Canadian Journal of Remote Sensing
Journal canadien de télédétection

Measurement of suspended sediment concentration at the outlet of the Yellow River Canyon: Using sentinel images and machine learning

Journal:	*Canadian Journal of Remote Sensing*
Manuscript ID	CJRS-23-0004
Manuscript Type:	Research Article
Date Submitted by the Author:	02-Feb-2023
Complete List of Authors:	Lei, Xinyu; Henan University - Jinming Campus, College of Geography and Environmental Science Song, Genxin; Henan University - Jinming Campus, College of Geography and Environmental Science Zhai, Shiyan; Henan University - Jinming Campus, College of Geography and Environmental Science Dong, Yubo; Henan University - Jinming Campus, College of Geography and Environmental Science He, Kangjie; Henan University - Jinming Campus, College of Geography and Environmental Science
Keyword:	

Measurement of suspended sediment concentration at the outlet of the Yellow River Canyon: Using sentinel images and machine learning

Xinyu Lei [a], Genxin Song [a]*, Shiyan Zhai [ab]*, Yubo Dong[a] and Kangjie He[a]

[a] *College of Geography and Environmental Science, Henan University, Kaifeng 475004, China;* [b] *Key Laboratory of Geospatial Technology for the Middle and Lower Yellow River Regions, Ministry of Education, Henan University, Kaifeng 475004, China*

***** Genxin Song**,** E-mail: gxsong@henu.edu.cn

***** Shiyan Zhai**,** E-mail: 10130079@vip.henu.edu.cn

Measurement of suspended sediment concentration at the outlet of the Yellow River Canyon: Using sentinel images and machine learning

Abstract. Monitoring and measuring suspended sediment concentration (SSC) along river systems greatly affects the balance of the river channel. Previous studies mainly focused on low-concentration sediment rivers. The traditional situ monitoring techniques and low-resolution imagery cannot meet the vastly spatial and temporal coverage. The Yellow River is famous for its high sediment content in the world. Therefore, this research considers the outlet of the Yellow River Canyon as a case study. This study develops a method for quantifying SSC in rivers with high-concentration sediment by combing High-Resolution Sentinel-2 products, corresponding SSC data obtained from the Yellow River Conservancy Commission of the Ministry of Water Resources, and machine learning techniques. The results showed that: (1) The estimation model of SSC derived from the random forest model performs well, and the prediction accuracy is high. The coefficient of determination (R^2) equals 0.94, These satisfy the requirements of high-precision SSC estimation. (2) The red and near-infrared bands are vital to the accuracy of SSC prediction. (3) Seasonal differences and spatial variation in sediment in Longmen waters are evident. The study's findings demonstrate that the random forest regression model is superior to traditional modeling methods for predicting SSC in sediment-rich rivers.

Keywords: suspended sediment concentration, Yellow River canyon outlet, random forest, Sentinel images

Introduction

The suspended sediment concentration (SSC) is one of the most critical water quality parameters in rivers, and sediment transport is an essential hydrological process (Francke, López-Tarazón, & Schröder, 2008). The SSC in water affects a series of water's optical properties, such as transparency, turbidity, and color (H. Wang et al., 2017). In addition, the sediment content of the river greatly influences the balance of the river channel, and local sediment deposition causes problems such as river bed deformation (Y. Wang, 2020). The UNESCO-ISI Online Training Workshop on Sediment Transport

Measurement and Monitoring was held from July 5-9, 2021, representing a vital initiative of the International Sediment Initiative (ISI) of UNESCO for 2021. The meeting highlighted the importance of measuring and monitoring sediment transport and managing and mitigating its negative effects (WASER, 2021-07-21).

The traditional SSC survey method measures point-by-point sampling with the ship on site, which is time-consuming and costly, and it is difficult to monitor a large area of water for a long time (Salama, Radwan, & van der Velde, 2012). With the advancement of science and technology, remote sensing technology possesses characteristics such as strong macroscopic and comprehensiveness, high comprehensive benefit, a large amount of information, a rapid update cycle, and a wide range of application fields. Therefore, re-mote sensing is valuable for quantifying SSC in river systems (Gordon, Brown, Brown, Evans, & Smith, 1988). The current SSC inversion is divided into parametric and non-parametric models. The parametric model is a mathematical and statistical model in which the uncertainty between the water body information and the remote sensing information is assumed to be a specific functional relationship. The model parameters are estimated by converting them into a multivariate linear function (W., 2012). Liao et al. (Liao, Zhang, & P.Y.Deschamps, 2005) developed an index model from atmospheric-corrected SeaWiFS data to monitor the spatial distribution of sediment along the eastern coast. Zhang et al. (M. Zhang & Guo, 2018) constructed a single-band empirical statistical model and a band combination empirical statistical model in the Zhoushan offshore area using GF-1 data, analyzed the correlation coefficients between various models, and verified their accuracy. They indicated that the inversion accuracy is higher in the quadratic model based on the (Band 3/Band 2) combination. The parametric model is intuitive, easy to understand, and has many applications. However, it usually requires the sample population to meet specific distribution characteristics (such as normal distribution), and the wide-spread spatiotemporal autocorrelation of water parameters cannot meet the assumptions(W., 2012). Therefore, this assumption has theoretical flaws(Zhang H X, 2002), which lower the model's prediction accuracy and generalization performance. Non-parametric models make no assumptions about the overall distribution of samples; they are a statistical analysis method for directly analyzing samples. They are widely used, highly reliable, and favored by many scholars. Li (Z. Li, 2012) built a relationship between marine suspended matter and remote sensing images utilizing a support vector machine (SVM). Kyle T. Peterson et al. (K. Peterson, Sagan, Sidike, Cox, & Martinez, 2018) used an extreme learning machine (ELM) in a neural network and Landsat satellite images to invert SSC in the Missouri and Mississippi rivers. In contrast, some machine learning also has some problems(L. Zhang, Zhang, & Du, 2016). For example, the performance of machine learning methods is mainly dependent on a large number of training samples, which are usually difficult to obtain in real-world scenarios; the complex structure is crucial (Szegedy et al., 2015). In addition, deeper networks result in high computational costs and overfitting, and the issue of balancing network depth and computational efficiency must be addressed(Sagan et al., 2020).

The Random Forest model (RF) is a machine learning method based on classification trees (Breiman, 2001). It is flexible, robust, practical, and efficient and can be used for regression, clustering, classification, prediction, and other analyses. Since the model has obvious advantages in parameter optimization, variable sorting, and subsequent variable analysis and interpretation, it is ideally suited for simulating suspended sediment. Dehkordi et al.(Dehkordi, Ghasemi, & Zoej, 2021) estimated the SSC along the Missouri River using both SVR and RF models and found that the RF model performed better. Gu et al.(Gu, Zhang, & Qiao, 2020) proposed a new RF-based river turbidity measurement model with excellent results, demonstrating the potential of RF for river detection.

Researchers have investigated the relationship between remote sensing reflectance data at visible and near-infrared wavelengths in the past few decades with SSC utilizing large rivers as the study area(Park & Latrubesse, 2015; Ritchie, Schiebe, & Mchenry, 1976). The current research still has some limitations. First, the sediment concentrations in large rivers are generally low and stable. For example, Yangtze (Fang et al., 2019), the Missouri River(Umar, Rhoads, & Greenberg, 2018), and the Mississippi River(K. Peterson et al., 2018) are current research hotspots. These rivers have an average sediment concentration between 83 and 395 mg/l. Second, many studies used low-resolution remote sensing im-age data to detect the long-time variation trend of river sediment (e.g., MODIS with a spatial resolution of 250–1000 m)(Espinoza Villar et al., 2013; Kilham & Roberts, 2011; Mangiarotti et al., 2013; K. Peterson et al., 2018). However, when the river channel in the study area is narrow, the resolution of these images cannot adequately meet the research requirements.

