REMOTE SENSING OF RESERVOIR WATER QUALITY: MODELLING OPTICALLY SENSITIVE AND NON-SENSITIVE PARAMETERS WITH LANDSAT

Suraj PAWAR [1], ***Rushikesh KULKARNI*** [1*], ***Kanchan KHARE*** [2] ***and Humera KHANUM*** [1]

DOI: 10.21163/GT_2026.211.05

ABSTRACT

This study integrated in situ water quality measurements with Landsat 8 Operational Land Imager (OLI) data to assess the water quality of the Khadakwasala Reservoir in Pune, India. Stepwise regression analysis was applied to develop models correlating Landsat-derived spectral indices with key water quality parameters, including turbidity, chlorophyll-a, dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD). The models achieved predictive accuracy with R2 values of 0.85 for turbidity, 0.90 for chlorophyll-a, 0.65 for DO, 0.64 for BOD, and 0.84 for COD. Turbidity was identified as a key predictor for the non-sensitive parameters (DO, BOD, COD). These models provide a scalable method for monitoring water quality, facilitating broader spatial assessments using satellite data. The dataset supports the reuse of remote sensing data for water quality management and environmental monitoring. The distributed images of water quality parameters can be obtained at the repository address mentioned in the data availability section.

Key-words: *Water Quality Assessment; Remote Sensing of Lakes; Landsat 8/9 OLI; Surface Water Quality Parameters; Stepwise Regression Model; Chlorophyll, Turbidity; Khadakwasla Reservoir*

1. INTRODUCTION

Access to clean and safe water is of paramount importance for human health and ecological integrity. Surface water bodies, including reservoirs, are vital freshwater resources, but their quality is increasingly threatened by anthropogenic activities and natural processes (Akhtar et al., 2021). Monitoring and evaluation of water quality parameters is necessary to Effective water resource management.

Traditional water quality monitoring approaches often rely on in situ sampling and laboratory analysis, which can be time consuming, expensive, and spatially limited (Palmer et al., 2015; Olmanson et al., 2015; Chang et al., 2014). Remote sensing techniques, employing satellite imagery, offer a spatially comprehensive and cost-effective alternative for assessing water quality parameters over large geographical areas, (Olmanson et al., 2015; Chang et al., 2014. Landsat satellites, with their long-term data archives and moderate spatial resolution, have proven invaluable for monitoring inland water bodies (Vakili & Amanollahi, 2020; Wang et al., 2019).

The cornerstone of water quality recovery relies on establishing correlations between the concentrations of water components and the scattered signals - in particular, the radiance emitted by the water as observed by sensors, (Kirk, 2010). Models for retrieval can be built based on the relationship of underlying optical properties (IOPs) to remote sensing reflectance, because IOPs are unique optical properties of water that are independent of external conditions and depend entirely on the composition of the water body. Consequently, it becomes feasible to directly retrieve water quality parameters such as Chlorophyll-a (Chl), suspended matter (SM), and colored dissolved organic matter (CDOM), (Palmer et al., 2015; Wang et al., 2017; Li et al., 2018).

---

[1] *Symbiosis Institute of Technology, Symbiosis International University, 412115 Pune, Maharashtra, India; suraj.pawar.phd2024@sitpune.edu.in (SP); *corresponding author rushikeshk@sitpune.edu.in (RK); humera.khanum@sitpune.edu.in (HK).*

[2] *Sevavardhini, Water Resource Department, Pune Division, 411030 Pune, Maharashtra, India, kanchankhare@sevavardhini.org (KK).*

Water constituents devoid of optical activity like Total Nitrogen (TN), Total Phosphorus (TP), Chemical Oxygen Demand (COD), and Dissolved Oxygen (DO), which lack direct optical characteristics, can be approached either by exploring interrelationships between various substances in water or by deploying Artificial Intelligence (AI) techniques, (Sun et al., 2014; Sharaf El Din et al., 2017).

This study focused on a conventional regression-based approach to link optically sensitive parameters (turbidity, chlorophyll a) with non-optically sensitive parameters (DO, BOD, COD). While several recent works have highlighted the potential of artificial intelligence (AI) frameworks for indirect estimation of non-optically sensitive parameters, such approaches often face data limitations and require large, diverse training sets that were not available for the Khadakwasla case study. Therefore, instead of AI-driven models, our study demonstrates how statistically robust relationships can be derived between optically sensitive parameters and optically non-sensitive parameters to predict. This approach is particularly advantageous in data-limited settings, as it avoids the overfitting risks and transferability issues often associated with AI models.

