Geoscience and Remote Sensing (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/geors.2025.080104
ISSN 2523-2592 Vol. 8 Num. 1

Monte Carlo Deep Learning Model for Quantitative Inversion of Total Nitrogen Concentration in the Source Area of the Yellow River Using Google Earth Engine

Ruichun Chang[1,2,a,*], Chi Zhang[1,2], Jian Xu[1,2], Zhe Chen[1,2,3,4], Wanquan Tuo[5]

[1]School of Mathematical Sciences, Chengdu University of Technology, Chengdu, 610066, Sichuan, China
[2]Digital Hu Line Research Institute, Chengdu University of Technology, Chengdu, 610066, Sichuan, China
[3]Department of Land, Environment, Agriculture and Forestry, University of Padova, Legnaro, PD 35020, Italy
[4]Aerospace Information Innovation Research Institute, Chinese Academy of Sciences, Beijing, 100089, China
[5]State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, CAS, Xi'an, 710061, China
[a]changruichun08@cdut.edu.cn
[*]Corresponding author

Keywords: Hyperspectral Remote Sensing; Monte Carlo Dropout Technique; Deep Learning; Total Nitrogen; Remote Sensing Quantitative Inversion

Abstract: As a vital ecological barrier in China, the Yellow River source area's water changes significantly impact the regional environment. Traditional remote sensing inversion methods face challenges like limited accuracy and complex data processing. This study uses Sentinel-2 remote sensing data and ground-based hyperspectral data, combined with an improved deep learning model (MC-DL), to establish an efficient framework for key water parameter inversion. Focusing on Ruoergai County, the MC-DL model, enhanced by Monte Carlo dropout, quantitatively inverts total nitrogen (TN) concentration. The MC-DL model outperforms Support Vector Regression (SVR) and Convolutional Neural Network (CNN) in accuracy and stability (R^2 = 0.95, MAE = 0.08, MBE = -0.004, RMSE = 0.13).This study provides a new technological approach for water monitoring in the Yellow River source area and supports ecological management and protection.

1. Introduction

The Yellow River source area, known as the "Water Tower of China," serves as a crucial ecological barrier for China's environmental protection. As the origin of the Yellow River, this region has faced immense ecological pressure due to global climate change and increasingly frequent human activities. In recent years, significant changes in the surface water quality and

quantity in the Yellow River source area have greatly impacted agricultural production, water security, environmental protection, and sustainable water resource development in the middle and lower reaches of the Yellow River basin[1]. Therefore, there is an urgent need for large-scale, rapid, and accurate monitoring of surface water in the Yellow River source area to ensure the continued functioning of its ecological role and the stability of the regional environment.

Water quality detection and monitoring have become key strategies to ensure the safety of the Yellow River source area. Traditional water quality detection methods rely on field sampling and laboratory analysis, which suffer from poor timeliness, high costs, and limited coverage. Hyperspectral remote sensing technology, by monitoring multiple spectral bands of surface water bodies, can obtain large-scale water quality data in real-time and quickly, thus compensating for the shortcomings of traditional methods [2].

Significant progress has been made in water quality inversion via remote sensing. Internationally, research focuses on large-scale water bodies using diverse remote sensing data and algorithms, achieving promising results. Domestically, studies emphasize regional issues like urban lakes and river pollution, utilizing high-resolution data and machine learning for refined monitoring and assessment.

In the field of parameter regression models, successfully applied the Extreme Gradient Boosting Tree model to retrieve TN concentration in the Liao River Basin ($R^2 \geq 0.575$), outperforming stepwise regression and random forest models[3].Employed the Partial Least Squares Regression model to retrieve TP and TN content, achieving significant results[4], highlighting the advantages of parameter regression models in water quality retrieval. In the field of empirical models, Used empirical models to retrieve TN concentration, achieving a maximum R^2 of 0.92 [5]. In the field of machine learning models, applied back-propagation neural networks (BP), Gaussian process regression (GPR), and random forest regression (RFR) to retrieve in-situ concentrations of TN, TP, and COD in Taihu Lake, Liangxi River, and Fuchuanjiang Reservoir, finding that the BP model performed the best, with accuracies exceeding 80% [6]. Achieved high-precision retrieval of water quality parameters from Landsat 8 OLI imagery data using a neural network model ($R^2 \geq 0.85$) [7].