The Yellow River, the mother river of the Chinese nation, is famous for its high sediment concentration. Its average sediment concentration is 1.44 × 104 mg/l, and the fluctuation is huge, with a peak value of 7 × 104 mg/l (Data source: http://www.yrcc.gov.cn/zwzc/gzgb/gb/nsgb/)(YRCC, 2022-11-18). The problem of "hanging river on the earth's surface" in the downstream areas is becoming increasingly severe, and dam failure poses a serious risk to human life. The SSC of the Yellow River has be-come a vital factor affecting the ecological environment of the Yellow River Basin. The Yellow River sediment control is an urgent task. The Qin Jin Canyon of the Yellow River is located on the Loess Plateau, and a large amount of sediment is transported downstream yearly, which is the primary source of the Yellow River sediment(Ning, Gao,

& Fu, 2022; P. Zhang, Cai, Zheng, & He, 2020). Hence, monitoring the concentration of suspended particulate matter at the outlet of the Yellow River Canyon is valuable for understanding the specific situation of Yellow River sediment, which is of great scientific interest. For the SSC study of the Yellow River, most current investigations focus on the estuary(C. Li, Yu, Gong, Yang, & Cao, 2020; Jin Li et al., 2021; Qiu et al., 2017; S. Wang et al., 2019), while the middle reaches of the Yellow River, especially the mountainous areas, are seldom. This is detrimental to the ecological protection of the Yellow River Basin. Previously, this type of re-search was hindered by the limitation of remote sensing image resolution. The higher resolution, shorter return period, and multiple bands Sentinel-2 satellite image can effectively address these issues(Batista, 2022; Tian, Li, Zhu, Xue, & Zhao, 2022; Tripathi, Pandey, & Parida, 2020).

The outlet of the Yellow River Canyon is selected as the study area to investigate the relationship between SSC and remote sensing reflectance in high-concentration sediment rivers. This study focuses on the SSC inversion and spatial distribution at the Yellow River Canyon outlet using the Sentinel-2 satellite image, data on the daily average sediment con-tent of the Yellow River, and a random forest model. The specific research objectives of this research are as follows:

(1) Building a non-parametric inversion model, dividing the sample point into sever-al types by attributes (season, hydrological situation), and finding the optimal inversion model through model checking.

(2) Applying the optimal inversion model, performing remote sensing estimation of SSC distribution in the study area, and analyzing the concentration of suspended sediment and its temporal and spatial variation in the water body in this region.

It provides essential information for managing and utilizing water resources in the middle reaches of the Yellow River and other inland rivers. It can also fill the gap of high-concentration sediment river inversion and introduce new concepts for remote sensing monitoring of the river environment.

Materials and Methods

Case study

The outlet of the Qin Jin Canyon of the Yellow River is YuMen Kou (110 ° 60′ E, 35 ° 65′ N), also known as Longmen. It is situated at the intersection of the Hejin City of Shanxi Province and the Hancheng City of Shaanxi Province. This region is a temperate continental climate (Fu, Zhang, Zheng, Yinghui, & Gao, 2016). More than 60% of the annual precipitation falls each year between June and September (Ouyang, Wang, Tian, & Tian, 2016). Upstream of YuMen kou is Qinjin Canyon, located on the Loess Plateau. The soil is loose, the terrain is broken, the vegetation coverage is poor, the rainstorm is concentrated, and the rainfall is large, providing favorable conditions for transporting a large amount of sediment. There is the primary source of the Yellow River sediment (X. Li, Jin, & Xu, 2012). The downstream of YuMen kou is Xiaobei Main Stream (from Yumenkou to Tongguan), which is a typical wandering river. In addition, the river bed of this reach is seriously silt-ed and uplifted, making it one of the most difficult reaches of the middle reaches of the Yellow River to control (Jie Li, Xia, & Zhu, 2022; X. Lin, Dong, Surin, & Hu, 2019; Xu, 2009).

1500 meters upstream of YuMen Kou is Longmen Hydrometric Station (110° 35′ E, 35° 40′ N). The test reach is approximately 400 meters in length and 130 meters in width. The bend and the checkpoint control the low-water channel and high-water section, as shown in Figure 1. Based on the data measured at the stations, the annual average runoff and sediment are 2.59 × 109 m3 and 6.40 × 108 t (Statistics 1987-2020), respectively. The monitoring data of Longmen Station are used as the basis for inversion modeling and model verification.

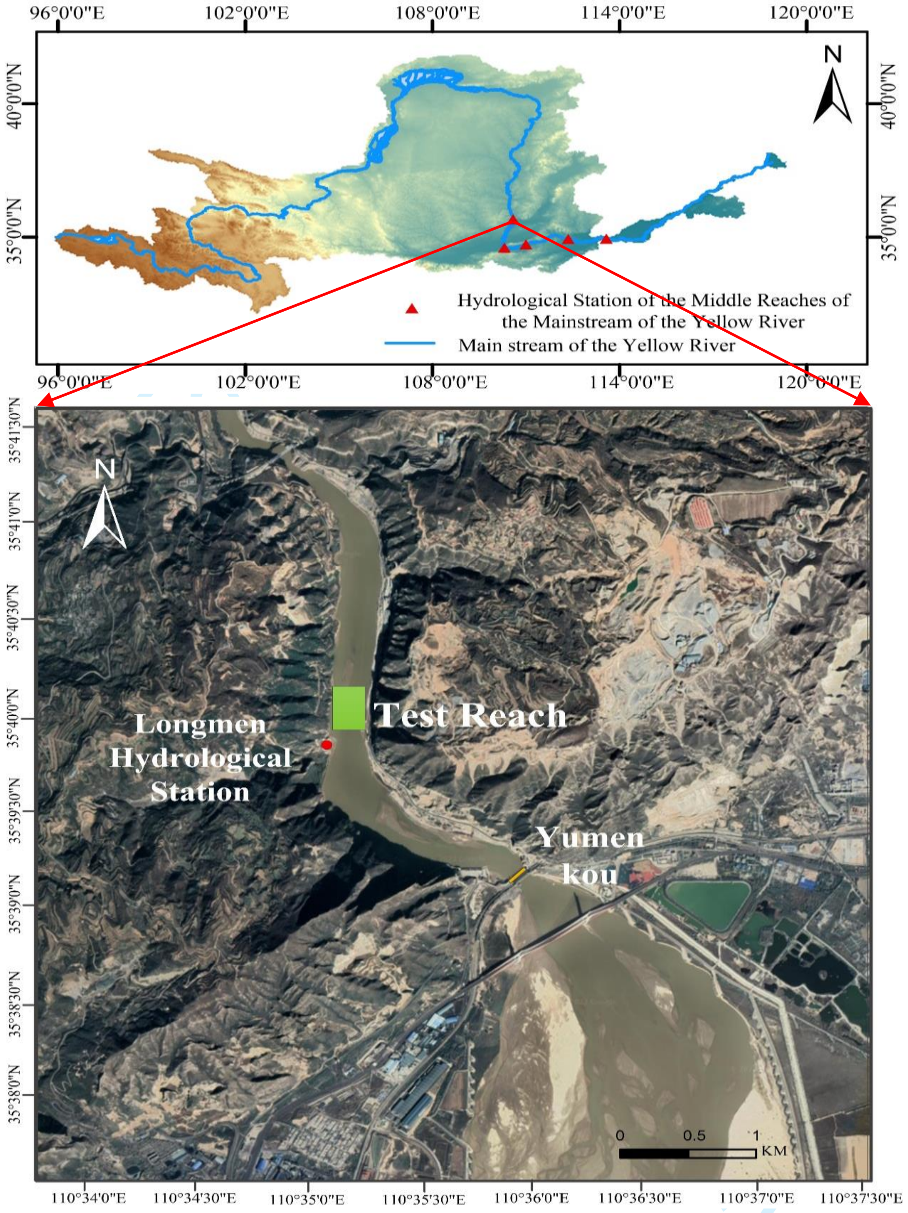


Figure 1. The geographical locations of the Yellow River Canyon and the Sampling reach.

