-situ water depth measurement data to form a matched dataset, which is used to establish the correlation between spectral information and actual water depth. We collected 2000 in-situ water depth points using a multi-beam echo sounder system, providing a reliable measured dataset for model training and validation.

For water depth estimation, we adopted two parallel modeling approaches. The first approach is based on traditional bathymetric models, including the Stumpf model, log-linear model, and random forest model, all of which are trained using the matched Rrs data and in-situ water depth data. Additionally, we proposed an innovative water depth estimation model namely the spectral-geospatial XGBoost regression model (SG-XGBoost). Beyond integrating spectral band information, this model also includes the latitude and longitude information of each data point. Leveraging the gradient boosting framework of XGBoost, the new model improves prediction accuracy through iterative optimization. Meanwhile, it accounts for spatial variations in the relationship between spectral reflectance and water depth a factor often ignored by traditional models.

We applied both the traditional models and the SG-XGBoost model to estimate water depth across the entire study area. Specifically, the XGBoost model, which relies on an ensemble of decision trees constructed by gradient boosting, enables robust water depth prediction even under the noise and outlier interference commonly found in satellite images. In the subsequent step, we compared the reserved in-situ measurement data with the model prediction results to complete model validation. This validation process not only evaluates the accuracy and performance of each model but also focuses on analyzing the performance advantages of the SG-XGBoost model over traditional ones.

Finally, post-processing is performed on the water depth prediction results. First, tidal corrections conducted to eliminate the impact of water level changes between the satellite image acquisition period and the in-situ measurement period. After that, the corrected water depth data are converted into an image format, ultimately generating a complete bathymetric map of the study area (see Fig. 2).

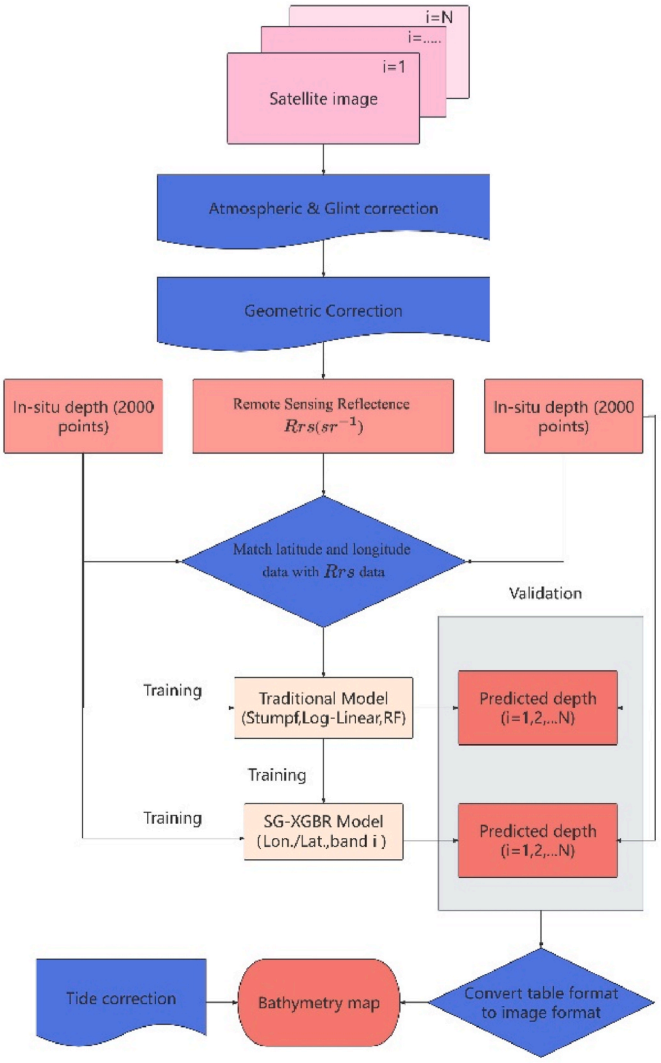

Fig. 2. The workflow of the Sancha lake bathymetry method.

3. Results

3.1. Comparative model performance

To present a comprehensive comparison of the bathymetric retrieval models, the following analysis of scatter plots comparing predicted depths against measured depths (Fig. 3). Among these traditional empirical models, the Stumpf model, despite its widespread application in bathymetric mapping, had only a low level of predictive capability, with a coefficient of determination R^2 of merely 0.02. This means the model could not explain the total water depth variance through its blue-green band ratio method. The Log-Linear model performed even worse, with an R^2 value of 0.01. Both models exhibited significant data scatter, a phenomenon of depth range prediction deviation. Their reliability decreased significantly, especially in Sancha Lake where the water depth gradient is relatively large. In contrast, the Random Forest model marked a substantial advancement over these empirical methods, delivering an R^2 of 0.86 and reducing systematic biases across depth ranges. Its ensemble structure, capable of capturing non-linear relationships between spectral signatures, spatial location and depth, enabled better performance in the mid-depth range 10–20 m where traditional models struggled, though it still faced limitations in the deepest regions exceeding 25 m where limited training samples n = 287 and weak optical signals increased prediction uncertainty, leading to subtle horizontal clustering of points around mean depth values.

The proposed SG-XGBoost model performed exceptionally well, with the R^2 achieving a value of 0.91 that reflects strong consistency between predicted and measured depths. Its scatter plot reveals tight clustering of points along the 1:1 line, reflecting consistent accuracy across the entire bathymetric range 2.3–31.7 m, a feature not observed in the other three models. It further demonstrated superiority in challenging transitional zones where rapid depth changes had previously confounded other approaches and delivered the lowest error metrics with a root mean square error RMSE of 1.66m and a mean absolute error MAE of 0.93m. This performance gain stems from the gradient boosting framework's iterative refinement process, which allowed SG-XGBoost to exploit subtle patterns in the spatial-spectral feature space that simpler models could not detect.

The bathymetric maps of each model (Fig. 4), reveal marked differences in underwater topography capture. Unlike traditional approaches that produced oversimplified results, SG-XGBoost delineated subtle bathymetric features at scales approaching the 10m pixel resolution, including submerged channels 10–30m wide, deltaic sediment deposits at dendritic tributary mouths, and gradual depth transitions reflecting historical lake littoral boundaries. it demonstrates enhanced ability to capture bathymetric variations at the pixel scale through effective integration of spatial context from surrounding pixels via coordinate-based features. The apparent detection of narrow linear features results from the model learning characteristic spatial signatures across multiple adjacent pixels rather than true sub-pixel resolution enhancement. The Log-Linear model had similar oversimplification plus additional artifacts linked to uncorrected atmospheric or sensor effects that empirical deep-water radiance correction failed to resolve.