However, quantifying a wide range of water quality parameters is challenging, particularly when calculating water-leaving reflectance due to the complex interactions of radiation among the atmosphere, water surface, and water body across various wavelengths. Generally, remote sensing-based water quality retrieval methods can be classified into empirical, physical, semi-empirical, and intelligent models, with AI methods standing out as a distinct empirical approach that uses statistical techniques (Palmer et al., 2015; Wang & Yang, 2019). Many reviews have highlighted the major advancements in water quality retrieval using remote sensing, especially focusing on different retrieval methods and the use of multi-source remote sensing data (Palmer et al., 2015; Chang et al., 2014; Wang et al., 2019; Chen et al., 2013; Kuhwald & Oppelt, 2016; Sagan et al., 2020; Chawla et al., 2020).

This study aims to examine the Khadakwasla Reservoir water quality. The reservoir serves as a primary drinking and farming water source in Pune, Maharashtra, which is found in India. The study has combined the in situ physical measurements of water quality parameters from Landsat 8 Operational Land Imager (OLI) surface reflectance data to acquire a wide-ranging knowledge about the lake's status on water quality. Furthermore, this integration leverages the strengths of both methods, resulting in a more comprehensive assessment with improved spatial accuracy.

This project involves the estimation of quality parameters of water such as chlorophyll (Chl), chemical oxygen demand (COD), turbidity (Tur), bio chemical oxygen demand (BOD), and dissolved oxygen (DO) based on sampling points of choice of the reservoir. The main task of the investigation is to use the pre-processed Landsat 8 Operational Land Imager (OLI) surface reflectance data to estimate the parameters of surface water quality (SWQP). Analytical approaches aimed at the statistical testing the relationship between the in-situ water quality measurements and the corresponding Landsat 8 OLI surface reflectance values will be the regression analytical methods.
The data obtained as a result will be aimed at providing a methodology to identify water quality in Khadakwasla Reservoir through the combination of in-situ measurements and remote sensing observations. The developed statistical association among parameters of in-situ water quality and those provided by the Landsat 8 OLI spectral reflectances make it possible to form the predictive models which will be able to approximate the spectra of water quality parameters in different across the reservoir sectors. Such a strategy is likely to work together with effective water resource management methods.

2. STUDY AREA AND METHODOLOGY

2.1. Study area

The Khadakwasala Reservoir is selected as a study area in this study due to easy accessibility for in-situ water sampling for testing water quality and wide spatial coverage (**Fig. 1**). As per the Water Resource Department data, the surface area of the reservoirs varies from ~ 8.5 Sq.km to ~12.0 Sq.km corresponding to the seasonal variation. This spatial extent is enough to use the high and moderate resolution satellite for analysis. The Pune metropolitan city and its surrounding areas primarily depend on the Khadakwasla reservoir for their water needs.

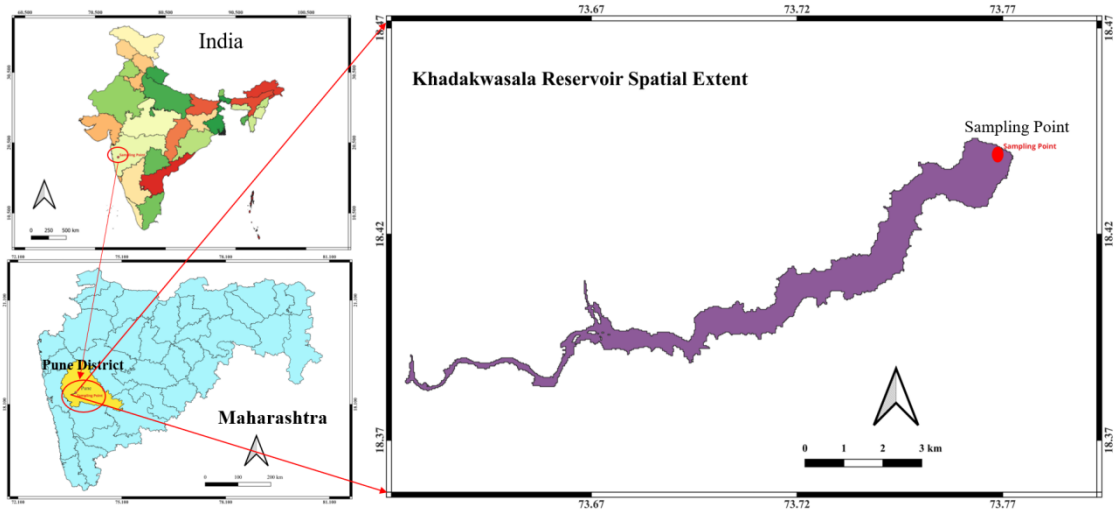

Fig. 1. Location of the Study Area: Khadakwasala Reservoir (Sampling Location :18.4390 N / 73.7720 E).