Sentinel-2 MSI data

The Sentinel-2 satellite is part of the Copernicus mission of the European Space Agency, which uses a constellation of satellites to observe Earth. Launched in June 2015 and March 2017, both satellites are equipped with MSI (Multispectral Imaging

System), covering 13 spectral bands and 290 km wide. The ground resolution is 10, 20, and 60 meters (Table 1.). The revisit period of one satellite is six days, and the two complement each other, and the revisit period is three days. It has various spatial resolutions, from visible and near-infrared to short-wave infrared.

Table 1 Spectral information for Sentinel-2 MSI.

Payload Band	Sentinel-2A		Sentinel- 2B		Pixel Size(m)
	central wavelength (nm)	**Spectral width (nm, half height)**	**central wavelength (nm)**	**Spectral width (nm, half height)**	
B1：Aerosols	442.7	21	442.2	21	60
B2：Blue	492.4	66	492.1	66	10
B3：Green	559.8	36	559.0	36	10
B4：Red	664.6	31	664.9	31	10
B5：Red Edge 1	704.1	15	703.8	16	20
B6：Red Edge 2	740.5	15	739.1	15	20
B7：Red Edge 3	782.8	20	779.7	20	20
B8：NIR	832.8	106	832.9	106	10
B8A：Red Edge 4	864.7	21	864.0	22	20
B9：Red Edge 4	945.1	20	943.2	21	60
B11：SWIR 1	1613.7	91	1610.4	94	20
B12：SWIR 2	2202.4	175	2185.7	185	20

Search through GEE (Google Earth Engine) for all Sentinel-2 MSI Level 2A products in 2019-2020. The product is the Bottom-of-Atmosphere created reflectance data that has been corrected for ortho, geometrical, and atmospheric correction. The following processing on the 295 images queried is performed:

(1) Image filtering. Through visual interpretation and the band (QA60) of cloud mask information included with Sentinel-2 data, images in the study area that cannot be collected due to cloudy weather are filtered out. Neighboring pixels are reselected as sampling points for images exposed to land due to a decreased water volume. Finally, 128 valid images were obtained.

(2) Band reflectance information extraction. Due to the weak reflection signal of water bodies, the edge region of the water bodies is easily affected by the reflection of land pixels, making it difficult to represent the actual water surface. In order to

eliminate the effect of land pixel reflection, considering the resolution of the Sentinel image and the width of the test river, the median filter template of 5 × 5 is selected (Pahlevan, Sarkar, Franz, Balasubramanian, & He, 2017). Finally, the center of the test river is chosen as the sampling point, and the image reflectance after median filtering is extracted as the final reflectance.

Sediment data

Longmen Hydrometric Station was established on the left bank outside YuMen kou on June 14, 1934 (Longmen I). After several changes in September 1971, it moved to the current site (Mawang Miao II). In 1974, the measurement method was changed from a gondola ropeway to a heavy lead fish electric ropeway and double-cable fixed hanging box, and it remains in use today. The daily average sediment concentration data at Longmen station from 2019 to 2020 was obtained through the Yellow River Conservancy Commission of the Ministry of Water Resources.

Due to the excessive sediment concentration in the Yellow River, the inversion effect is not good by directly using the original sediment concentration data. In order to improve the accuracy of the model and ensure that the regression parameter estimator has good statistical properties, the natural logarithm of sediment concentration (expressed in terms of ln (SSC)) is applied to fit the reflectivity of each band(Smith & Croke, 2005).

Random Forest Model

Modeling Process

After a series of experiments, it found that the band combination with a better inversion effect after linear fitting has a better inversion impact in random forest model regression, and the overall accuracy is higher than that of correlation, index fitting and other indicators. Fit the spectral reflectance of each band and ln (SSC) with a linear model, and select the band combination with a better fitting effect to participate in constructing the random forest model. Select 70% of the sample data from the sediment concentration data set as the training sample set for the random forest regression model, and use the remaining 30% as the test sample set for the model.

The importance of each variable obtained by the random forest algorithm is sorted. Variable importance refers to the contribution rate of the predictor variable to the prediction accuracy. The larger the value, the more important the variable is. Variable importance in the random forest model is relative and sums to 1. After the variables are sorted by importance, they must be screened to ensure that the model has fewer predictors and more accurate prediction effects. This not only simplifies the model, but also facilitates the interpretation of subsequent models(Ismail, Mutanga, & Kumar, 2010). The variable selection adopts the backward elimination method, i.e., after sorting based on the importance, all variables are used as prediction parameters to construct a random forest algorithm. The model is then applied to estimate the test sample set, and its prediction accuracy is recorded. Finally, each variable is reduced individually, and the above process is repeated. The prediction accuracy of each model under different variable combinations is compared, and the combination with the highest accuracy is selected as the optimal variable group. The random forest algorithm model is then reconstructed to test the model's prediction accuracy.

Model checking

The evaluation indices of this paper are mainly R^2 (1) of determination coefficient, (2) MSE (mean-square error) and (3) MAPE (mean absolute percentage error) as auxiliary

(Bian et al., 2013), jointly assessing the accuracy of the model, comprehensively evaluating the prediction ability of all regression models, and testing the prediction results of regression models. MSE is the mean value of the sum of the squared errors of the predicted data and the original data; MAPE is the percentage used to express model errors, which is relatively more intuitive and unaffected by the range of values of the original data. It is more suitable for comparing different data (subsets).

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

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

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{\hat{y}_i}\right| \tag{3}$$

where n is the total number of samples, $\bar{y}$ is the average value of the natural logarithm of the concentration of sediment in the data set, y_i is the natural logarithm of the concentration of sediment in the i (i =1,2,3,⋯,n) sample point in the data set, and y_i is the natural logarithm of the concentration of sediment in the i (i =1,2,3,⋯,n) sample point predicted by the prediction model. The higher R^2, the lower MSE and MAPE, the better the model fitting effect and the higher the inversion precision.

Results

Suspended Sediment Modeling

According to(Wen, 2017; Yang, Fan, Zhang, Yu, & Zhu, 2019), this study combines the bands by division, subtraction, and the normalized difference index form to fit the ln (SSC). Each band is entered into the model for linear fitting using the exhaustive method. The combination bands with relatively high R^2 and low MSE will be selected. Table 2 displays the band combination results with a better fitting effect.

Table 2. Linear fit effect.

Modeling variables	MSE	MAPE	R^2	Formula
X=B8	0.99	11.05	0.453	Y=0.001269 X + 5.361
X=B8A	1.03	11.43	0.426	Y= 0.001112 X + 5.629
X=B5/B3	0.53	8.26	0.706	Y=4.937 X+ 1.196
X=B6/B1	0.53	8.60	0.704	Y= 1.808 X+ 4.242
X=B6/B2	0.50	8.28	0.722	Y= 2.24 X+ 4.193
X=B7/B1	0.52	8.39	0.712	Y= 1.672 X + 4.427
X=B7/B2	0.51	8.35	0.717	Y= 2.034 X + 4.422
X=B8/B1	0.53	8.58	0.706	Y= 1.724 X + 4.529
X=B8/B2	0.51	8.31	0.717	Y=2.118 X + 4.502
X=B8A/B1	0.51	8.51	0.716	Y=1.617 X + 4.897
X=B8-B1	0.52	8.65	0.715	Y= 0.002047 X + 6.239
X=B8-B2	0.50	8.59	0.722	Y= 0.002132 X + 6.53
X=B8-B3	0.51	8.63	0.718	Y= 0.002345 X + 7.356
X=B11-B6	0.52	8.26	0.713	Y= -0.002 X + 5.265
X=B11-B8A	0.49	8.24	0.732	Y=-0.001851 X + 5.677
X=B12-B6	0.51	8.26	0.721	Y=-0.001979 X + 5.159
X=B12-B7	0.50	8.29	0.722	Y= -0.001831 X + 5.26
X=B12-B8	0.49	8.20	0.729	Y= -0.001935 X + 5.351
X=B12-B8A	0.49	8.40	0.730	Y= -0.001809 X + 5.591
X=(B4-B3)/(B4+B3)	0.59	8.44	0.675	Y= 15.95 X + 5.958
X=(B5-B3)/(B5+B3)	0.58	8.66	0.681	Y= 11.39 X + 6.198
X=(B6-B2)/(B6+B2)	0.59	8.91	0.672	Y= 5.599 X + 6.631
X=(B7-B2)/(B7+B2)	0.62	9.10	0.658	Y= 5.087 X + 6.679
X=(B8-B2)/(B8+B2)	0.62	9.203	0.655	Y= 4.997 X + 6.871

Note: X represents the band or band combination, B is the abbreviation of the band, and Y represents ln (SSC).