3.2. Depth-stratified performance analysis

The depth-stratified analysis in Table 1 reveals performance variations among models across different depth ranges, with a key feature that SG-XGBoost maintained the lowest errors across the entire depth spectrum, whereas traditional models showed notable depth-dependent

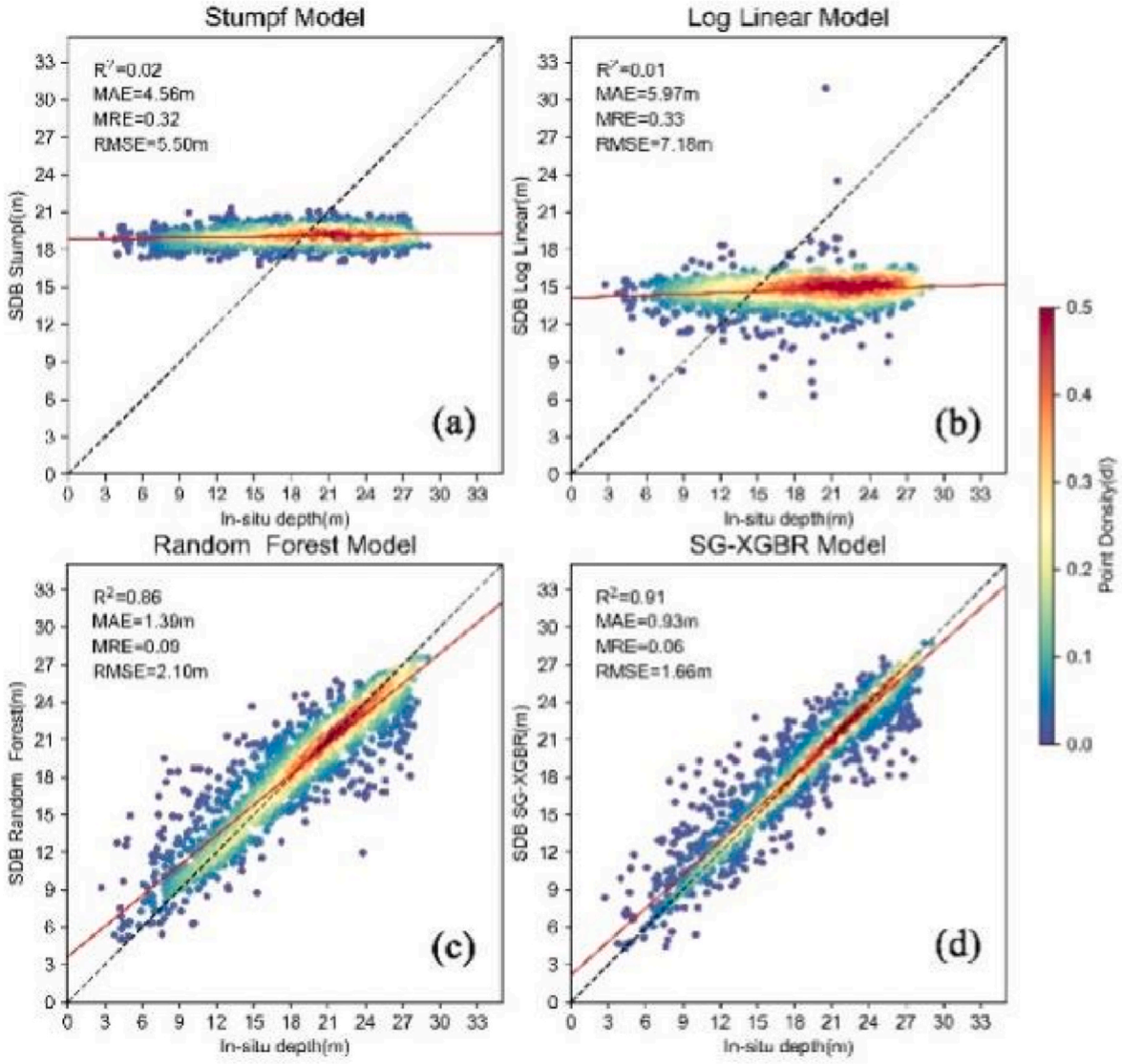

Fig. 3. Comparative analysis of bathymetric estimation methods.

performance decline (see Table 2).

Within the shallow water zone (spanning 0–10 m), a region where bottom reflectance and aquatic vegetation lead to complex spectral signatures, the SG-XGBoost model reached an RMSE of 2.70 m. The relatively higher error in this zone compared to deeper waters reflects the substantial spatial heterogeneity characteristic of littoral environments rather than model weakness. Shallow nearshore areas experience multiple confounding factors including: (1) mixed pixel effects from heterogeneous bottom substrates (rock, sediment, vegetation) within 10m Sentinel-2 pixels; (2) variable spectral signatures from emergent and submerged aquatic vegetation unrelated to depth; (3) higher turbidity variability from agricultural runoff and sediment resuspension; and (4) highly non-linear depth-reflectance relationships in very shallow waters (<3m) where optical saturation occurs. While we acknowledge the lack of comprehensive in-situ measurements of bottom type and vegetation distribution at individual sampling locations for definitive validation, these patterns are consistent with established characteristics of shallow inland waters. By comparison, the Stumpf model had an RMSE of 11 m, the Log-Linear model had an RMSE of 6.71 m, and the Random Forest model's RMSE was higher than SG-XGBoost's, all reflecting SG-XGBoost's better performance despite these challenges. This advantage originates not only from its capacity to integrate spatial context for distinguishing between depth-related spectral variations and those induced by bottom type or vegetation, thereby resolving uncertainties in single spectral signals, but also from the XGBoost gradient boosting framework's ability to model complex non-linear relationships between spectral features and depth (Traganos et al., 2018), a capability of critical value for applications like nearshore habitat mapping and development of tourism. If future research opportunities arise, we plan to conduct dedicated field campaigns collecting co-located measurements of depth, bottom type, vegetation coverage, and water quality parameters to explicitly test and validate these hypothesized mechanisms.

For the intermediate depth transition zone (covering 10–15 m), a region with the largest sample size (n = 331), where water column optical properties dominate and bottom influence fades due to limited light penetration, all models faced difficulties, even though bottom detection remained feasible. SG-XGBoost still delivered outstanding performance with an RMSE of 2.00 m. The Stumpf model's RMSE in this zone was 6.54 m, the Log-Linear model had an RMSE of 2.78 m, and the Random Forest model's RMSE was 2.25 m, all confirming SG-XGBoost's superior performance in this zone. The improved accuracy in this zone directly provides more reliable decision support for management scenarios such as reservoir volume calculation and sediment accumulation assessment; this performance advantage stems from the model's ability to adaptively weight spectral and spatial features based on local

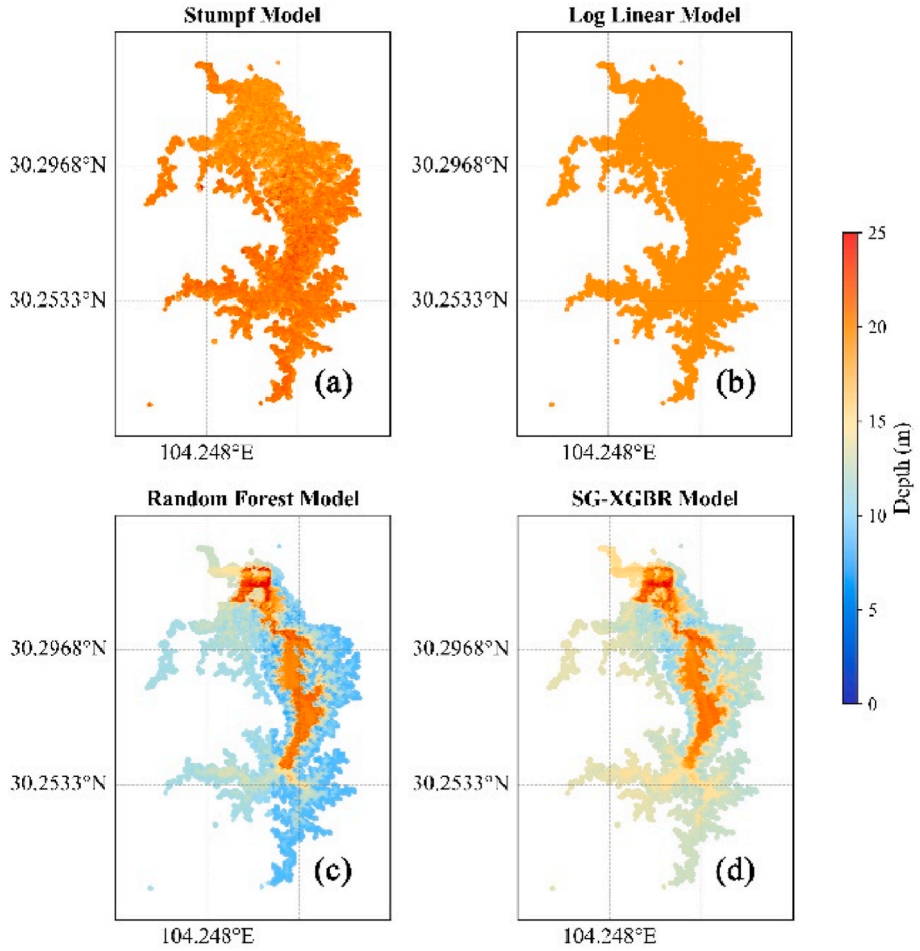

Fig. 4. Bathymetric Maps of the Sancha lake Using Various Estimation model(a) Stumpf Model(b) Log-Linear Model(c) Random Forest Model(d) Spectral-Geospatial XGBoost Regression (SG-XGBoost).

conditions.