With research advancements, deep learning's advantages in water quality retrieval are increasingly evident. ANN and LR models were used to analyze Landsat 8 OLI data for TP and TN in the Geshlagh Reservoir. Results showed a strong correlation between TN, TP, and Chl-a, with ANN achieving higher accuracy (TP: 0.81, TN: 0.93) than LR[8].

Traditional regression models achieve some success in water quality retrieval but struggle with nonlinear complexity and environmental variability. Deep learning offers superior nonlinear modeling and large-scale data processing but risks overfitting. This study enhances deep learning models using GEE and hyperspectral data to improve TN retrieval accuracy. Key challenges include integrating remote sensing with deep learning for TN distribution and optimizing models to address data scarcity and observation variability. These improvements enhance water quality monitoring and ecological management.

2. Materials

2.1 Study Area

The Yellow River source area, located in the eastern Roof of the World, spans 122,000 square kilometers (16.4% of the Yellow River basin) and contributes about 38% of the river's annual runoff. It is the basin's most critical water conservation and runoff replenishment area and a vital ecological barrier in China. Ruoergai County, situated in the southeastern Roof of the World, is one of the Yellow River's origins and a key ecological functional zone in its upper reaches.

The parameters used to detect water environmental conditions are diverse, covering various

indicators such as TN, TP, chemical oxygen demand, etc. However, existing water quality inversion models mostly focus on monitoring common parameters such as suspended solids, chlorophyll-a, etc., in static water bodies such as lakes and reservoirs[9,10].In the high-altitude, cold, and low-oxygen Yellow River source area, traditional techniques are limited. Despite environmental variations, TN, TP, COD, suspended solids, and chlorophyll-a inversion follow similar processes. TN, a key indicator of eutrophication and pollution, is crucial for water quality monitoring. This study focuses on TN to improve water resource management and ecological protection. TN variation is influenced by alpine meadows, climate change, and human activities like farming and land use changes. The research area is the confluence of the Yellow, Heihe, and Baihe Rivers in Ruoergai County.

2.2 Remote sensing data preprocessing

2.2.1 Remote sensing data sources

Sentinel-2, launched by the European Space Agency (ESA), comprises two high-resolution optical satellites with 13 spectral bands and a 290 km swath width. Each satellite has a 10-day revisit cycle, offering frequent observations with resolutions ranging from 10 to 60 meters. Its near-infrared bands are particularly sensitive to water bodies, making it highly effective for monitoring water quality changes, detecting pollution, and assessing aquatic ecosystem health.For this study, Sentinel-2 data synchronized with field sampling time (±5 days) from July 12 to July 24, 2022, was used for model construction and validation.

2.2.2 GEE cloud computing platform and remote sensing data source preprocessing

In recent years, with the continuous development of cloud computing technology and the open sharing of remote sensing data, cloud platforms such as Google Earth Engine (GEE) have become important tools for remote sensing data processing and analysis [11]. GEE provides access to massive amounts of remote sensing data, powerful computational capabilities, and rich analysis functions, offering researchers an efficient and convenient data processing platform. This has greatly facilitated the development and application of remote sensing water quality inversion techniques.

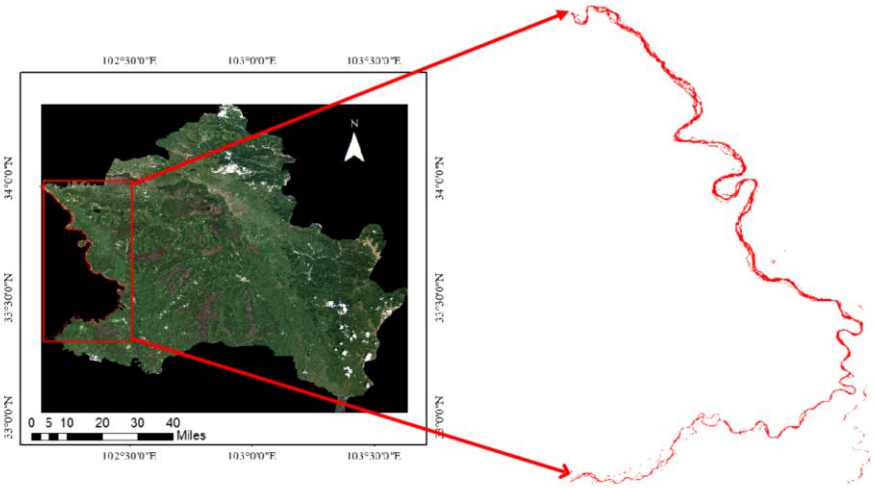


Fig.1 Water extraction area in the study area

In GEE, Sentinel-2 preprocessing includes atmospheric correction, radiometric calibration, and geometric correction to ensure data quality and model accuracy. Atmospheric correction removes