Table 2 demonstrates that the fitting effect of a single band is subpar. For example, the R^2 of B8 (NIR) and B8A (Red Edge 4) is approximately 0.45. Combining the two bands can effectively improve the fitting accuracy. Table 2 indicates that B11-B8A and B12-B8 are the best inversion band combinations. The highest R^2 is 0.732, the minimum MAPE is 8.20%, and the lowest MSE is 0.49.

Sediment inversion

The sample points are divided into several groups by season or hydrological situation. First, this research conducted the linear fitting for each band individually. Then, the optimal band combination (between 25 and 30) will be selected and added to the random forest regression model.

Table 3. Results of random forest regression based on different classifications.

Basis for classification	Sample number	Training ln(SSC) R^2	Testing ln (SSC) R^2	Testing SSC R^2
Overall	124	0.956	0.873	0.956
High water period	45	0.965	0.683	0.881
Low water period	79	0.895	0.808	0.709
Spring	31	0.957	0.717	0.726
Summer	32	0.916	0.742	0.562
Autumn	30	0.924	0.754	0.917
Winter	31	0.841	0.551	0.803

Note: High water period is from June to October, and the rest of the time is Low water period.

Table 3 lists the results of the regression modeling. The random forest method proved to be quite robust, as expected. All testing SSC, excluding summer, have an R^2 greater than 0.7 and an average R^2 of 0.79. It shows that using the random forest model is feasible to monitor the suspended sediment concentration.

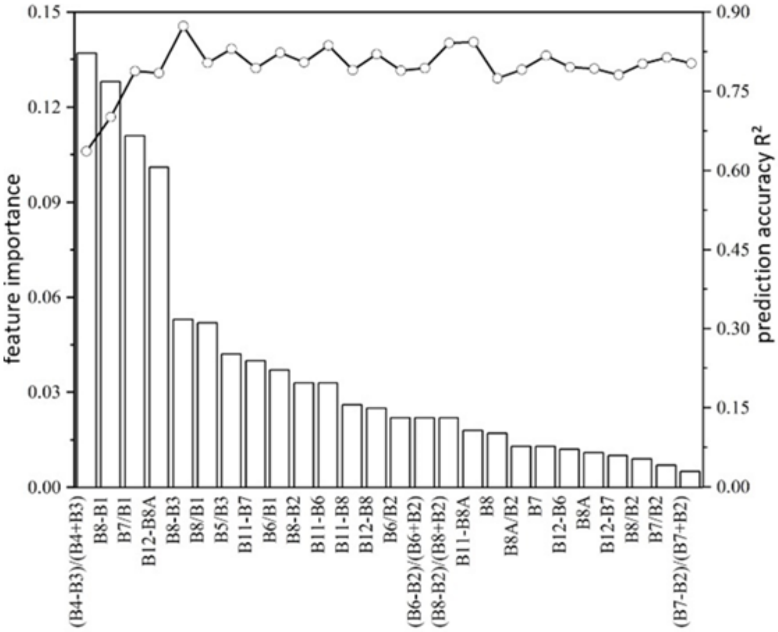
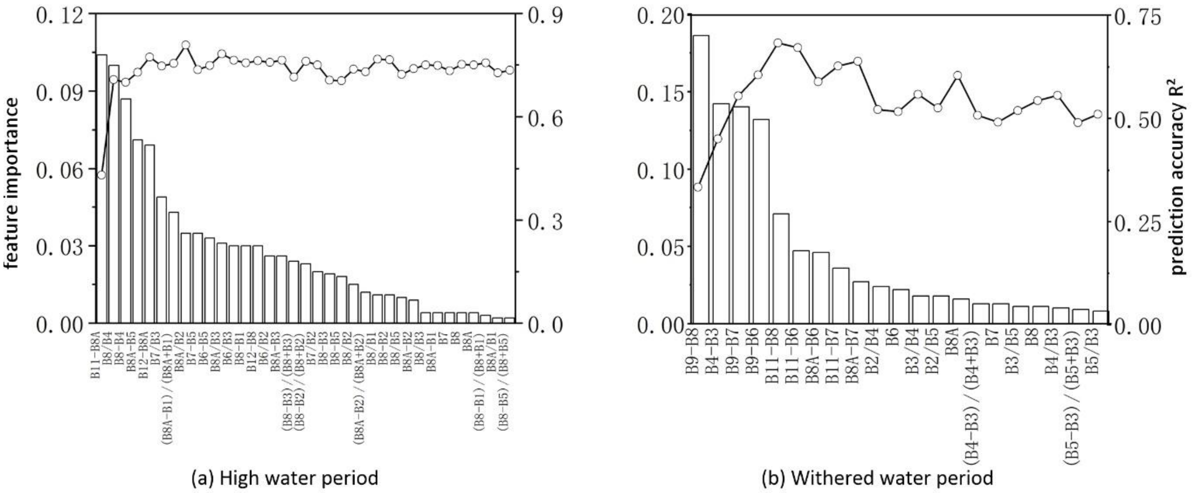


Figure 2. Random forest model SSC inversion result (overall, modeling sample number n=124). The bar graph represents the importance order of each predictor variable in the random forest regression model, and the line graph represents the fitting accuracy of the model after performing backward variable selection.

The random forest model determines the relative importance of each variable in the model by evaluating and ranking the contribution of each variable to the improvement of the model's prediction accuracy. The greater the variable's importance, the higher its contribution to the model's prediction accuracy.

Figure 2 depicts the overall inversion results of SSC. It indicates that as the number of variables changes, the prediction accuracy of each sub-model will gradually reach a relatively stable state. When all band combinations are added to the model, the obtained pre-diction accuracy is not the best. In contrast, the most accurate model prediction will result from choosing several variables. In addition, too many variables will lead to difficulty in interpretation. Therefore, when rebuilding the model, it is unnecessary to add too many variables to achieve maximum accuracy. When all samples participate in modeling, the band combination of (B4-B3)/(B4+B3), B8-B1, B7/B1, B12-B8A, B8-B3, has the best inversion effect, and R^2 is 0.873, where the band combination with the greatest feature importance is (B4-B3)/(B4+B3).