Within the deep water zone (ranging from 15 to 20 m), where the main deep channel constitutes 27 % of the lake's total volume, all models reached their highest level of relative accuracy, yet notable differences in absolute errors still persisted. Under weak optical signal conditions, SG-XGBoost still achieved a low RMSE of 1.40 m. In contrast, the Stumpf model had an RMSE of 2.11 m, the Random Forest model had an RMSE of 1.72 m, and the Log-Linear model's RMSE was 3.65 m; all these values are higher than SG-XGBoost's, with the Log-Linear model's being notably higher. This result indicates that as optical signals approach noise levels, spatial patterns become increasingly crucial for bathymetric inversion, and SG-XGBoost can more sophisticatedly utilize these spatial relationships through the iterative refinement process of gradient boosting.

In the ultra-deep water zone (>20 m), where bottom reflection is negligible, depth estimation depends entirely on water column properties and spatial context, making optical bathymetry most challenging. Here, the performance gap among models further widened: traditional models essentially failed, with the Stumpf model's RMSE 5.13 m and the Log-Linear model's RMSE 9.45 m both approaching the range of depths being predicted, an indication of near-random predictions. The Random Forest model maintained reasonable performance with an RMSE of 1.98 m by leveraging spatial patterns, while SG-XGBoost achieved the best results with an RMSE of 1.44 m through the optimal integration of weak spectral signals and strong spatial priors learned from training data.

It is important to clarify the prediction mechanism in this depth zone: the model essentially performs sophisticated spatial interpolation informed by geomorphological structures encoded in coordinate features, rather than direct optical retrieval of bottom features. This represents a hybrid optical-spatial approach where, in shallow-moderate depths (<15m), the model combines optical signals with spatial context, while in deep waters (>20m), it transitions to spatial interpolation guided by morphological patterns learned from shallower training samples. This mechanism does not diminish the practical validity of predictions—spatial interpolation constrained by learned bathymetric structure is a legitimate and operationally valuable approach, conceptually similar to geostatistical methods like kriging but with advantages in capturing non-linear spatial patterns through gradient boosting. The key practical criterion is prediction accuracy for operational mapping, where SG-XGBoost's maintained RMSE = 1.44m represents a significant advantage over purely optical methods that fail completely in deep turbid waters. The lower RMSE of SG-XGBoost compared to Random Forest in deep waters underscores the advantage of gradient boosting's sequential error correction: this mechanism allows the model to concentrate its capacity on challenging samples that simpler ensemble methods may underweight, thereby sustaining reliable performance in ultra-deep waters.

3.3. *Feature importance analysis*

Fig. 5 displays the feature importance distribution calculated after training the SG-XGBoost model. Using standardized importance scores, this analysis quantifies the average gain in prediction accuracy contributed by each feature during the splitting process of all decision trees within the ensemble model. The importance score of the latitude-longitude interaction term reaches 0.62, nearly four times higher than that of any single spectral feature and even exceeding the combined

Table 1
Complete inventory of all features used in the RSG-XGBoost model.

Category	Feature Name	Description	Formula/Definition
Original Spectral Bands (n=10)	Band_1	Coastal Aerosol (443 nm)	Sentinel-2 Band 1 reflectance
	Band_2	Blue (490 nm)	Sentinel-2 Band 2 reflectance
	Band_3	Green (560 nm)	Sentinel-2 Band 3 reflectance
	Band_4	Red (665 nm)	Sentinel-2 Band 4 reflectance
	Band_5	Red Edge 1 (705 nm)	Sentinel-2 Band 5 reflectance
	Band_6	Red Edge 2 (740 nm)	Sentinel-2 Band 6 reflectance
	Band_7	Red Edge 3 (783 nm)	Sentinel-2 Band 7 reflectance
	Band_8	Near-Infrared (842 nm)	Sentinel-2 Band 8 reflectance
	Band_9	Water Vapor (945 nm)	Sentinel-2 Band 9 reflectance
	Band_10	SWIR-Cirrus (1375 nm)	Sentinel-2 Band 10 reflectance
Spectral Band Ratios (n=6)	Blue_Green_Ratio	Blue to Green reflectance ratio	Band_2/Band_3
	Red_Green_Ratio	Red to Green reflectance ratio	Band_4/Band_3
	NIR_Red_Ratio	NIR to Red reflectance ratio	Band_8/Band_4
	NIR_Green_Ratio	NIR to Green reflectance ratio	Band_8/Band_3
	RedEdge_Red_Ratio	Red Edge to Red reflectance ratio	Band_5/Band_4
	SWIR_NIR_Ratio	SWIR to NIR reflectance ratio	Band_10/Band_8
Spectral Indices (n=4)	NDWI	Normalized Difference Water Index	(Band_3 - Band_8)/(Band_3 + Band_8)
	NDVI	Normalized Difference Vegetation Index	(Band_8 - Band_4)/(Band_8 + Band_4)
	EVI	Enhanced Vegetation Index	2.5 × (Band_8 - Band_4)/(Band_8 + 6 × Band_4–7.5 × Band_2 + 1)
	MNDWI	Modified NDWI	(Band_3 - Band_10)/(Band_3 + Band_10)
Non-linear Band Transformations (n=10)	Band_1_Squared	Squared Coastal Aerosol band	$Band\_1^2$
	Band_2_Squared	Squared Blue band	$Band\_2^2$
	Band_3_Squared	Squared Green band	$Band\_3^2$
	Band_4_Squared	Squared Red band	$Band\_4^2$
	Band_5_Squared	Squared Red Edge 1 band	$Band\_5^2$
	Band_1_Log	Natural logarithm of Coastal Aerosol	ln (Band_1 + ε)
	Band_2_Log	Natural logarithm of Blue band	ln (Band_2 + ε)
	Band_3_Log	Natural logarithm of Green band	ln (Band_3 + ε)
	Band_4_Log	Natural logarithm of Red band	ln (Band_4 + ε)
	Band_8_Log	Natural logarithm of NIR band	ln (Band_8 + ε)
Statistical Band Features (n=3)	Band_Mean	Mean reflectance across all bands	(Band_1 + Band_2 + … + Band_10)/10
	Band_Sum	Sum of all band reflectances	Band_1 + Band_2 + … + Band_10
	Band_StdDev	Standard deviation across all bands	σ(Band_1, Band_2, …, Band_10)
Primary Geospatial Features (n=4)	Longitude	Geographic longitude coordinate	UTM-projected X coordinate (m)
	Latitude	Geographic latitude coordinate	UTM-projected Y coordinate (m)
	Dist_From_Origin	Empirical distance from reference point	$\sqrt{(Longitude^2 + Latitude^2)}$
	Along_Channel_Distance	Distance along main river channel	Cumulative distance from upstream reference
Geospatial Interaction Terms (n=3)	Lon_Lat_Product	Interaction between longitude and latitude	Longitude × Latitude
	Lon_Squared	Squared longitude	$Longitude^2$
	Lat_Squared	Squared latitude	$Latitude^2$
Derived Spatial Features (n=7)	Local_Slope	Local terrain gradient	Computed from DEM
	Distance_to_Bank	Distance to nearest river bank	Euclidean distance (m)
	Channel_Width	Local channel width	Cross-sectional width (m)
	Curvature	Local channel curvature	Radius of curvature (m^{-1})
	Flow_Direction	Primary flow direction angle	Azimuth (degrees)
	Sinuosity	Local channel sinuosity	Channel length/straight-line distance
	Braiding_Index	Multi-channel braiding intensity	Number of active channels
Total Features	**n = 47**		

Notes.
$\varepsilon = 1 \times 10^{-6}$ added to prevent log(0) errors in logarithmic transformations.
UTM projection used for Longitude and Latitude to ensure metric consistency.
All reflectance values normalized to [0, 1] range prior to feature computation.
Geospatial features extracted from 30-m resolution DEM and river centerline vector data.
Feature correlation filtering applied: features with $|r| > 0.85$ excluded to prevent multicollinearity.

predictive power of all spectral features collectively, challenging our fundamental assumptions in the traditional field of optical bathymetry.