aerosol and water vapor effects using surface reflectance products; radiometric calibration converts raw values into reflectance via spectral coefficients; and geometric correction aligns spatial positions. Water body screening is performed using indices like NDWI, MNDWI, and NDSI, classifying pixels as water or non-water based on thresholds[12]. Water extraction areas are generated from index classification results, and NDWI is used for road extraction. This process is implemented through GEE's image calculation and classification functions, producing a final water extraction layer with accurately delineated water bodies, as shown in Fig. 1.

2.2.3 Develop a route for collecting hyperspectral data of field objects

This study employed a field route planning method based on Geographic Information System (GIS) technology to effectively carry out field hyperspectral data collection tasks. By integrating water body and road extraction results from the GEE platform, terrain data was analyzed in GIS for factors like elevation, slope, and aspect. Considering water sources, vegetation, terrain, transportation, and safety, the optimal field route was determined to ensure practicality and feasibility.

2.3 Hyperspectral remote sensing data sampling and processing

In this experiment, spectral water samples were collected using the American ASD FieldSpec spectrometer , with a detection spectral range of 350-2500 nm. The spectral resolution was 3 nm at 700 nm and 8 nm from 1400-2100 nm, with wavelength accuracy of 0.5 nm and wavelength repeatability of 0.1 nm. A total of 56 water samples were collected (Fig. 2) and analyzed using the ASD spectrometer, followed by laboratory analysis using the ultraviolet spectrophotometry method with alkaline potassium persulfate digestion, as per the HJ 636-2012 standard for TN in water quality.

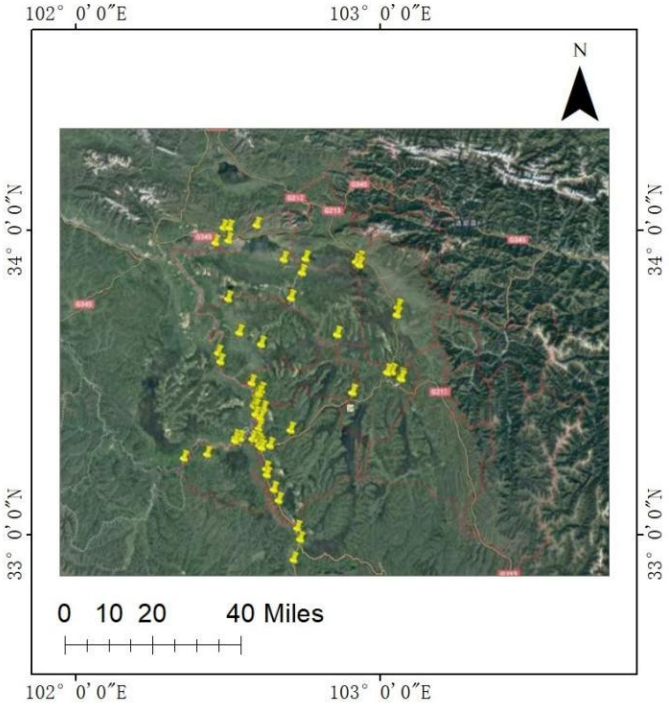


Fig.2 Sampling point distribution

3. Proposed model

3.1 Random Enhancement Sampling

Random augmentation sampling introduces randomness into datasets to enhance model generalization, addressing issues like data imbalance, overfitting, and limited generalization. In complex real-world scenarios, original data may lack diversity and quantity for effective model learning. This method improves robustness by applying random transformations (e.g., rotation, cropping) to training samples, increasing dataset diversity. The principle is represented as:

$$X^{'} = T(X, \theta) \tag{1}$$

where, X is the original data, $X^{'}$ is the data after random transformation, T is the random enhancement transformation, and θ is the parameter of the transformation.

3.2 MC-DL inversion model construction

Deep learning is a branch of machine learning that aims to mimic the structure of the human brain's neural networks to achieve the ability to learn from data and extract complex patterns.