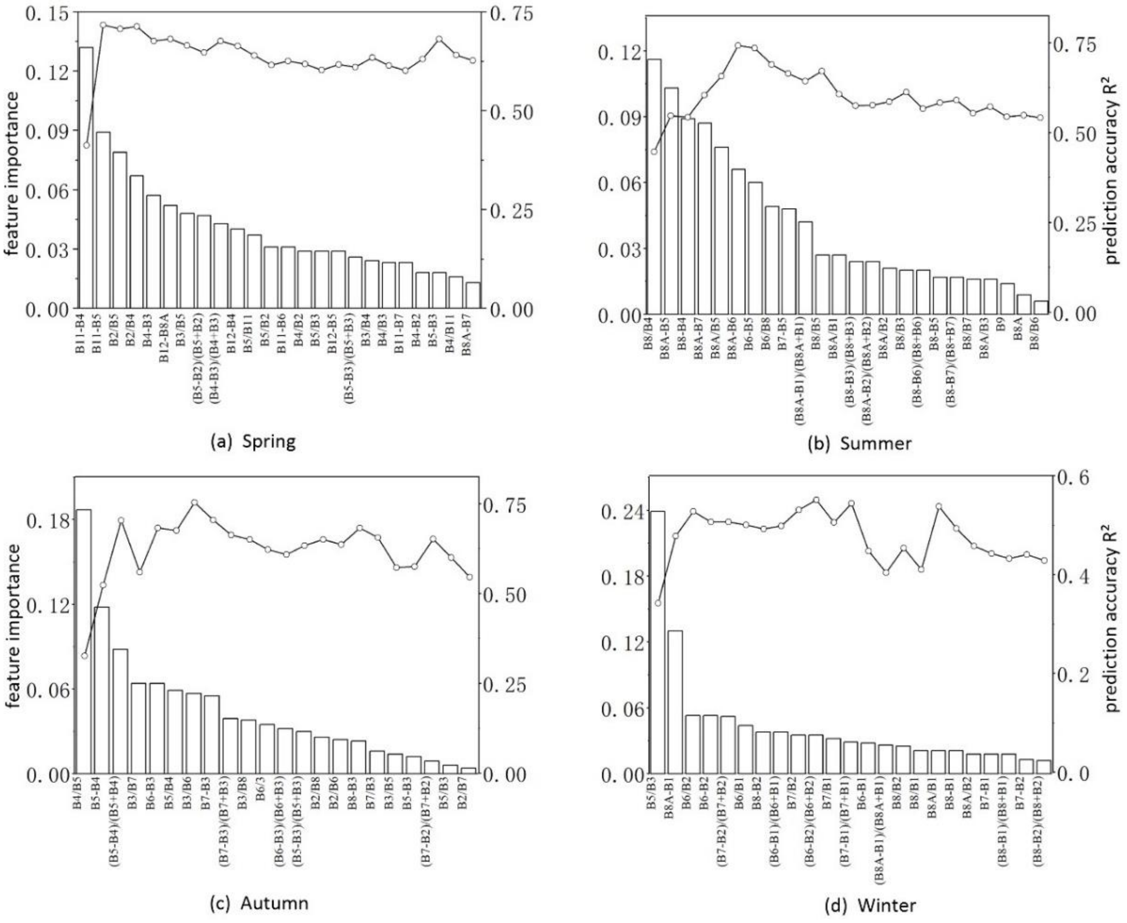
(a) High water period

(b) Withered water period

Figure 3. Random forest model SSC inversion results. (hydrological situation, n=79,45).

Figure 3 displays the inversion results of SSC by water regime classification. In the dry season, the band combination of B11-B8A, B8/B4, B8-B4, B8A-B5, B12-B8A, B7/B3, (B8A-B1)/(B8A+B1), B8A/B2 has the best inversion effect, R^2 is 0.808. The band combination with the greatest feature importance is B11-B8A. In the wet season, the band combi-nation of B9-B8, B4-B3, B9-B7, B9-B6 B11-B8 has the best inversion effect; R^2 is 0.683, and the band combination with the most important feature is B9-B8.

Figure 4 shows the inversion results of SSC by season classification. The inversion effect varies from season to season. The best inversion effect is in autumn, when the band combination is B4/B5, B5-B4, (B5-B4)/(B5+B4), B3/B7, B6-B3, B5/B4, B3/B6, R^2 is 0.754. The worst inversion effect is in winter, when its R^2 is the highest at 0.551, The band combination with the most important feature is B5/B3, the value of feature importance reaches 0.24. In summer, the inversion effect of 6-7 band combinations is the best, and the R^2 is above 0.73. The redundant band combinations will make the inversion effect worse. In spring, the inversion results are good when the second band combination is added.



(a) Spring

(b) Summer

(c) Autumn

(d) Winter

Figure 4. Random forest model SSC inversion results (season, n=31,32,30,31).

Comparison by classifying with different attributes, the optimal feature combinations obtained in Figures 2 through 4 are different. The Red band B4 (664 nm) and the Red Edge band B8A (864 nm) appear most frequently in the variable combinations when each model achieves the highest prediction accuracy, indicating that they are the most influential bands in the random forest prediction model.

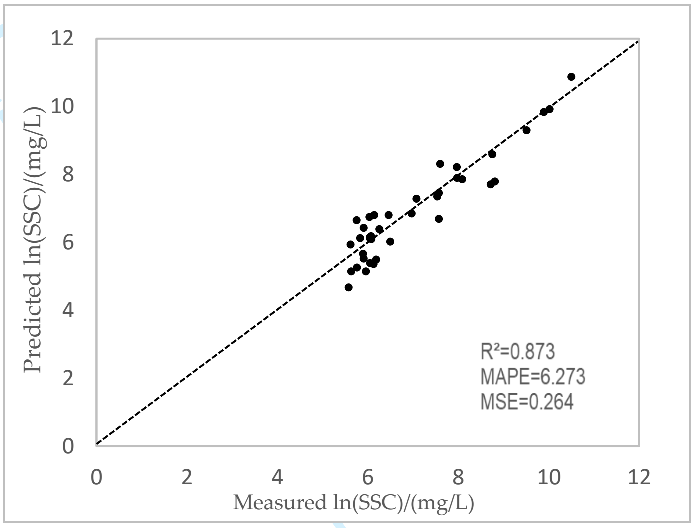


Figure 5. Overall estimation of suspended sediment in Longmen Station. The dotted line in the figure is the 1:1 line.

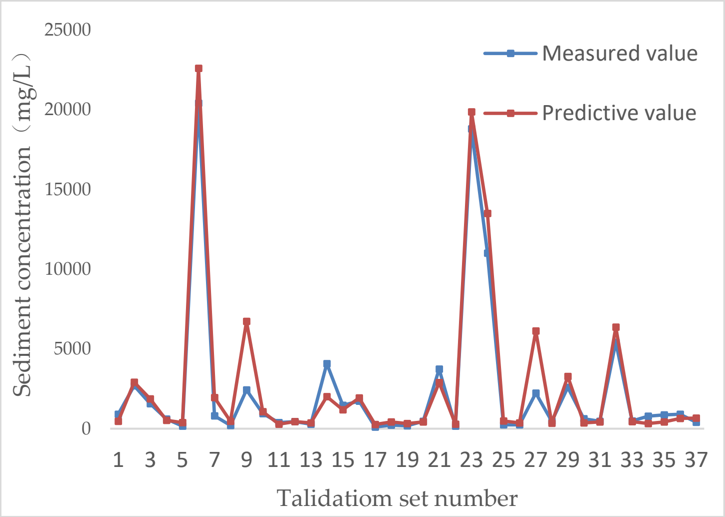


Figure 6. Effect of Suspended Sediment Validation Set at Longmen Station.

When all sample points are added to the random forest, the combination of (B4-B3)/(B4+B3), B8-B1, B7/B1, B12-B8A, and B8-B3 produces the most effective inversion impact. Figure 5 illustrates a scatter plot of predicted ln (SSC) and actual ln

(SSC). It indicates that R^2 is 0.873, MSE is 6.273%, and MAPE is 0.264 for the ln(SSC) calculated by random forest. It also shows that using the random forest model is feasible to monitor the suspended sediment concentration.

Figure 6 displays the impact of the suspension sediment validation set at the Long-men Station. At this time, the R^2 is 0.94, and the average absolute percentage error is 38.32%. It indicates that, in most cases, this model has relatively accurate results and can predict the suspended sediment concentration. However, there are also instances where the inversion accuracy is poor (points 9, 14, and 27). This shows that the model is still not perfect for prediction, and there is room for further development.

Sediment spatial distribution

The scene images with the least cloud amount and the best inversion effect in each season are selected, and the sediment concentration at the outlet of the Yellow River Canyon in 2019 is inverted by using ENVI and ArcGIS.