The next most influential features are the longitude term (importance score = 0.33) and the latitude term (importance score = 0.26). This notable dominance of spatial features reflects two critical aspects of Sancha Lake's morphology. On one hand, it mirrors the northwest-southeast orientation of the lake's key morphological characteristics, including the main navigation channel distributed along the pre-impoundment river valley and the tributary valleys that are at right angles to the main channel, forming the lake's branch-like structure. On the other hand, it embodies the underlying geological structure and sedimentation patterns of the lake basin (Su et al., 2023). Turning to spectral features, the 8th band exhibits the highest importance (0.18), indicating that the near-infrared band possesses stronger penetration capabilities in deep water. This property balances the band's sensitivity across a wide range of water depths while maintaining sufficient signal strength for reliable measurements, thanks to its excellent performance in complex water bodies. By contrast, coastal water environments typically rely on blue and green bands as the dominant signals for bathymetry. Sancha Lake, however, has high concentrations of dissolved organic matter (DOM), which preferentially absorbs shorter wavelengths such as blue light. For this reason, the longer wavelength of the

Table 2
A comparison of the RMSE errors for different water depths and different bathymetry methods.

Training Method	RMSE				
	0–10m (153Points)	10–15 m (331 Points)	15–20 m (549 Points)	>20m (967 Points)	Overall (2000 Points)
Stumpf	11	6.54	2.11	5.13	5.5
Log-Linear	6.71	2.78	3.65	9.45	7.18
Random Forest	3.4	2.25	1.72	1.98	2.10
SG-XGBoost	2.70	2.00	1.40	1.44	1.66

band 8 is more suitable for bathymetric inversion in the lake's complex water conditions.

3.4. Residual error distribution analysis

The residual error analyses, which are conducted through histograms and spatial distribution maps, are presented in Fig. 6, revealing key patterns in model performance that guide result interpretation and identify areas for potential improvement. The residual histogram of the SG-XGBoost model shows a near-normal distribution centered at −0.08 m with a standard deviation of 1.66 m, indicating slight systematic underestimation but generally unbiased predictions across the depth range.This narrow, symmetric distribution contrasts sharply with the positively skewed distributions of empirical models such as the Stumpf model and the Log-Linear model; the former has a mean residual of 1.83 m and a skewness of 1.24 while the latter exhibits a mean residual of 2.14 m and a skewness of 1.41, both consistently underestimating depths especially in deeper waters where optical signals weaken. The Random Forest model displays intermediate residual characteristics, with a mean of 0.43 m and a standard deviation of 3.12 m, and its slight positive skew suggests minor systematic overestimation.

Spatial mapping of residuals further reveals geographic patterns in prediction errors, providing insights into model limitations and environmental factors affecting performance. The SG-XGBoost model maintains a relatively uniform error distribution across the lake, though slightly elevated errors concentrate in three specific areas, namely the northwestern tributary mouth, where high turbidity from agricultural runoff creates challenging optical conditions with Secchi depth below 1 m during rain events; the transition zone between shallow eastern platforms and the central deep basin, where rapid depth changes over distances smaller than the 10 m pixel resolution generate mixed pixels that hinder depth retrieval; and isolated deep holes exceeding 28 m, where limited training samples and negligible optical signals force the model to rely entirely on spatial interpolation. These error patterns, while present, are substantially less pronounced than those in other models, where large contiguous areas of systematic overestimation or underestimation produce unreliable bathymetric maps.

Beyond the combined insights gleaned from histograms and spatial maps, residual histograms in Fig. 5 further emphasize critical differences in model precision that matter for operational use. The SG-XGBoost model, which exhibits a near-normal distribution centered near zero (mean absolute error = 0.94 m, standard deviation = 1.65 m), confirms its unbiased prediction capability, which is a clear distinction from traditional models that show pronounced positive skewness and consequent systematic depth underestimation. A key metric of practical reliability further highlights SG-XGBoost's superiority, namely that 95 % of its errors fall within ±4.8 m, a much narrower range than the ±11 m observed for the Stumpf model. This enhanced precision validates SG-XGBoost's greater reliability for operational bathymetric mapping applications as supported by prior research.

Spatial mapping of residuals further reveals geographic patterns in prediction errors, providing insights into model limitations and environmental factors affecting performance. The SG-XGBoost model maintains a relatively uniform error distribution across the lake, though slightly elevated errors concentrate in three specific areas: (1) the northwestern tributary mouth, where high turbidity from agricultural

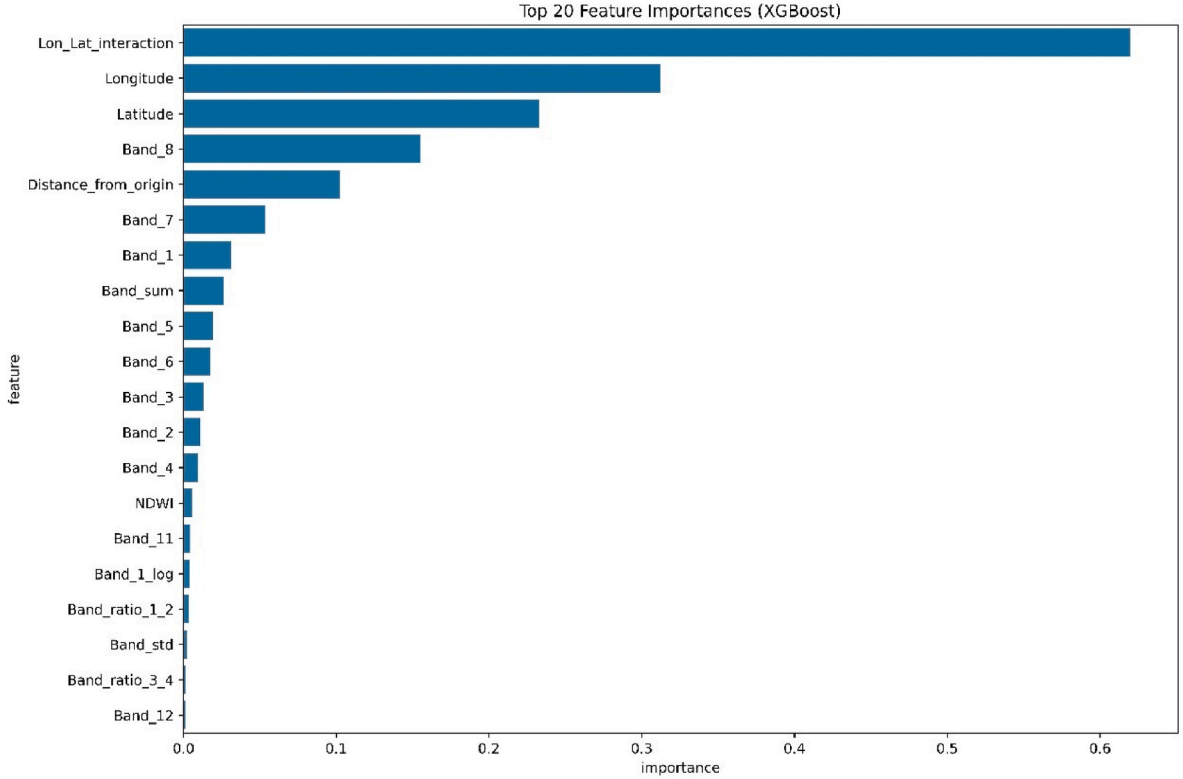

Fig. 5. Feature importance analysis from the SG-XGBoost model showing the relative contribution of each feature to bathymetry prediction.