It is a source of water for domestic needs, agricultural needs, industrial needs, and hydroelectric power generation. For these reasons, understanding Khadakwasala's storage potential becomes integral to managing water resources within the locality. The Khadakwasala Reservoir was built in 1879 near Pune. Maharashtra, India, primarily to meet Pune's drinking water source and irrigation needs. It is located on the Mutha River and holds an approximate storage capacity of 2.46 million cubic meters.

2.2. Methodology

The methodology presented in the flowchart (**Fig. 2**) describes an integrated approach combining remote sensing and in-situ data for the prediction of surface water quality parameters (SWQPs). This process involves the use of Landsat 8/9 imagery along with physical water quality measurements. Satellite data goes through pre-processing steps, including geometric, radiometric, and atmospheric corrections, as well as masking of water bodies to obtain the reflectance of the upper atmosphere (ToA). Concurrently, in-situ measurements are gathered from the Khadakwasla Dam through physical sampling and laboratory analysis to determine values for turbidity (Tur), chlorophyll-a (Chl), dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD).

The relationship between spectral data and SWQPs is established using stepwise regression, a statistical method that adds or removes predictor variables in an iterative manner based on their significance. At each step, the p-value of the independent variable is evaluated and only the value that is less than the specified threshold is retained in the model. This helps identify the most relevant spectral bands or indices that contribute significantly to the prediction of each water quality parameter, improving the accuracy and interpretation of the model. The SWQP's are divided into optically sensitive (Tur, Chl) and non-sensitive (DO, BOD, COD) groups, followed by training (75%) and testing (25%) datasets. The derived models are verified through accuracy analysis using test data.

2.3. Data Acquisition

2.3.1. Landsat OLI Images

The images of Landsat 8/9 OLI were accessed from the Earth Explorer platform *(http://earthexplorer.usgs.gov)* operated by the United States Geological Survey (USGS). The Landsat imagery for the Khadakwasala Reservoir was selected based on the specific path 147 and row 47 criteria. This led to a collection of 23 cloud-free Landsat OLI 8/9 images, which were used for the analysis. The downloaded images covered the period from 20th October 2022 to 22nd April 2023.