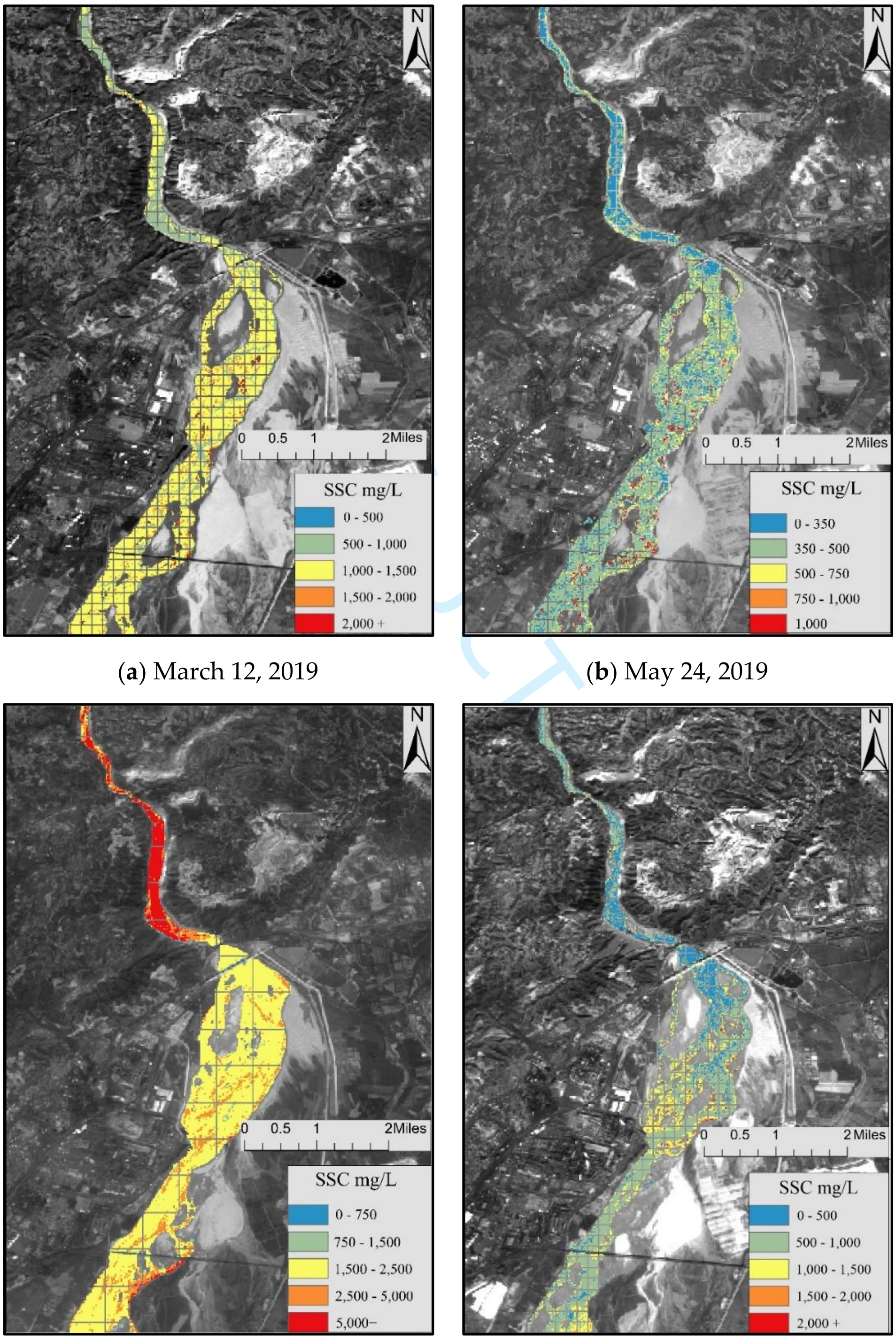
(1) Generally, the river's width is greatest in September, and the sediment concentration is much higher than in other months, followed by March. May has the lowest sediment concentration and the smallest river area.

(2) The sediment in the lower reaches of Yumen kou has been deposited for many years to form floodplains exposed to the water surface. Compared to other water bodies, sediment concentration is higher in the water bodies near the riverside and around the floodplain.

(3) The difference in sediment concentration between the upstream and downstream of the Yumen kou is evident in all the inversion images. During September, when the average sediment concentration was high, the upstream sediment

concentration was higher; during other months, when the sediment concentration was low, the upstream sediment concentration was lower.

In general, the random forest algorithm can accurately represent the sediment concentration in the waters near the Longmen site.



(**a**) March 12, 2019 (**b**) May 24, 2019

(**c**) September 6, 2019 (**d**) December 10, 2019

Figure 7. Estimated map of Longmen sediment distribution in 2019.

Discussion

This study found that the B4, B7, B8A, and B9 bands are sensitive to high-concentration sediment rivers. Moreover, combining the sensitive and non-sensitive bands, calculated by their spectral indices' ratio, difference, or NDVI, can achieve a good prediction effect. This result is consistent with the findings of Shen et al.(Shen, 2018) and Gitelson, A et al.(Gitelson et al., 2007). since the change in the wavelength leads to a change in the water body reflectivity. The blue-green band absorbs less sunlight, while the near-infrared band has a strong absorption ability. When sediment particles exist in the water, the reflection spectrum curve changes based on the sediment scattering, and the band reflectance increases. The peaks of the band reflectance appear in the yellow and red bands, which are consistent with the changes in the water body remote sensing reflectance data measured by others in the field(L. Lin & Wu, 2016). Therefore, the red and the near-infrared bands are sensitive to the reaction of suspended sediment in the water and can be better identified.

Various specific operations are derived from different bands in the variable combination, which means that the estimation of suspended sediment involves many variables and is affected by different factors. It is a complex process in which using certain estimated frequency bands is not simple(K. T. Peterson et al., 2019). There is a relatively high correlation between the predicted value calculated using the random forest model and the measured value. Besides, both the MSE and MAPE are ideal, indicating that using the random forest model is feasible to monitor the suspended sediment concentration. However, the accuracy of the data in the summer is still low, and the coefficient of determination is below 0.7. It is speculated that the reasons may include the following aspects:

1. The nature of suspended sediment itself, where suspended sediment particle size is small, unstable, and easily affected by river fluctuations. The suspended sediment concentration in regional river sections may change significantly in a short period of time. The satellite image obtains instantaneous information regarding the water body, and the provided band reflection information represents the suspended matter's reflectivity on the water body's surface. However, the measured sediment concentration at the monitoring site is the average suspended matter concentration within a specific depth range in each section. The relationship between the two is not equivalent(Pahlevan et al., 2017).

2. The sample size is limited. The sentinel image data obtained from Google Earth Engine belongs to the recording from 2019, and the useless images are filtered out. There are only 128 images involved in the modeling and about 30 images per season. Training such machine learning models to observe effective features is challenging.

3. Interference from other factors. The monitoring river section of the Longmen station is located in a mountainous area. The main factors affecting the distribution of surface suspended sediment are runoff and topography. While other environmental factors, such as meteorology and hydrology, also have an impact. Therefore, further analysis needs to integrate more influencing factors to make remote sensing monitoring more scientific and objective(Barnes, Hu, Bailey, Pahlevan, & Franz, 2021; Sun et al., 2022).

In August-September, there is much rain compared to other months, bringing a large amount of sediment from the upstream bank into the river, causing the sediment concentration in the river to rise(Feng, Zhao, Mu, & Tian, 2020). Therefore, the river area in September is the widest throughout the year, and the sediment concentration is much higher than in other months. The weather warmed up in March, and the snow and ice in the mountains near the source of the Yellow River melted a lot, and the amount of water and sediment in the middle and lower reaches increased.

Taking YuMen Kou as the boundary, the reason for the significant difference in sediment distribution between the upstream and downstream is that the upstream terrain is narrow, and the downstream is open. When the sediment concentration is high in summer, the upstream sediment is deposited, and the downstream is dispersed. On the other hand, when the overall sediment concentration is low in the other seasons, the turbulent upstream water brings the sediment downstream, and the floodplain in the downstream river still has a large sediment amount brought into the water.

Due to the late launch of the Sentinel-2 satellite, the sample period that can be obtained is short. In addition, the study area is primarily mountainous and has much cloudy and foggy weather; thus, the number of samples that can participate in the modeling after screening is limited. This is reflected in the seasonal inversion process, which is this paper's primary deficiency. In the initial variable selection, this paper selected the band combinations that operate in pairs only. For the multispectral satellite Sentinel-2, a combination of three or even four bands is achievable. However, the combination of more bands also makes interpreting the model more complex. This requires further research in later attempts to explore.