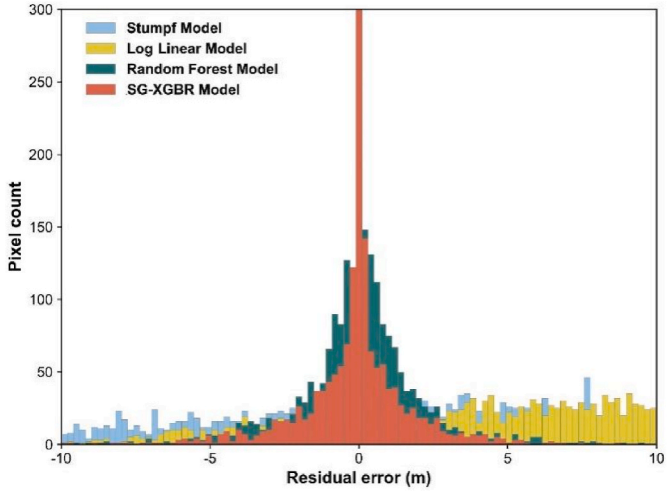

Fig. 6. The histogram map of the residual error obtained from different bathymetry methods.

runoff creates challenging optical conditions; (2) the transition zone between shallow eastern platforms and the central deep basin, where rapid depth changes generate mixed pixels; and (3) isolated deep holes exceeding 28 m, where limited training samples force spatial interpolation. These error patterns, while present, are substantially less pronounced than those in other models, where large contiguous areas of systematic overestimation or underestimation produce unreliable bathymetric maps.

Regarding spatial autocorrelation in residuals, we acknowledge that formal quantitative assessment using Moran's I or semivariogram analysis would strengthen validation. While we did not originally compute these statistics, several lines of evidence suggest limited problematic spatial bias: (1) Visual inspection of spatial residual patterns (Fig. 5) reveals relatively dispersed errors without large contiguous regions of systematic bias, indicating the model captures broad spatial trends effectively; (2) Our 5-fold cross-validation with spatially distributed training/testing splits demonstrates consistent accuracy (R^2 = 0.91) across geographically separated test points—if substantial residual spatial autocorrelation indicated overfitting to training locations, we would expect degraded performance on distant test samples; (3) Some degree of residual spatial autocorrelation is inherently expected when deliberately using spatial features for prediction, reflecting the model's intended exploitation of geomorphological structure rather than a flaw. The critical diagnostic is distinguishing between residual patterns indicating unmodeled spatial structure (underfitting) versus overlearning of specific training locations (overfitting). The combination of visual residual dispersion and maintained cross-validation accuracy suggests our model achieves appropriate balance. Nevertheless, we recognize that computing formal spatial autocorrelation statistics (Moran's I, Geary's C, semivariograms) for residuals represents an important methodological enhancement for future applications of spatial machine learning to bathymetry, providing quantitative confirmation of spatial bias characteristics.

4. Discussion

4.1. Transformation in bathymetric remote sensing

In our SG-XGBoost model, the significant importance of geospatial features compared with spectral variables has changed our understanding of the remote sensing of inland water depth. Previously, the core of research had always been optimizing the performance of spectral band combinations and atmospheric correction tools, which has overturned the traditional understanding that water depth is estimated based on the law of light attenuation in water. Furthermore, it also indicates that for complex-shaped inland water bodies like Sancha Lake, geographic location can often provide more reliable water depth information than optical measurement data.

The next most influential features are the longitude term (importance score = 0.33) and the latitude term (importance score = 0.26), with longitude exhibiting notably higher predictive power than latitude. This asymmetry reflects fundamental geomorphological controls rooted in China's regional topography and Sancha Lake's specific valley orientation.

Geomorphological Justification for Longitude Dominance: China's terrain exhibits a systematic three-step topographic gradient descending from west to east: the Tibetan Plateau (>4000m elevation) steps down through mountainous regions (1000–2000m) to coastal plains (<500m). This fundamental west-to-east gradient, created by Cenozoic tectonic uplift and subsequent fluvial erosion, governs river systems and sediment transport throughout the Yangtze watershed. Sancha Lake, located in the Upper Yangtze region of Sichuan Province, occupies a valley that reflects this broader topographic pattern. The pre-impoundment river channel flowed predominantly east-southeast, carving a deep valley aligned with the regional slope that now forms the lake's primary bathymetric axis.

Longitude serves as an effective proxy for this stable longitudinal valley structure: depth systematically increases moving eastward along the 15 km main channel from shallow western tributaries (~5m) to the deepest eastern basin (~32m near the dam). This relationship persists despite 45 years of reservoir operation because the underlying geological structure—bedrock valley walls and thalweg alignment—remains stable even as surficial sediments accumulate. The gradient boosting algorithm learns this consistent longitude-depth relationship through the training data, effectively using longitude as a surrogate for distance along the primary valley axis.

In contrast, latitude captures secondary tributary valleys oriented roughly perpendicular (north-south) to the main channel. While these tributaries create bathymetric variations across latitudes, they exhibit less systematic depth patterns compared to the dominant longitudinal valley. Furthermore, latitude is more strongly associated with temporally variable water quality conditions: turbidity and dissolved organic matter (DOM) concentrations vary more across latitudes due to differential tributary inputs, with northern tributaries receiving agricultural runoff with high suspended sediment loads (Secchi depth <1m during rainfall) while southern areas remain clearer (Secchi depth 2–3m). These latitudinal water quality gradients are seasonally dynamic, changing with rainfall and agricultural cycles, making latitude-based spectral relationships less stable for bathymetric prediction.

Longitude thus provides more reliable depth prediction than latitude or spectral features because it correlates with permanent topographic controls (valley morphology) rather than transient water quality conditions (turbidity, DOM). The XGBoost model effectively learns that longitude offers consistent depth information regardless of seasonal spectral variations, explaining its 27 % higher feature importance (0.33) compared to latitude (0.26). This finding has important implications for transfer learning: while absolute longitude values are site-specific and non-transferable, the principle of identifying stable geomorphological axes as primary features could generalize to other reservoirs with strong directional valley structure.

In the research of this paper, geographic coordinates can not only capture the unique water depth distribution patterns of specific regions, such as those formed by the dendritic tributary system of the lake but also mask the dynamic water quality conditions that are directly associated with spectral data and water depth. Even when the optical signal is poor, they can help us obtain reliable results of indirect water depth estimation. The importance of geospatial features stems from the unique combination of three factors in Sancha Lake, that is its specific terrain, anthropogenic factors, and optical characteristics.

The water depth distribution of this lake retains the valley

morphology before the dam was built, and the influence of this terrain remains highly stable. Even after 45 years of sedimentation, water level fluctuations, and changes in lake morphology, the horizontal position is still associated with the stable form of the underground geological structure. This characteristic makes geographic coordinates a basis for water depth estimation.