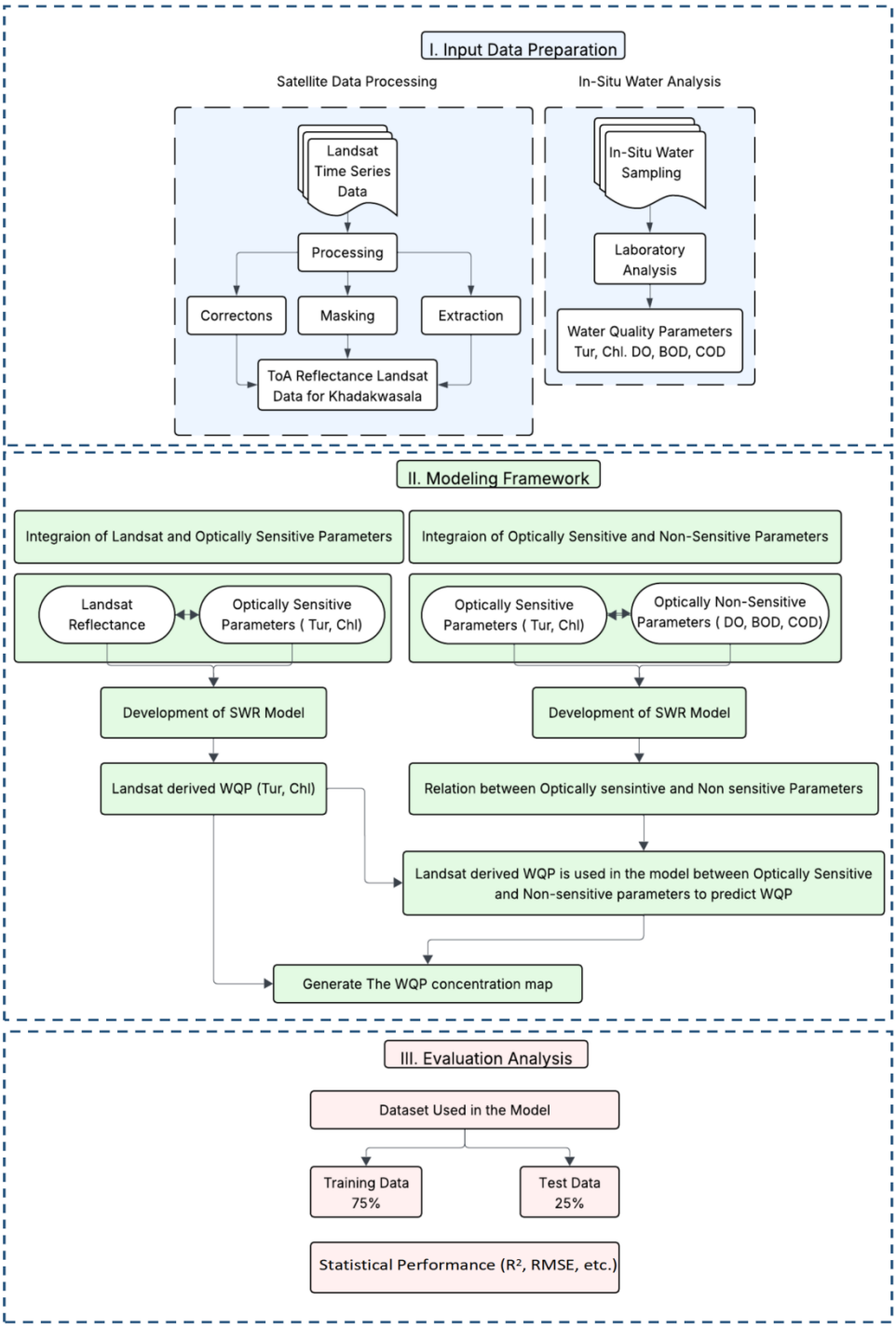

Fig. 2. Methodology Workflow for estimation of water quality parameters from Landsat 8/9 OLI using SWR.

2.3.2. Data Pre-processing

The first step in the methodology is data acquisition, which includes two types of data: in-situ water quality parameters and satellite imagery. The in-situ water quality data was specifically collected on dates when corresponding Landsat 8/9 OLI images were available to maintain data

consistency. The next process is the image processing of raw images for further analysis. The Landsat images are processed to reduce the interference caused by the atmosphere and sensors, allowing for accurate interpretation and analysis of water quality parameters. The image processing involves radiometric corrections, geometrical corrections, atmospheric corrections, and Dark Object Subtraction.

2.3.3. In-situ water quality parameter data

Physical water samples from the Khadakwasala Reservoir were collected at coordinates 18.4390 N and 73.7720 E. Sampling was conducted from October 20, 2022, to April 22, 2023, aligning with the satellite imagery acquisition period. These samples were analysed for various water quality parameters, including Chl, Turbidity, DO, BOD, and COD. Turbidity was measured on-site using a standard turbidity meter, while the other parameters were tested in the laboratory following the American Public Health Association (APHA) standards.

To ensure consistency and minimize contamination from surface debris and atmospheric effects, all measurements were taken at a depth of 0.25 meters. The clean, sterile, and non-reactive polyethylene bottles were used for collecting water samples. After collection, the samples were placed in an ice container and transported to the laboratory within the holding time specified by APHA guidelines.

2.3.4. Analytical Methods

The analysis of the sampled water was carried out following the Standard Methods outlined by APHA. DO levels were quantified utilizing the Winkler titration method and measured in milligrams per liter (mg/L). BOD was determined using the 5-day BOD test, which measures the amount of oxygen consumed by microorganisms during the aerobic breakdown of organic material. These concentrations were reported as milligram per liter (mg/liter). The chemical oxygen demand (COD) was measured using the closed reflux titrimetry which is used to measure the oxygen equivalent that is needed to completely oxidize all the organic and inorganic matter that exists in the aqueous sample and the reflux reaction result is expressed in mg/L. The level of chlorophyll which is a measure of the phytoplankton biomass was determined using a pigment extraction technique. The obtained figures were recorded in micrograms per liter (µg/L). Turbidity used as a surrogate of suspended solids was estimated with the help of a nephelometric turbidity unit (NTU) meter. The device measures the light scattering of suspended particles hence recording an indirect measure of turbidity on the NTU.

2.4. Stepwise Regression Technique

Stepwise regression (SWR) is one of the statistical methods where the influential predictor variables in a set of candidates are given sequentially, to form a strong predictive model. The Akaike Information Criterion (AIC) is an indispensable factor in model selection which gives a balance between the accuracy of prediction and the complexity of the model, thus alleviating the problem of overfitting. SWR makes use of forward selection and backward elimination steps which is driven by AIC. In the case of optical sensitive parameters, Landsat spectral bands and spectral ratios are considered as the independent variables and turbidity and chlorophyll concentrations are the dependent variables. In contrast, the chlorophyll and turbidity are taken as independent variables in optically non-sensitive parameters, whereas, the dissolved oxygen, biochemical oxygen demand, and chemical oxygen demand are dependent variables. The procedure chooses the variable that produces least AIC which makes it the most applicable predictor and includes it in the model. k suing the models, k -fold cross- validation was used to test the validity and reliability of the models. The best SWR was chosen based on maximization of adjusted R^2, minimum root-mean-square error (RMSE), mean absolute error (MAE) among another basis such as the AIC criterion.