Elastic Net is a regularization method that combines Lasso Regression and Ridge Regression by adding both L_1 and L_2 penalty terms. This method is widely used in feature selection and regression problems, enhancing the model's generalization ability and robustness. Its formal representation is as follows:

$$\min_{\omega} \frac{1}{2n_{samples}} \left\| X\omega - y \right\|_2^2 + \lambda\rho \left\| \omega \right\|_1 + \frac{\lambda(1-\rho)}{2} \left\| \omega \right\|_2^2 \tag{2}$$

where, ω is the weight vector to be solved, X is the input eigenmatrix, y is the target variable vector, $n_{samples}$ is the sample number, $\| \|_1$ represents the L_1 norm (sum of absolute values), $\| \|_2$ represents the L_2 norm (Euclidean distance), λ is the regularization coefficient, used to balance the influence of fitting error and regularization terms, ρ is the mixing ratio, when $\rho = 1$, The model is equivalent to Lasso regression; When $\rho = 0$, the model is equivalent to Ridge regression. By adjusting the value of ρ, you can find a balance between L_1 and L_2 to better accommodate various data characteristics.

Monte Carlo (MC) dropout technique provides a scalable approach for predicting distributions in deep learning [13]. The working principle of MC dropout involves randomly dropping neurons in the neural network to regularize the network. Each dropout configuration corresponds to a sample from a different approximate parameterized posterior distribution:

$$q(\Theta|D) \tag{3}$$

$$\Theta_t \sim q(\Theta|D) \tag{4}$$

where, Θ_t corresponds to the Dropout configuration sampled from the approximate parameterized posteriori, or equivalently corresponds to the simulation $q(\Theta|D)$, and sampling q from the approximate posteriori $q(\Theta|D)$ allows Monte Carlo integration of the model's likelihood to reveal the predicted distribution, as shown below:

$$p(x|y) \approx \int_{\Omega} p(y|x,\Theta)q(\Theta|D)d\Theta$$

$$\overset{MC}{\approx} \frac{1}{T}\sum_{t=1}^{T} p(y|x,\Theta_t), \quad \text{s.t.} \quad \Theta_t \sim q(\Theta|D) \tag{5}$$

For simplicity, we can assume that the likelihood is Gaussian:

$$P(y \mid x,\Theta) = \mathrm{N}(f(x,\Theta), s^2(x,\Theta)) \tag{6}$$

Using the Gaussian function N specified by the mean $f(x,\Theta)$ and the variance $s^2(x,\Theta)$ parameters output by the Monte Carlo dropout BNN simulation:

$$f(x,\Theta), s^2(x,\Theta) \sim MonteCarloDropout(x) \tag{7}$$

Fig.3 shows the case of MCdropout. Each dropout configuration generates different outputs by randomly turning neurons off (gray circles) and on (colored circles) each time they propagate forward. Multiple forward passes with different dropout configurations produce a predicted distribution of the mean $p(f(x,\Theta))$.

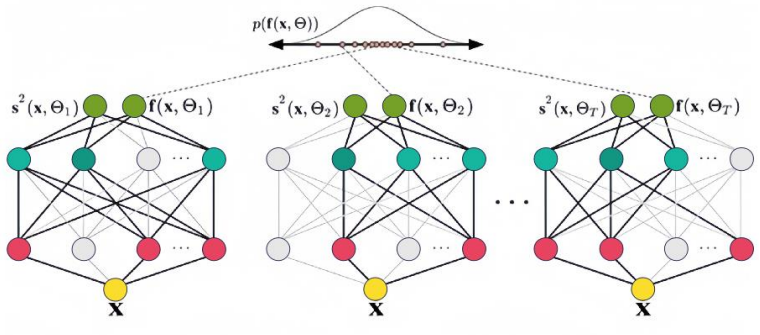


Fig.3 MC dropout schematic

This study builds the MC-DL model using TensorFlow and Keras frameworks (Fig. 4), training and testing it on the Colab platform. The model inputs spectral reflectance from Sentinel-2's 13 bands and outputs predicted TN concentration. Through multiple experiments, the optimal number of hidden layers and neurons per layer is determined, selecting the parameter combination that maximizes accuracy on both training and testing sets, as shown in Table 1.