5. Conclusion

Machine learning is a hot and core issue in remote sensing image processing. Random forests have attracted much attention due to their high efficiency and accuracy and are widely used in various industries. The Yellow River has been plagued by sediment problems for thousands of years. Although much more attention has been paid to the

Yellow River's ecological problems in recent years, the sediment problem has been improved, but it is still a vital problem that requires attention(Hu & Zhang, 2022). Suspended matter in water is an essential parameter for evaluating water quality. The satellite remote sensing monitoring and research provide a promising approach for macroscopically understanding the changing laws of suspended sediment in the Yellow River. Based on the monitoring data of the Longmen sediment station and satellite remote sensing reflectance data, this paper firstly used linear fitting to select suitable band combinations, secondly constructed random forest non-parametric models, and finally estimated the SSC in the reaches around the Yellow River Canyon outlet.

The application range of the linear regression model is limited and can only deal with linear relationships, which is challenging to apply in complex and changeable remote sensing of water environment parameters. The random forest model is flexible, robust, simple, and convenient and has multiple advantages in parameter optimization, variable sorting, and subsequent variable analysis and interpretation. Moreover, it is suitable for remote sensing estimation of suspended matter monitoring in the Yellow River. In Sentinel-2 MSI, the red band B4 and the near-infrared band B8A are significant predictors of suspended sediment concentration in the random forest model. At the same time, the combination of a sensitive band and a non-sensitive band can effectively improve the inversion accuracy. It can be seen that there are many variables involved in the suspended sediment estimation, which is affected by many factors. Therefore, it is impossible to easily use a specific band for estimation, and the remote sensing estimation of suspended sediment requires the participation of multiple variables.

Acknowledgments

We Acknowledgment Henan University and the Yellow River Water Conservancy Commission for providing data

Disclosure statement

The authors declare no conflict of interest.

Funding

Key scientific research projects of colleges and universities in Henan Province (Grant No. 21A170007), and the National Natural Science Foundation of China (Grant No. 32130066).

References:

Barnes, B. B., Hu, C., Bailey, S. W., Pahlevan, N., & Franz, B. A. (2021). Cross-calibration of MODIS and VIIRS long near infrared bands for ocean color science and applications. *Remote Sensing of Environment, 260*, 112439. doi:https://doi.org/10.1016/j.rse.2021.112439

Batista, L. V. (2022). Turbidity classification of the Paraopeba River using machine learning and Sentinel-2 images. *IEEE Latin America Transactions, 20*(5), 799-805. doi:10.1109/TLA.2022.9693564

Bian, M., Skidmore, A. K., Schlerf, M., Wang, T., Liu, Y., Zeng, R., & Fei, T. (2013). Predicting foliar biochemistry of tea (Camellia sinensis) using reflectance spectra measured at powder, leaf and canopy levels. *ISPRS Journal of Photogrammetry and Remote Sensing, 78*, 148-156. doi:https://doi.org/10.1016/j.isprsjprs.2013.02.002

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Dehkordi, A. T., Ghasemi, H., & Zoej, M. J. V. (2021, 29-30 Dec. 2021). *Machine Learning-Based Estimation of Suspended Sediment Concentration along Missouri River using Remote Sensing Imageries in Google Earth Engine.* Paper presented at the 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS).

Espinoza Villar, R., Martinez, J.-M., Le Texier, M., Guyot, J.-L., Fraizy, P., Meneses, P. R., & Oliveira, E. d. (2013). A study of sediment transport in the Madeira River, Brazil, using MODIS remote-sensing images. *Journal of South American Earth Sciences, 44*, 45-54. doi:https://doi.org/10.1016/j.jsames.2012.11.006

Fang, X., Wen, Z., Chen, J., Wu, S., Huang, Y., & Ma, M. (2019). Remote sensing estimation of suspended sediment concentration based on Random Forest Regression Model. *Journal of Remote Sensing, 23*(04), 756-772.

Feng, J., Zhao, G., Mu, X., & Tian, P. (2020). Characteristics and mechanism of sediment transport in the Middle Yellow River. *Journal of Sediment Research, 45*(05), 34-41. doi:10.16239/j.cnki.0468-155x.2020.05.006

Francke, T., López-Tarazón, J. A., & Schröder, B. (2008). Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrological Processes, 22*(25), 4892-4904. doi:10.1002/hyp.7110

Fu, J., Zhang, P., Zheng, F., Yinghui, K., & Gao, Y. (2016). Dynamic Change Analysis of RainfallErosivity and River Sediment Discharge of He-Long Reach of the Yellow River from 1957 to 2011. *Transactions of the Chinese Society for Agricultural Machinery, 47*(02), 185-192+207. Retrieved from https://kns.cnki.net/kcms/detail/11.1964.s.20151019.1056.004.html

Gitelson, A., Garbuzov, G., Szilagyi, F., Mittenzwey, K. H., Karnieli, A., & Kaiser, A. (2007). Quantitative remote sensing methods for real-time monitoring of inland waters quality. *International Journal of Remote Sensing, 14*(7), 1269-1295. doi:10.1080/01431169308953956

Gordon, H., Brown, J., Brown, O., Evans, R., & Smith, R. (1988). A semianalytic radiance model of ocean color. *J. Geophys. Res., 93*. doi:10.1029/JD093iD09p10909

Gu, K., Zhang, Y., & Qiao, J. (2020). Random Forest Ensemble for River Turbidity Measurement From Space Remote Sensing Data. *IEEE Transactions on Instrumentation and Measurement, 69*(11), 9028-9036. doi:10.1109/TIM.2020.2998615

Hu, C., & Zhang, X. (2022). Research on the change of river water and sediment, regulation of water and sediment and utilization of sediment resources in China's rivers in recent ten years. *China Water Resources*(19), 24-28.

Ismail, R., Mutanga, O., & Kumar, L. (2010). Modeling the Potential Distribution of Pine Forests Susceptible to Sirex Noctilio Infestations in Mpumalanga, South Africa. *Transactions in GIS, 14*(5), 709-726. doi:https://doi.org/10.1111/j.1467-9671.2010.01229.x

Kilham, N. E., & Roberts, D. (2011). Amazon River time series of surface sediment concentration from MODIS. *International Journal of Remote Sensing, 32*(10), 2659-2679. doi:10.1080/01431161003713044

Li, C., Yu, Q., Gong, X., Yang, L., & Cao, Y. (2020). Remote Sensing Monitoring of Sediment Content Variation in Lower Reach of Yellow River since 1980s. *Environmental Science and Management, 45*(02), 165-170.

Li, J., Hao, Y., Zhang, Z., Li, Z., Yu, R., & Sun, Y. (2021). Analyzing the distribution and variation of Suspended Particulate Matter (SPM) in the Yellow River Estuary (YRE) using Landsat 8 OLI. *Regional Studies in Marine Science, 48*. doi:10.1016/j.rsma.2021.102064

Li, J., Xia, J., & Zhu, C. (2022). Characteristics and Influencing Factors of Thalweg Migration in the Xiaobeiganliu Reach of the Yellow River During the Period of Continuous Channel Aggradation. *Journal of Basic Science and Engineering, 30*(04), 883-892. doi:10.16058/j.issn.1005-0930.2022.04.008

Li, X., Jin, S., & Xu, J. (2012). Conservation Projects Impacts on Flood and Sediment in Hekouzhen to Longmen Region. *Yellow River, 34*(04), 87-89.

Li, Z. (2012). *Research on Remote Sensing Inversion Model of Marine Suspended Matter Concentration Based on Support Vector Machine.* (Master). China University of Geosciences (Beijing), Available from Cnki

Liao, Y., Zhang, W., & P.Y.Deschamps. (2005). Remote sensing of suspended sediments in China east coastal waters from Sea WiFS data. *Chinese Journal of Hydrodynamics*(05), 558-564.

Lin, L., & Wu, H. (2016). Study on Suspended Sediment Concentration Model in Case II Water for the Yellow River Estuary Based on the Spectral Reflectance. *Jiangsu Science & Technology Information*(02), 52-55.