2.4.1. Relationship of Landsat with optically sensitive water quality parameter

This research used SWR to find out the relations between surface reflectance measured using Landsat and in-situ of water quality factors, that is, turbidity (Tur) and chlorophyll-a concentration (Chl), in which the first five reflection bands and several band combinations as independent variables.

The combinations used in the model were (Band 3 / Band 2), (Band 2 + Band 3), (Band 3 + Band 4), and (Band 2 / Band 1). The model uses established research from the literature to select specific

bands and combinations for estimating water quality parameters like Tur and Chl, with in-situ measurements as dependent variables.

The selection of band combinations in this study was not arbitrary but guided by established findings in remote sensing of inland waters. Band ratios such as B3/B2 (green/blue) and B2/B1 (blue/coastal aerosol) are widely reported as effective indicators of turbidity and total suspended solids (TSS), since shorter wavelengths are strongly scattered by suspended particles, while green reflectance is more stable under varying concentrations. Similarly, combinations involving B3 and B4 (green + red) are sensitive to chlorophyll-a because of the strong absorption of red light by chlorophyll pigments and the differential scattering in the green band. The additive indices (e.g., B2 + B3) exploit the complementary absorption and scattering properties of blue and green bands, which enhance sensitivity to phytoplankton biomass and light attenuation. Accordingly, the combinations tested in this study—(B3/B2), (B2 + B3), (B3 + B4), and (B2/B1)—were adopted from prior literature demonstrating their utility in estimating optically active parameters like chlorophyll-a and turbidity in reservoirs and lakes (Vakili & Amanollahi, 2020; Kirk, 2010; Chen et al., 2013; Sagan et al., 2020). These indices were therefore incorporated into the stepwise regression framework to ensure comparability with earlier studies and to leverage their proven biophysical sensitivity, rather than being developed empirically from scratch. This approach both strengthens model interpretability and ensures consistency with established remote sensing methodologies.

2.4.2. Relationship of Landsat with optically non-sensitive water quality parameters

The first step in our analysis is to establish a correlation between the optically sensitive parameters, derived from Landsat, and the optically non-sensitive parameters. We treat the optically sensitive parameters as independent variables and the non-sensitive parameters as dependent variables. Using the relationships from the previous step, we derive the optically sensitive parameters from Landsat data, allowing us to estimate the optically non-sensitive parameters despite their lack of direct correlation with Landsat reflectance.

3. DATA RECORD

This data record provides a distinctive evaluation of the surface water quality of the Khadakwasala reservoir by utilizing two separate datasets: physically sampled and measured surface water quality data, and water quality data obtained from Landsat 8 and 9 OLI. Landsat 8/9 OLI-derived data is assessed by integrating physical measurements with top-of-atmospheric reflectance, enabling a thorough spatial analysis of water quality parameters.

3.1. Data Record

In-situ water quality data for the Khadakwasala Reservoir, including measurements of DO, BOD, COD, Chl, and Tur, were collected from October 2022 to April 2023 (**Table 1**). Chl serves as a vital indicator of phytoplankton density and productivity in aquatic ecosystems, reflecting oligotrophic to mesotrophic states. Reduced Chl concentrations signify limited algal blooms, thereby maintaining water quality and ecological equilibrium. Turbidity, a factor influencing light penetration and habitat quality, varies from 1.82 to 3.42 NTU, signifying clear water conditions. DO concentrations ranging from 7.48 to 7.84 mg/l signify optimal conditions for aquatic organisms. The BOD values fluctuate between 3.70 and 4.25 mg/l, signifying heightened organic pollution. The COD values span from 14.99 to 24.25 mg/l, signifying the existence and concentration of various pollutants, including those resistant to microbial degradation. Consistent monitoring of BOD facilitates the comprehension of the impacts of activities on water quality and the identification of trends in organic matter accumulation. Elevated COD levels may signify sources of non-biodegradable pollutants.

The temporal trends of the water quality data set show seasonal variation following precipitation, temperature fluctuation, and human activities. Higher levels of COD and BOD may occur during periods of increased runoff, while temperature stratification during different months or years can lead to varying DO levels. Throughout the period, the Khadakwasala Reservoir maintained quality levels that are safe enough to support life in the water bodies without showing any sign of intense pollution or eutrophication.

Table 1.

In-situ SWQPs measured at Khadakwasala reservoir.