Table 1 MC-DL model parameters

Hyperparameters	Parameter Value
Number of fully connected hidden layers	3
Number of neurons in each hidden layer	128,64,32
MCDropout layer	0.1
Hidden layer activation function	ReLU
Regularization coefficient L_2	0.001
Optimizer	Adam
loss	MSE
epochs	900
batch size	20
validation_split	0.2

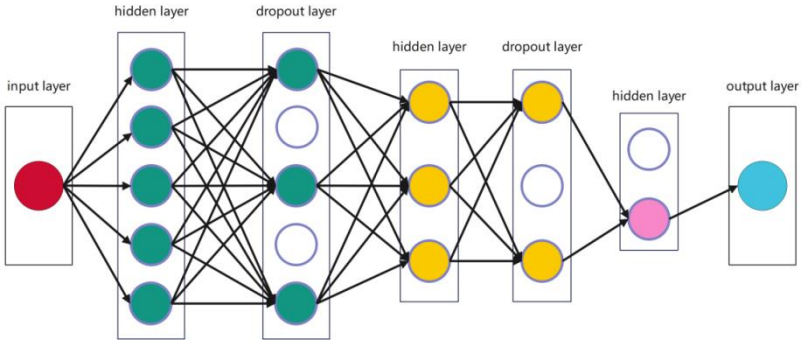


Fig.4 MC-DL model principle

3.3 Evaluation indicators

In this paper, coefficient of determination (R^2), mean absolute error (MAE), mean deviation error (MSE) and root mean square error (RMSE) are used to measure the fitting degree of the model to the true value and the accuracy of the prediction. The calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

where n is the number of observations in the data set, y_i is the true value, $\hat{y}_i$ is the predicted value of the model, and $\bar{y}$ is the average of the true values.

4. Result

4.1 Model accuracy evaluation results

This study trains and tests three models (SVR, CNN, and deep learning) using the same dataset, with 70% of sample points for training and 30% for testing. Model performance is evaluated using R ², MAE, MBE, and RMSE (Table 2). A sensitivity analysis assesses the impact of input uncertainties on prediction accuracy by perturbing key parameters and observing output changes.

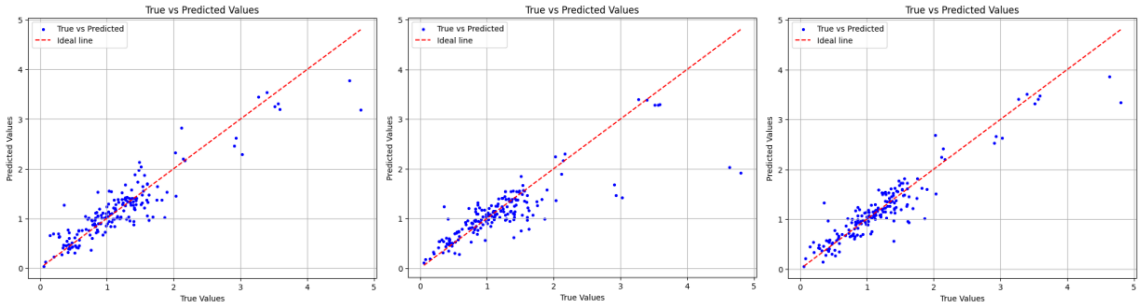
The SVR model performs well in training (R ²= 0.88) but drops in testing (R ²= 0.78), indicating overfitting and sensitivity to outliers (MAE, RMSE, MBE: 0.02 training, 0.06 test). The CNN model captures key features effectively (R ² 0.93 training, 0.85 test), with low errors and near-zero MBE, ensuring consistency. The MC-DL model, using Dense layers, excels across metrics, surpassing SVR and slightly trailing CNN in testing (R ²= 0.95 training, 0.89 test; MAE = 0.08 training, 0.18 test). Its RMSE (0.24) and MBE (0.02) suggest strong generalization but require further validation.

Table 2 Training and validation statistics of TN concentration based on hyperspectral Remote sensing(R^2, MAE, MBE, RMSE)

		SVR	CNN	MC-DL
Training set	R^2	0.88	0.93	0.95
	MAE	0.12	0.08	0.08
	MBE	0.02	-0.03	-0.004
	RMSE	0.21	0.16	0.13
Validation set	R^2	0.78	0.85	0.89
	MAE	0.21	0.20	0.18
	MBE	0.06	-0.01	0.02
	RMSE	0.34	0.29	0.24

4.2 Model performance evaluation results

Fig. 5 compares the three models' TN prediction performance using scatter plots and linear fitting. The x-axis represents true values, and the y-axis shows predictions. Each blue point is a sample, ideally aligning with the red dashed diagonal (perfect agreement). The closer points are to the regression line (slope = 1) and the denser their distribution, the more accurate the model.