The water depth distribution of this lake retains the valley morphology before the dam was built, and the influence of this terrain remains highly stable. Even after 45 years of sedimentation, water level fluctuations, and changes in lake morphology, the horizontal position is still associated with the stable form of the underground geological structure. This characteristic makes geographic coordinates a basis for water depth estimation. The XGBoost model, when combined with location-based features, can effectively capture this information. Moreover, affected by the content of substances in the water, the lake also has a complex optical environment.

In complex water areas like Sancha Lake, the inversion capability of traditional water depth inversion models is significantly limited, making it difficult for them to deal with the complex optical and topographic conditions in the water. In contrast, the SG-XGBoost model we adopted, with its strong nonlinear mapping capability, can effectively handle the intertwined nonlinear relationships among multiple factors in complex water areas, providing a more suitable algorithmic basis for water depth inversion. At the same time, the near-infrared band exhibits unique advantages in this water area. Specifically, its penetration ability in water is basically unaffected by changes in chlorophyll concentration, and it can maintain stable optical propagation characteristics even when facing the impacts of water quality variations in different seasons and regions of Sancha Lake. The combination of this characteristic and the advantages of the SG-XGBoost model further enhances the application value of the near-infrared band in water depth inversion of Sancha Lake and significantly improves the overall inversion accuracy.

Regarding the concern about distinguishing genuine spatial patterns from site-specific overfitting, several factors support the validity of our geospatial features:

Empirical Feature Engineering Framework: In machine learning applications, particularly tree-based models like XGBoost, features do not necessarily require strict physical interpretation. The model learns non-linear relationships and interactions that may not be immediately intuitive but contribute to predictive performance. This represents a fundamental distinction between data-driven and physics-based modeling paradigms.

Capturing Geographic Patterns: The longitude-latitude interaction effectively encodes the systematic northwest-southeast orientation of Sancha Lake's main bathymetric channel following the pre-impoundment river valley, with perpendicular tributary valleys creating the lake's dendritic structure. This geospatial feature captures genuine morphological trends that persist across temporal variations in water level and sedimentation patterns.

Mixed Pixel Context: In remote sensing applications, each pixel represents mixed signals from suspended sediments and other water constituents. Geographic trend features enable XGBoost to implicitly account for spatial variation patterns in these mixed pixel compositions, aligned with dominant flow directions and sediment transport pathways.

Cross-Validation Evidence: Our 5-fold cross-validation with spatially distributed splits demonstrates the model maintains consistent accuracy ($R^2 = 0.91$) across test points geographically separated from training data. This indicates the model learns transferable spatial relationships rather than memorizing specific coordinate-depth pairs. While ideal validation would involve training on one lake region and testing on another entirely separate region without coordinate information, the maintained performance across spatially distinct folds provides evidence against simple overfitting.

Model Performance Justification: The empirical value of including geospatial features is demonstrated through improved model accuracy, indicating successful capture of meaningful geographic variation patterns relevant to the prediction task. While we acknowledge the site-specificity limitation discussed in Section 4.3, within Sancha Lake these features encode genuine bathymetric structure rather than spurious correlations.

4.2. Advantages of gradient boosting framework

Based on the same feature dataset, the SG-XGBoost algorithm outperforms Random Forest, highlighting the advantages of gradient boosting algorithms in bathymetric prediction for complex water bodies. These advantages stem from three key mechanisms.

A core strength of SG-XGBoost lies in its iterative learning framework, which prioritizes samples with prediction errors to iteratively enhance predictive accuracy. In bathymetric estimation, errors tend to cluster in specific regions such as deep-water areas and turbid waters, and these regions precisely demand targeted model optimization. Unlike Random Forest, which constructs trees in parallel, SG-XGBoost addresses clustered errors through its error-focused correction mechanism, a capability critical to achieving robust bathymetric mapping.

Another pivotal characteristic of SG-XGBoost resides in its regularization design, incorporating built-in L1 (Lasso) and L2 (Ridge) mechanisms. These not only effectively suppress overfitting but also retain sufficient model complexity to capture subtle variations in bathymetric data. This balance is particularly vital when processing noisy satellite imagery, as it mitigates the impact of atmospheric interference such as haze and aerosols, as well as water surface interference such as waves and glare, on spectral measurements, thereby preventing overfitting from undermining the stability of bathymetric predictions.

Additionally, the tree-based structure of SG-XGBoost enables it to capture complex feature interactions without manual specification. In this study, the model revealed that interactive features between longitude and latitude contain more informative content than individual coordinate features, a result that directly indicates the existence of nonlinear spatial patterns in lake bathymetry [50]. This ability to autonomously extract meaningful feature correlations gives SG-XGBoost a distinct edge over Random Forest, which lacks comparable proficiency in modeling such nonlinear dependencies.

4.3. Limitations and future directions

While the SG-XGBoost model demonstrates good performance in water depth estimation for Sancha Lake, it has several inherent limitations that hold significant reference value for clarifying its current application scope and future research directions. Constrained by optical penetration depth, when the water depth exceeds 25 m, bottom reflection becomes negligible, and water depth prediction relies solely on water column properties and spatial interpolation, leading to a decline in model performance. In extremely turbid conditions such as flood periods or algal blooms, optical methods fail completely regardless of algorithmic improvements, which means active sensors like Light Detection and Ranging (LiDAR) or Synthetic Aperture Radar (SAR) need to be integrated to expand mapping capabilities. Another key limitation of the model is its site specificity, which originates from the model's reliance on absolute geographic coordinates. The association between latitude-longitude interaction terms and water depth identified by the model originates from the unique morphological characteristics and geological background of Sancha Lake, making it difficult to transfer the model to other water bodies. Transfer learning approaches that use the Sancha Lake model as a base and perform calibration with local data still require further research to determine the optimal strategy. Additionally, static bathymetric maps fail to capture sediment dynamics, morphological changes, or temporal instability of reservoir systems on inter-annual to decadal time scales. Therefore, conducting multi-temporal analysis using time series of satellite imagery or realizing regular model updates through periodic field truth data collection is crucial for

maintaining long-term operational reliability of the model and improving its predictive capabilities.

Beyond addressing the aforementioned limitations, the methodological framework of SG-XGBoost holds broad application potential in the field of aquatic environment remote sensing, and its impact can extend far beyond bathymetric mapping itself. In terms of water quality monitoring, this framework can be used to retrieve concentrations of chlorophyll-a, suspended sediments, and dissolved organic matter. In optically complex waters where traditional bio-optical algorithms fail to perform effectively, integrating spatial information is expected to improve retrieval accuracy. In habitat mapping, water depth is a key factor determining the distribution of aquatic habitats, and the high-resolution bathymetric maps generated by SG-XGBoost can provide important support for species distribution modeling and conservation planning. In climate change assessment, long-term water depth monitoring based on this model can capture sedimentation rates and the responses of water body morphology to changes in precipitation patterns. Meanwhile, the efficiency of satellite-based mapping enables regular data updates, which is crucial for identifying long-term, gradual changes that are ecologically significant.

Based on the same feature dataset, the SG-XGBoost algorithm outperforms Random Forest, highlighting the advantages of gradient boosting algorithms in bathymetric prediction for complex water bodies. These advantages stem from three key mechanisms that are intrinsic to the gradient boosting framework rather than simply hyperparameter tuning differences. To ensure fair comparison, both models underwent rigorous hyperparameter optimization using Bayesian optimization (Optuna framework, 100 iterations) with 5-fold cross-validation, with Random Forest optimized over comparable parameter search spaces.