Date	DO (mg/L)	BOD (mg/L)	COD (mg/L)	Chl (ug/L)	Tur (NTU)
20-10-2022	7.68	3.97	16.61	0.15	2.03
28-10-2022	7.76	3.93	24.25	0.14	1.87
05-11-2022	7.84	4.04	17.99	0.18	1.82
13-11-2022	7.70	3.84	18.91	0.14	2.06
21-11-2022	7.67	3.92	20.88	0.15	2.10
29-11-2022	7.65	3.83	21.09	0.15	2.07
07-12-2022	7.73	3.93	17.15	0.16	1.91
23-12-2022	7.70	3.79	21.30	0.15	2.09
31-12-2022	7.67	3.97	20.16	0.15	2.04
08-01-2023	7.67	4.25	23.43	0.15	1.86
16-01-2023	7.59	3.83	19.86	0.15	2.06
24-01-2023	7.76	3.97	18.10	0.15	1.99
01-02-2023	7.68	3.87	16.60	0.17	2.07
09-02-2023	7.61	3.80	17.59	0.15	2.15
17-02-2023	7.48	3.80	15.11	0.15	2.21
25-02-2023	7.51	3.73	14.99	0.16	2.19
05-03-2023	7.57	3.90	17.70	0.19	1.95
13-03-2023	7.61	3.70	16.20	0.21	2.12
21-03-2023	7.50	3.70	15.90	0.22	3.42
29-03-2023	7.59	4.00	19.00	0.17	2.06
06-04-2023	7.48	3.90	17.40	0.21	2.32
14-04-2023	7.58	3.80	17.70	0.20	2.08
22-04-2023	7.75	4.00	16.20	0.22	1.85

3.2. Landsat 8 OLI-derived Water Quality Data

Landsat 8 OLI-derived data were utilized to assess SWQPs in the Khadakwasala Reservoir. A total of 24 samples were collected during the Landsat satellite pass from October 20, 2022, to April 22, 2023, corresponding to the post-monsoon through pre-monsoon dry period in Pune. The sample from March 21, 2023, was excluded from the analysis due to cloud cover, which compromised the accuracy of the Landsat image for estimating surface water quality parameters. Thus, the analysis was performed on the 23 remaining samples. Consequently, the regression models are calibrated for dry-season conditions and should be applied to monsoon periods only with caution, ideally following season-specific validation or recalibration.

The differentiation between optically sensitive (Tur and Chl) and optically non-sensitive (DO, COD, and BOD) parameters underscores the distinct approaches required for their estimation from remote sensing data. The optically sensitive parameters were directly estimated from the spectral data due to their strong correlation with surface reflectance. In contrast, the optically non-sensitive parameters, although not directly retrievable, were modeled using indirect approaches that leverage the relationship with the optically sensitive parameters.

3.2.1. Landsat-derived Turbidity

The Tur concentration in the Khadakwasla Reservoir was estimated using a SWR analysis. The SWR analysis identified the band ratio of B2 (blue) to B1 (coastal aerosol) as the most significant variable for predicting turbidity concentration. The results of the multiple linear regression model using the selected variables are shown in **Table 2**. The resulting regression equation for Tur concentration is given by Eq. (1):

$$Turbidity = 1.01 \times \frac{B2}{B1} + 1.02 \quad (1)$$

This equation indicates that the ratio of reflectance between the B2 and B1 bands is a strong predictor of turbidity concentration, explaining 84.5% of the variance in the measured turbidity values ($R^2 = 0.85$). The model's statistical significance is reinforced by an F-statistic of 76.04 and a p-value below 0.05, signifying a highly significant correlation between the predictor variable and turbidity concentration.

The regression coefficients, comprising the constant term (1.02) and the coefficient for the (B2/B1) band ratio (1.01), were determined to be statistically significant at the 0.05 level. The t-values (8.76 and 8.72) and their corresponding p-values (both below 0.05) confirm the strength of these coefficients. The 95% confidence intervals for the coefficients ([0.02, 0.98] for both) further substantiate their reliability and the robustness of the predictive relationship.

The results validate that the (B2/B1) band ratio is a robust and statistically significant predictor of turbidity concentration (see **Fig. 3a**) in the Khadakwasala Reservoir, illustrating the efficacy of remote sensing methodologies for water quality evaluation.