Lin, X., Dong, C., Surin, M., & Hu, T. (2019). Analysis of the Relationship Between Scouring and Silting and Response of Water and Sediment in Xiaobeiganliu Reach of the Yellow River. *Yellow River, 41*(05), 5-8.

Mangiarotti, S., Martinez, J. M., Bonnet, M. P., Buarque, D. C., Filizola, N., & Mazzega, P. (2013). Discharge and suspended sediment flux estimated along the mainstream of the Amazon and the Madeira Rivers (from in situ and MODIS Satellite Data). *International Journal of Applied Earth Observation and Geoinformation, 21*, 341-355. doi:https://doi.org/10.1016/j.jag.2012.07.015

Ning, Z., Gao, G., & Fu, B. (2022). Characteristics and Attribution Analysis of Sediment Yield Changes in Helong Region of the Yellow River. *Research of Soil and Water Conservation, 29*(03), 38-42. doi:10.13869/j.cnki.rswc.2022.03.024

Ouyang, C., Wang, W., Tian, Y., & Tian, S. (2016). Evaluation on the variation of water-sediment and human activities in the He-Long Reach of the Yellow River over the past 60 years. *Journal of Sediment Research*(04), 55-61. doi:10.16239/j.cnki.0468-155x.2016.04.009

Pahlevan, N., Sarkar, S., Franz, B. A., Balasubramanian, S. V., & He, J. (2017). Sentinel-2 MultiSpectral Instrument (MSI) data processing for aquatic science applications: Demonstrations and validations. *Remote Sensing of Environment, 201*.

Park, E., & Latrubesse, E. M. (2015). Surface water types and sediment distribution patterns at the confluence of mega rivers: The Solimões-Amazon and Negro Rivers junction. *Water Resources Research, 51*(8), 6197-6213. doi:10.1002/2014wr016757

Peterson, K., Sagan, V., Sidike, P., Cox, A., & Martinez, M. (2018). Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. *Remote Sensing, 10*(10). doi:10.3390/rs10101503

Peterson, K. T., Sagan, V., Sidike, P., Hasenmueller, E. A., Sloan, J. J., & Knouft, J. H. (2019). Machine Learning-Based Ensemble Prediction of Water-quality Variables Using Feature-level and Decision-level Fusion with Proximal Remote Sensing. *Photogrammetric Engineering & Remote Sensing*.

Qiu, Z., Xiao, C., Perrie, W., Sun, D., Wang, S., Shen, H., . . . He, Y. (2017). Using Landsat 8 data to estimate suspended particulate matter in the Yellow River estuary. *Journal of Geophysical Research: Oceans, 122*(1), 276-290. doi:10.1002/2016jc012412

Ritchie, J. C., Schiebe, F. R., & Mchenry, J. R. (1976). *REMOTE SENSING OF SUSPENDED SEDIMENTS IN SURFACE WATERS*.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., . . . Adams, C. (2020). Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews, 205*. doi:10.1016/j.earscirev.2020.103187

Salama, M. S., Radwan, M., & van der Velde, R. (2012). A hydro-optical model for deriving water quality variables from satellite images (HydroSat): A case study of the Nile River demonstrating the future Sentinel-2 capabilities. *Physics and Chemistry of the Earth, Parts A/B/C, 50-52*, 224-232. doi:10.1016/j.pce.2012.08.013

Shen, M. (2018). *Research on remote sensing monitoring model of water environment in river basin.* (Master). University of Chinese Academy of Sciences (Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences), Available from Cnki

Smith, C., & Croke, B. F. W. (2005). Sources of uncertainty in estimating suspended sediment load. *IAHS-AISH publication*, 136-143.

Sun, X., Zhang, Y., Shi, K., Zhang, Y., Li, N., Wang, W., . . . Qin, B. (2022). Monitoring water quality using proximal remote sensing technology. *Science of The Total Environment, 803*, 149805. doi:https://doi.org/10.1016/j.scitotenv.2021.149805

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015, 7-12 June 2015). *Going deeper with convolutions.* Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tian, Z., Li, Z., Zhu, J., Xue, Z., & Zhao, Y. (2022, 17-22 July 2022). *Seasonal Variation of Suspended Sediments in the Yongding New River Estuary from 2017 to 2021.* Paper presented at the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium.

Tripathi, G., Pandey, A. C., & Parida, B. R. (2020, 1-4 Dec. 2020). *Spatio- Temporal Analysis of Turbidity in Ganga River in Patna, Bihar Using Sentinel-2 Satellite Data Linked with Covid-19 Pandemic.* Paper presented at the 2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS).

Umar, M., Rhoads, B. L., & Greenberg, J. A. (2018). Use of multispectral satellite remote sensing to assess mixing of suspended sediment downstream of large river confluences. *Journal of Hydrology, 556*, 325-338. doi:10.1016/j.jhydrol.2017.11.026

W., Z. (2012). *Remote Sensing Retrieval of Suspended Matters Based on Intelligent Calculation——A Case Study of Middle Yangtze River.* (Master). Beijing: University of Chinese Academy of Sciences,

Wang, H., Wu, X., Bi, N., Li, S., Yuan, P., Wang, A., . . . Nittrouer, J. (2017). Impacts of the dam-orientated water-sediment regulation scheme on the lower reaches and delta of the Yellow River (Huanghe): A review. *Global and Planetary Change, 157*, 93-113. doi:https://doi.org/10.1016/j.gloplacha.2017.08.005

Wang, S., Shen, M., Ma, Y., Chen, G., You, Y., & Liu, W. (2019). Application of Remote Sensing to Identify and Monitor Seasonal and Interannual Changes of Water Turbidity in Yellow River Estuary, China. *Journal of Geophysical Research: Oceans, 124*(7), 4904-4917. doi:10.1029/2019jc015106

Wang, Y. (2020). Analysis of Influence of Sediment Content on Monitoring Results of River Water Quality. *Shaanxi Water Resources*(12), 114-116. doi:10.16747/j.cnki.cn61-1109/tv.2020.12.041

WASER. (2021-07-21). UNESCO-ISI Online Training Workshop on Sediment Transport Measurement and Monitoring, July 5-9, 2021, successfully held. Retrieved from http://www.waser.cn/waser/NAA/webinfo/2021/07/1627854511860865.htm

Wen, Z. (2017). *Spatial and temporal variation of above-ground net primary productivity and its influencing factors in the fluctuating zone of the Three Gorges Reservoir.* (PhD). University of Chinese Academy of Sciences (Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences), Available from Cnki

Xu, J. (2009). A Study of Sediment Sink between Longmen and Sanmenxia on the Yellow River. *Acta Geographica Sinica, 64*(05), 515-530.

Yang, J., Fan, J., Zhang, Q., Yu, J., & Zhu, X. (2019). Review of suspended sediment content recognition in case II waters by remote sensing. *Yangtze River, 50*(07), 98-103. doi:10.16232/j.cnki.1001-4179.2019.07.016

YRCC. (2022-11-18). Yellow River Sediment Bulletin. Retrieved from http://www.yrcc.gov.cn/zwzc/gzgb/gb/nsgb/

Zhang H X, G. J. L., Zhu J Y and Yu J F. (2002). *Multivariate Data Analysis Methods and Applications with Few Observations*. Xi'an: Northwestern Polytechnical University Press.

Zhang, L., Zhang, L., & Du, B. (2016). Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geoscience and Remote Sensing Magazine, 4*(2), 22-40. doi:10.1109/MGRS.2016.2540798

Zhang, M., & Guo, B. (2018). Retrieval of Suspended Sediment Concentration in Zhoushan Coastal Area Satellite Based on GF-1. *Ocean Development and Management, 35*(01), 126-131. doi:10.20016/j.cnki.hykfygl.2018.01.022

Zhang, P., Cai, Q., Zheng, M., & He, T. (2020). Spatial and Temporal Distribution of Precipitation in Hekou-Longmen Region and lts Relationship with Sediment Yield. *Bulletin of Soil and Water Conservation, 40*(04), 25-31. doi:10.13961/j.cnki.stbctb.2020.04.004