A core strength of SG-XGBoost lies in its iterative learning framework, which prioritizes samples with prediction errors to iteratively enhance predictive accuracy. Unlike Random Forest, which constructs trees independently in parallel and aggregates their predictions through simple averaging, XGBoost builds trees sequentially, with each subsequent tree explicitly focusing on correcting the residual errors of the ensemble built so far. In bathymetric estimation, errors tend to cluster in specific challenging regions such as deep-water areas where optical signals are weak, turbid waters where suspended sediments interfere with spectral measurements, and transitional zones where rapid depth changes occur. These regions precisely demand targeted model optimization. XGBoost addresses these clustered errors through its error-focused correction mechanism by assigning higher weights to difficult samples during each boosting iteration, forcing the model to concentrate learning capacity on previously misclassified regions. This sequential refinement capability is critical to achieving robust bathymetric mapping in optically complex inland waters.

Another pivotal characteristic of SG-XGBoost resides in its regularization design, incorporating built-in L1 (Lasso) and L2 (Ridge) regularization terms directly in its objective function: $Obj = L(y, \hat{y}) + \Omega(f)$, where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \Sigma w^2$. These regularization mechanisms not only effectively suppress overfitting to training noise but also retain sufficient model complexity to capture subtle variations in bathymetric data—a delicate balance. This is particularly vital when processing noisy satellite imagery, as it mitigates the impact of atmospheric interference such as haze and aerosols, as well as water surface interference such as waves and sun glint, on spectral measurements, thereby preventing overfitting from undermining the stability of bathymetric predictions. In contrast, Random Forest relies primarily on tree pruning and ensemble averaging for regularization, which provides less direct control over the bias-variance tradeoff.

Additionally, the tree-based structure of SG-XGBoost combined with gradient boosting enables it to capture complex feature interactions without manual specification. The gradient boosting mechanism, through its additive model structure $f(x) = \Sigma f_k(x)$, where each f_k is optimized to reduce residuals via gradient descent on the loss function, naturally discovers and exploits higher-order interactions between features. In this study, the model revealed that interactive features between longitude and latitude contain more informative content than individual coordinate features (62 % importance for $\lambda \times \varphi$ vs. 33 % for λ alone and 26 % for φ alone), a result that directly indicates the existence of nonlinear spatial patterns in lake bathymetry aligned with the northwest-southeast valley orientation and perpendicular tributary structure. This ability to autonomously extract meaningful feature correlations through the sequential boosting process gives SG-XGBoost a distinct edge over Random Forest, which, despite handling feature interactions through individual tree splits, lacks the focused optimization capacity to systematically exploit such nonlinear dependencies across the entire ensemble. The 21 % RMSE improvement (from 2.10m to 1.66m) thus fundamentally arises from these algorithmic mechanisms inherent to gradient boosting rather than simply more aggressive hyperparameter tuning.

Additional limitations include temporal uncertainty from the 8-day gap between field survey and satellite acquisition, which, despite occurring during stable winter conditions, may introduce minor depth errors from water level fluctuations or optical property changes. Future studies should prioritize simultaneous or near-simultaneous data acquisition, ideally within 24–48 h, and incorporate continuous water level monitoring to enable more precise temporal corrections.

Future research could explore integration with emerging technologies including hyperspectral imaging systems and advanced deep learning architectures such as transformer-based models. These approaches could potentially extract more sophisticated spectral-spatial relationships, particularly in optically complex waters. However, such advances must be balanced against practical considerations including data availability, computational requirements, and transferability to operational monitoring programs in developing regions where Sentinel-2 remains the most accessible high-resolution multispectral platform.

5. Conclusion

This study confirms that the Spectral-Geospatial Extreme Gradient Boosting Regression (SG-XGBoost) model has significant application potential in the bathymetric inversion of complex inland waters. In the bathymetric inversion experiment of Sancha Lake, this model outperforms traditional empirical methods and conventional machine learning algorithms, greatly improving inversion accuracy. The overall root mean square error (RMSE) is as low as 1.66 m, representing a 21 % improvement in accuracy compared with the traditional random forest algorithm.

SG-XGBoost is built on a gradient boosting framework. Its unique iterative error correction mechanism and sophisticated regularization methods enable it to exhibit high effectiveness in uncovering the spatial distribution patterns of water depth. Meanwhile, the model maintains high inversion accuracy across different water depth ranges. This characteristic fully reflects its stability and further proves its applicability in practical scenarios, especially in cases where on-site data collection is constrained by both cost and spatial limitations. Furthermore, SG-XGBoost can leverage freely available Sentinel-2 images to generate high-resolution bathymetric maps. This advantage allows underdeveloped regions to easily access critical water depth information, which holds profound significance for local water resource management and environmental monitoring work. By regularly updating water depth data, the model can also be widely applied to various practical fields, such as sediment budget assessment, cage culture, development of tourism and ecological habitat mapping.

Currently, the model still has obvious limitations. On one hand, the "regional specificity" of spatial features restricts the model's transferability, making it difficult to apply directly to other regions. Therefore, the development of more robust transfer learning methods is urgently required. On the other hand, the specific impacts of seasonal changes in water quality and long-term geomorphic evolution on model performance remain unclear, and the temporal stability of the model

under dynamic environmental conditions still needs further verification. In the future, if SG-XGBoost can be integrated with other complementary data sources such as synthetic aperture radar (SAR) with all-weather monitoring capabilities and lidar data from the Ice, Cloud, and land Elevation Satellite-2 (ICESat-2), which enables validation under optically complex conditions, to construct a multi-source data fusion-based bathymetric inversion model, it is expected to further enhance the accuracy and applicability of bathymetric mapping.

Notably, the SG-XGBoost framework opens up a new path for the development of quantitative remote sensing technology in aquatic environments. The experience of successfully integrating spatial context information with spectral information can be further applied to related applications, such as water quality parameter inversion, benthic habitat classification, and underwater vegetation mapping, providing technical support for aquatic environment research. With the increasing impacts of climate change and human activities on freshwater resources, SG-XGBoost has become an important technical tool for understanding and managing key aquatic ecosystems.

In conclusion, the SG-XGBoost developed in this study provides a practical and effective solution for the field of bathymetric remote sensing. The excellent performance demonstrated by the model in the Sancha Lake experiment makes it a transformative tool for practical bathymetric mapping of inland waters. In the future, by continuously addressing the model's limitations and deeply integrating advanced machine learning technologies with remote sensing expertise, our ability to perceive and understand aquatic environments will undoubtedly be further enhanced.

CRediT authorship contribution statement

Xiaojuan Li: Writing – original draft, Resources, Investigation, Funding acquisition. **Wei Zhang:** Writing – review & editing, Investigation, Funding acquisition, Zhihua Mao, Funding acquisition, Conceptualization. **Hongrui Zheng:** Writing – review & editing, Investigation. **Zhongqiang Wu:** Visualization, Software, Data curation, Conceptualization. **Hongliang Lu:** Writing – review & editing.

Funding

This work was supported by the 2023 Hainan Province "South China Sea New Star" Science and Technology Innovation Talent Platform Project (NHXXRCXM202316),in part by Hainan Natural Science Foundation of China (No.424QN253,No.620RC602),National Natural Science Foundation of China (No.61966013),in part by the National Natural Science Foundation of China (42401411, 61991454); the Sichuan Science and Technology Program (2024NSFSC0782); the Natural Science Foundation of Jiangsu Province (BK20240282); in part by the National Natural Science Foundation of China under Grant 61991454, in part by the National Key Research and Development Program of China under Grant 2023YFC3107605, in part by the Oceanic Interdisciplinary Program of Shanghai Jiao Tong University under Grant SL2022ZD206 , and in part by the Scientific Research Fund of Second Institute of Oceanography, MNR under Grant SL230, and in part by the 2024 College Students Innovation and Entrepreneurship Training Program Project S202411658034.

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.

Acknowledgments

The authors sincerely thank the European Space Agency (ESA) for providing high-quality Sentinel-2 remote sensing imagery. As a core of the Copernicus Program, Sentinel-2 high resolution, multi-spectral capability, and free access were pivotal to our research.

Data availability

Data will be made available on request.