(a) (b) (c)

Fig.5 Machine learning and deep learning algorithms are used to evaluate the performance of TN retrieval, where (a), (b) and (c) are the verification results of SVR, CNN and MC-DL algorithms, respectively

4.3 Model inversion result

This study focuses on six segments (a, b, c, d, e, f) extracted from the Yellow River source area in Zoige County. Through in-depth analysis of their data, the aim is to explore the differences and characteristics among these river segments(Fig.6).

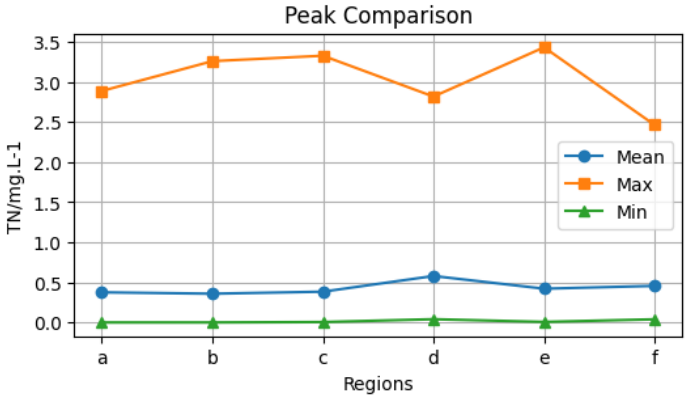


Fig.6 6 regional peak comparison statistical line chart

Fig.7 shows the local spatial distribution of TN concentration in the source region of the Yellow River obtained by model inversion. The figure clearly shows the difference of TN concentration in different regions, which helps us to better understand and analyze the water quality and its spatial variation characteristics in the source region of the Yellow River.

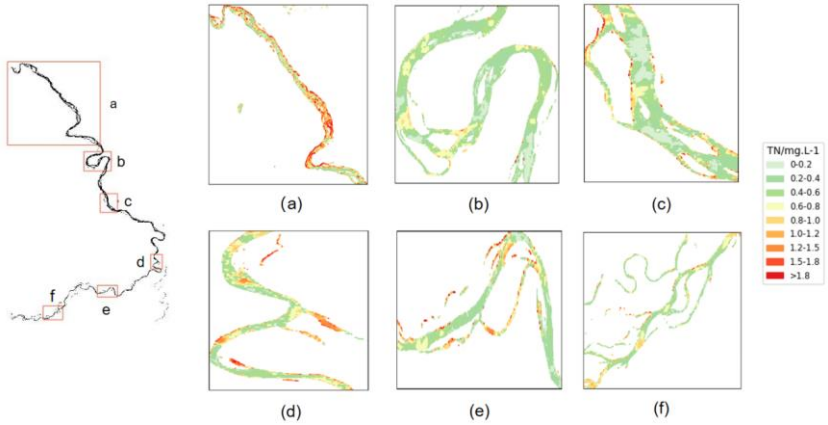


Fig.7 The local spatial distribution of TN concentration in the source region of the Yellow River was retrieved by the model

5. Discussion

Upstream (a) features plateau mountains with high precipitation and fast runoff, resulting in higher TN concentrations (0.3-0.4).Mid-upper reaches (b, c) are sparsely populated with limited farming, maintaining stable TN levels (0.2-0.4). Wetland vegetation and soil microbes reduce nitrogen content. Mid-lower reaches (d), at the confluence of the Yellow and Baihe Rivers, transition to plains with intensive agriculture and urbanization, leading to high TN (0.579) due to non-point source pollution and wastewater. Downstream (e, f) consist of the Yellow River plains, where fewer wetlands still effectively reduce TN to around 0.4.

6. Conclusions

This study developed a framework using Sentinel-2 and hyperspectral data to retrieve water quality parameters in Ruoergai County, Yellow River Source Region. The MC-DL model, incorporating Monte Carlo dropout, outperformed traditional methods (e.g., SVR, CNN) in predicting TN concentration. The MC-DL model improved accuracy and handled complex data, supporting water monitoring and ecological management. However, it relies on high-quality data, is sensitive to input parameters, and has limited generalization, requiring recalibration for different regions. Training demands high computational resources.