3.2.2. Landsat-derived Chlorophyll

The Chl data for the Khadakwasala Reservoir, derived from Landsat 8 OLI, were acquired via a SWR analysis. This analy sis identified six principal spectral variables as significant predictors of Chl concentration: B3/B2, B4, B3+B4, B2, B5, and B2+B3. The results of the multiple linear regression model using the selected variables are shown in **Table 2**. The regression model is represented by the subsequent (Eq. 2):

$$Chlorophyll = -0.16 \times \left(\frac{B3}{B2}\right) - 0.96 \times B4 + 1.54 \times (B3 + B4) - 2.12 \times B2 - 0.49 \times B5 + 0.37 \times B2 + 0.31 \quad (2)$$

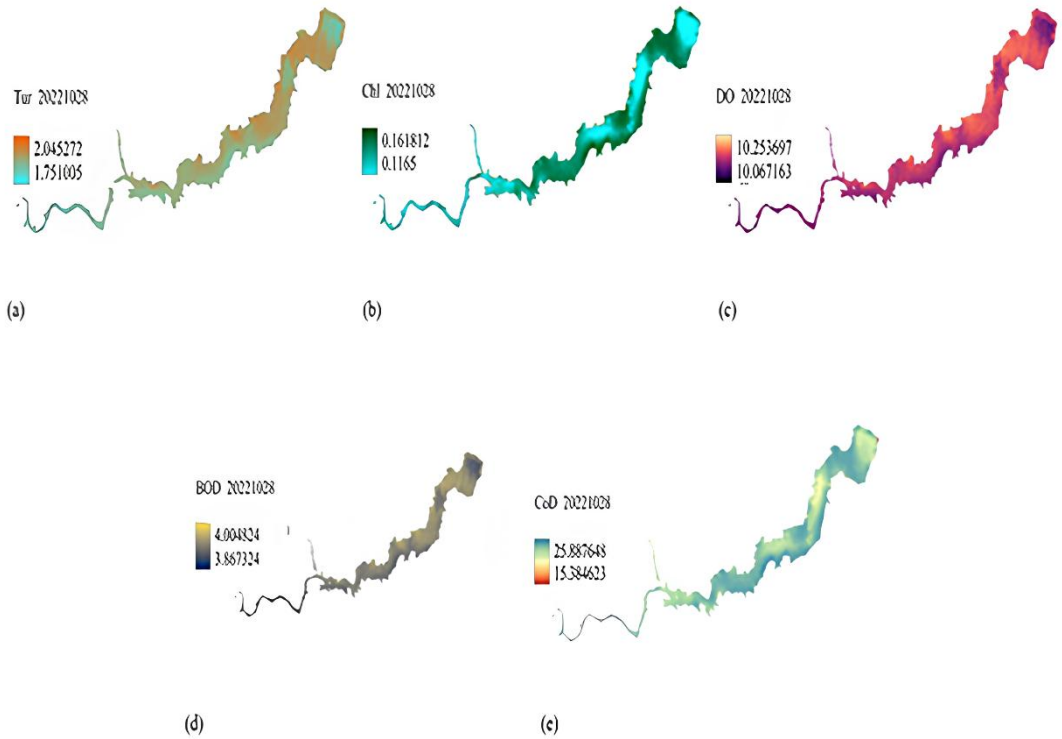

Fig. 3. Sample images of the Spatial variability of water quality parameters concentration across Khadakwasala reservoir sampled on 28-10-2022: (a) Turbidity (NTU); (b) Chlorophyll a, (µg/L); (c) DO (mg/L); (d) BOD (mg/L); (e) COD (mg/L).

This model achieved a R^2 value of 0.90, indicating that 89.6% of the variability in chlorophyll concentration can be explained by the selected spectral predictors. The adjusted R^2 of 0.85 demonstrates the model's robustness, accounting for the number of predictors while maintaining a high level of explanatory power. The F-statistic of 17.31 and the associated p-value less than 0.05 confirm the statistical significance of the model, underscoring the effectiveness of the selected spectral bands and ratios in predicting chlorophyll concentration.

The strong predictive capability of the SWR model, as evidenced by the high R^2 value, emphasizes the utility of Landsat 8 OLI spectral data in assessing Chl concentrations in surface water at Khadakwasala Reservoir. The identified relationships align with known optical properties of Chl and suspended matter, where the B2 and B3 bands are sensitive to Chl due to their response to specific light absorption and scattering characteristics. The Landsat 8 OLI-derived Chl data contribute valuable insights into the spatial and temporal distribution of chlorophyll concentration in the Khadakwasala Reservoir (refer to **Fig. 3b**), offering a reliable tool for water quality monitoring and management.

3.2.3. Landsat-derived DO

The SWR analysis used to estimate DO concentration identified turbidity as the sole predictor. The results of the multiple linear regression model using the selected variables are shown in **Table 2**. The linear regression model developed is described by the following equation (Eq.3):

$$DO = -0.63 \times Tur + 8.96 \quad (3)$$

The R^2 value of this model was 0.65 which shows that 64.9percent of the variance of dissolved oxygen concentration can be explained by the level of turbidity. The adjusted R 2 of 0.62, has validated the fact that the model still has a fairly adequate explanatory power whilst still taking into consideration the number of the predictors used. Statistical significance of the model is further confirmed by this F statistic that equals 25.89 and p value below, 0.005 and shows that turbidity is the significant predictor of dissolved oxygen concentration (refer **Fig. 3c**). The coefficient for Tur in the regression model is -0.63, with a p-value less than 0.001, demonstrating that Tur has a statistically significant negative effect on DO concentration. This negative relationship implies that as Tur increases, the DO concentration decreases. This is an ecologically sound trend in the sense that increased turbidity reduces light penetration into the water column thus repressing the photosynthetic mechanism of aquatic plants and algae of the water column, which eventually leads to low oxygen generation.