References

Agrafiotis, P., et al., 2024. MAGICBATHYNET: a multimodal remote sensing dataset for bathymetry prediction and pixel-based classification in shallow waters. In: IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE.
Agrafiotis, P., Demir, B., 2025. Deep learning-based bathymetry retrieval without in-situ depths using remote sensing imagery and SfM-MVS DSMs with data gaps. ISPRS J. Photogrammetry Remote Sens. 225, 341–361.
Al Najar, M., et al., 2023. Satellite derived bathymetry using deep learning. Mach. Learn. 112 (4), 1107–1130.
Alevizos, E., 2020. A combined machine learning and residual analysis approach for improved retrieval of shallow bathymetry from hyperspectral imagery and sparse ground truth data. Remote Sens. 12 (21), 3489.
AlHossainy, R.H., et al., 2025. Inferring bathymetry from Sentinel-2 satellite images using machine learning algorithms based on chlorophyll concentration data in the absence of ground measurement. Arabian J. Sci. Eng. 1–18.
Ashphaq, M., et al., 2021. Review of near-shore satellite derived bathymetry: classification and account of five decades of coastal bathymetry research. J. Ocean Eng. Sci. 6 (4), 340–359.
Barnes, B.B., et al., 2018. Multi-band spectral matching inversion algorithm to derive water column properties in optically shallow waters: an optimization of parameterization. Rem. Sens. Environ. 204, 424–438.
Benshila, R., et al., 2020. A deep learning approach for estimation of the nearshore bathymetry. J. Coast Res. 95 (SI), 1011–1015.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining.
Chen, B., et al., 2019. A dual band algorithm for shallow water depth retrieval from high spatial resolution imagery with no ground truth. ISPRS J. Photogrammetry Remote Sens. 151, 1–13.
Drusch, M., et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Rem. Sens. Environ. 120, 25–36.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Gao, J., 2009. Bathymetric mapping by means of remote sensing: methods, accuracy and limitations. Prog. Phys. Geogr. 33 (1), 103–116.
Garcia, R.A., et al., 2020. Benthic classification and IOP retrievals in shallow water environments using MERIS imagery. Rem. Sens. Environ. 249, 112015.
Hedley, J.D., et al., 2016. Remote sensing of coral reefs for monitoring and management: a review. Remote Sens. 8 (2), 118.
Huang, R., et al., 2017. Bathymetry of the coral reefs of Weizhou Island based on multispectral satellite images. Remote Sens. 9 (7), 750.
Kerr, J.M., Purkis, S., 2018. An algorithm for optically-deriving water depth from multispectral imagery in coral reef landscapes in the absence of ground-truth data. Rem. Sens. Environ. 210, 307–324.
Kutser, T., et al., 2020. Remote sensing of shallow waters–A 50 year retrospective and future directions. Rem. Sens. Environ. 240, 111619.
Lee, Z., Weidemann, A., Arnone, R., 2012. Combined effect of reduced band number and increased bandwidth on shallow water remote sensing: the case of WorldView 2. IEEE Trans. Geosci. Rem. Sens. 51 (5), 2577–2586.
Legleiter, C.J., Kinzel, P.J., Overstreet, B.T., 2011. Evaluating the potential for remote bathymetric mapping of a turbid, sand-bed river: 1. Field spectroscopy and radiative transfer modeling. Water Resour. Res. 47 (9).
Li, Y., et al., 2022. Diversity and phosphate solubilizing characteristics of cultivable organophosphorus-mineralizing bacteria in the sediments of sancha lake. Int. J. Environ. Res. Publ. Health 19 (4), 2320.
Li, N., et al., 2023a. Satellite-derived bathymetry integrating spatial and spectral information of multispectral images. Appl. Opt. 62 (8), 2017–2029.
Li, Y., et al., 2023b. The spatio-temporal distribution of alkaline phosphatase activity and phoD gene abundance and diversity in sediment of Sancha Lake. Sci. Rep. 13 (1), 3121.
Li, J., et al., 2025. Multi-temporal image fusion-based shallow-water bathymetry inversion method using active and passive satellite remote sensing data. Remote Sens. 17 (2), 265.
Liu, Y., et al., 2021. A downscaled bathymetric mapping approach combining multitemporal Landsat-8 and high spatial resolution imagery: demonstrations from clear to turbid waters. ISPRS J. Photogrammetry Remote Sens. 180, 65–81.
Liu, Y., et al., 2025a. A virtual coastal band-driven optimization-based bathymetric approach (VOBA) for optically shallow water with multispectral imagery. GIScience Remote Sens. 62 (1), 2506191.
Liu, Y., et al., 2025b. A virtual coastal band-driven optimization-based bathymetric approach (VOBA) for optically shallow water with multispectral imagery. GIScience Remote Sens. 62 (1), 2506191.
Lv, Z., et al., 2025. BathyFormer: a transformer-based deep learning method to map nearshore bathymetry with high-resolution multispectral satellite imagery. Remote Sens. 17 (7), 1195.

Lyzenga, D.R., 1978. Passive remote sensing techniques for mapping water depth and bottom features. Appl. Opt. 17 (3), 379–383.
Lyzenga, D.R., 1985. Shallow-water bathymetry using combined lidar and passive multispectral scanner data. Int. J. Rem. Sens. 6 (1), 115–125.
Ma, Y., et al., 2020. Satellite-derived bathymetry using the ICESat-2 lidar and Sentinel-2 imagery datasets. Rem. Sens. Environ. 250, 112047.
Main-Knorn, M., et al., 2017. Sen2Cor for sentinel-2. In: Image and Signal Processing for Remote Sensing XXIII. SPIE.
McFeeters, S.K., 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Rem. Sens. 17 (7), 1425–1432.
Misra, A., et al., 2018. Shallow water bathymetry mapping using support vector machine (SVM) technique and multispectral imagery. Int. J. Rem. Sens. 39 (13), 4431–4450.
Reichstein, M., et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204.
Sagawa, T., et al., 2019. Satellite derived bathymetry using machine learning and multi-temporal satellite images. Remote Sens. 11 (10), 1155.
Stumpf, R.P., Holderied, K., Sinclair, M., 2003. Determination of water depth with high-resolution satellite imagery over variable bottom types. Limnol. Oceanogr. 48 (1part2), 547–556.
Su, K., et al., 2023. Application and comparison of four assessment methods for water quality of Sancha Lake in central Sichuan Province, China. Water Pract. Technol. 18 (11), 2797–2808.
Traganos, D., et al., 2018. Estimating satellite-derived bathymetry (SDB) with the google Earth engine and Sentinel-2. Remote Sens. 10 (6), 859.
Vanhellemont, Q., 2019. Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the landsat and Sentinel-2 archives. Rem. Sens. Environ. 225, 175–192.
Vanhellemont, Q., Ruddick, K., 2021. Atmospheric correction of Sentinel-3/OLCI data for mapping of suspended particulate matter and chlorophyll-a concentration in Belgian turbid coastal waters. Rem. Sens. Environ. 256, 112284.
Vinayaraj, P., Raghavan, V., Masumoto, S., 2016. Satellite-derived bathymetry using adaptive geographically weighted regression model. Mar. Geod. 39 (6), 458–478.
Wan, J., Ma, Y., 2021. Shallow water bathymetry mapping of Xinji Island based on multispectral satellite image using deep learning. J. Indian Soc. Remote Sens. 49 (9), 2019–2032.
Wu, Z., et al., 2024. Integration of geographic features and bathymetric inversion in the Yangtze River's Nantong channel using gradient boosting machine algorithm with ZY-1E satellite and multibeam data. Geomatica (Ott.) 76 (2), 100027.
Xi, X., Guo, G., Gu, J., 2025. Band weight-optimized BiGRU model for large-area bathymetry inversion using satellite images. J. Mar. Sci. Eng. 13 (2).