Future work should optimize the model for diverse environments and expand data collection to enhance performance. This study advances water quality monitoring and supports ecological protection.

Analysis in this study was completed using Google Colaboratory (https://colab.research.google.com/). The analyzed remote sensing date can be downloaded free of charge from Google Earth Engine (https://earthengine.google.com/).

CRediT authorship contribution statement

Ruichun Chang: Data curation, Investigation, Writing – original draft
Chi Zhang: Concept presentation, Methodology, Investigation, Writing – review & editing

Jian Xu: Software, Writing – review & editing
Zhe Chen: Methodology, Writing – review & editing
Wanquan Tuo: Supervision

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.

Data availability

Data will be made available on request.

Acknowledgments

This study was financially supported by State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, CAS (SKLLOG2314).We would like to thank all the partners for their participation in field sample collection and experimental analysis. The authors are also grateful to the processing editors and anonymous reviewers for their patient communication and enlightening suggestions.

References

[1] Zhang, T., Liang, S., Zhao, G., et al., 2023. Evolution of Pattern and Service Functions of Ecosystem in the Source Region of the Yellow River. People's Yellow River. 45 (9), 70-76. https://doi.org/10.3969/j.issn.1000-1379.2023.09.012.
[2] Yang, L., Yang, M., Yang, Y., 2023. Simulation and Application Progress of Water Environment Parameters Based on Multi source Remote Sensing. Leather Manuf. Environ. Prot. Technol. 4 (21), 81-83. https://doi.org/10.20025/j.cnki.CN10-1679.2023-21-28.
[3] Wang, W., Li, Y., Lei, K., et al., 2022. Remote Sensing Retrieval of Total Nitrogen Concentration in the Mainstream and Part of Tributaries in the Liaohe Watershed. Chin. Rural Water Hydropower. 7, 32-40.
[4] Chen, J., Zhang, L., Zhang, H., et al., 2023. Comparative study on the hyperspectral estimation models of TP and TN in Baiyangdian water body. Nat. Remote Sens. Bull. 27 (7), 1642-1652. https://doi.org/10.11834/jrs.20210575.
[5] Yang, H., Kong, J., Hu, H., et al., 2022. A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges. Remote Sens. 14 (8), 1770. https://doi.org/10.3390/rs14081770.
[6] Zhang, Y., 2022. Monitoring water quality using proximal remote sensing technology. Sci. Total Environ. 803, 149805. https://doi.org/10.1016/j.scitotenv.2021.149805.
[7] Wu, H., Guo, Q., Zang, J., et al., 2021. Study on Water Quality Parameter Inversion based on Landsat 8 and Measured Data. Remote Sens. Technol. Appl. 36 (4), 898-907. https://doi.org/10.11873/j.issn.1004-0323.2021.4.0898.
[8] Vakili, T., Amanollahi, J., 2020. Determination of optically inactive water quality variables using Landsat 8 data: A case study in Geshlagh reservoir affected by agricultural land use. J. Clean. Prod. 247. https://doi.org/10.1016/j.jclepro.2019.119134.
[9] Liang, W., Wu, Y., Shi, Y., et al., 2024. Retrieval of water quality in the Taipu River based on UAV hyperspectral imagery. Bull. Surv. Mapp. 29-34. https://doi.org/10.13474/j.cnki.11-2246.2024.0406.
[10] Liu, X., Zhang, M., Xie, T., et al., 2024. Spatial-temporal changes of chlorophyll a and turbidity in Honghu Wetland based on multi-source data and machine learning. Resour. Environ. Yangtze Basin, 1-15 http://kns.cnki.net/kcms/detail/42.1320.X.20240429.1133.002.html.
[11] Gorelick, N., Hancher, M., Dixon, M., et al., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18-27. https://doi.org/10.1016/j.rse.2017.06.031.
[12] Li, Y., Sun, X., Guo, Y., et al., 2020. Remote Sensing Retrieval of Water Quality Parameters in Poyang Lake Based on the Gradient Boosting Decision Tree Algorithm. Spacecraft Recov. Remote Sens. 41 (6), 90-102. https://doi.org/10.3969/j.issn.1009-8518.2020.06.009.
[13] Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Proc. Int. Conf. Mach. Learn. 1050-1059. https://doi.org/10.48550/arXiv.1506.02142.