3.2.4. Landsat-derived BOD

The stepwise regression analysis for estimating BOD concentration identified turbidity as the only significant predictor. The results of the multiple linear regression model using the selected variables are shown in **Table 2**. The linear regression model is described by the following equation (Eq.4):

$$BOD = -0.87 \times Tur + 5.67 \quad (4)$$

This model attained a R^2 value of 0.64, signifying that 64.4% of the variability in BOD concentration is explicable by turbidity levels. The adjusted R^2 value of 0.62 indicates that the model possesses a substantial degree of explanatory power, considering the number of predictors employed. The F-statistic of 25.30 and the corresponding p-value below 0.05 establish the statistical significance of the model, signifying that turbidity is a significant predictor of BOD concentration. The coefficient for turbidity in the regression model is -0.869, with a p-value less than 0.001, demonstrating that turbidity has a statistically significant negative effect on BOD concentration. This negative relationship implies that as turbidity increases, BOD concentration decreases. This inverse relationship may be explained by the fact that higher turbidity levels can lead to reduced light

penetration in the water column, which may limit primary productivity and reduce the availability of organic matter for microbial decomposition, ultimately resulting in lower BOD concentrations (refer Fig. 3d).

The significant relationship between Tur and BOD concentration underscores the utility of remote sensing data, particularly Landsat 8 OLI-derived Tur, in estimating BOD in surface waters. While Tur alone was identified as a significant predictor, the complexity of BOD dynamics suggests that additional optically sensitive parameters or more advanced modelling approaches could further enhance the predictive accuracy of the model.

3.2.5. Landsat-derived COD

The stepwise regression analysis identified Tur and Chl as significant predictors of COD concentration. The results of the multiple linear regression model using the selected variables are shown in **Table 2**. The resulting linear regression model is described by the following equation (Eq.5):

$$COD = -233.18 \times Chl - 17.16 \times Tur + 89.44 \quad (5)$$

This model achieved R^2 value of 0.84, indicating that 83.5% of the variability in COD concentration can be explained by turbidity and chlorophyll levels. The adjusted R^2 value of 0.81 confirms that the model retains a high level of explanatory power while accounting for the number of predictors used. The F-statistic of 32.79 and the associated p-value less than 0.05 validate the statistical significance of the model, highlighting that Tur and Chl are meaningful predictors of COD concentration (refer **Fig. 3e**).

The regression coefficients for Chl (-233.18) and Tur (-17.16) are both negative, with p-values less than 0.001, indicating their statistically significant contributions to the model. The negative coefficients suggest that as chlorophyll and turbidity increase, COD concentration decreases. This inverse relationship can be attributed to the fact that higher Chl concentrations are indicative of increased photosynthetic activity, which can enhance oxygen production and thus reduce COD levels. Additionally, higher Tur levels may correspond with an abundance of suspended particles that limit light penetration, potentially decreasing the availability of organic matter and, consequently, COD concentrations. Identifying TUR and CHL as significant predictors emphasizes the utility of Landsat-derived data for assessing COD concentrations in surface waters. This approach allows for indirect estimation of COD, which is challenging to measure directly through remote sensing due to its optical insensitivity.

Table 2.
Ordinary Least Squares result obtained by SWR analysis between Landsat spectral reflectance bands and SWQPs at Khadakwasala.

SWQPs	Selected Features	R^2	Adj. R^2	F-Statistics	p-value
Turbidity	'B2 / B1'	0.85	0.83	76.04	< 0.05
Chlorophyll	'B3/B2', 'B4', 'B3+B4', 'B2', 'B5', 'B2+B3'	0.90	0.85	17.31	< 0.05
DO	'Turbidity'	0.65	0.62	24.89	< 0.05
BOD	'Turbidity'	0.65	0.62	25.30	< 0.05
COD	'Chlorophyll', 'Turbidity'	0.84	0.81	23.45	< 0.05

4. TECHNICAL VALIDATION

The Water quality parameters of Khadakwasla reservoir were estimated using Landsat-based SWR models and validated using a training dataset of 17 records and a testing dataset of six records. The models were assessed for predictive efficacy using metrics like R2, RMSE, and p-values to determine the correlation between predictor variables and water quality parameters.

4.1. Model Validation Results

Model of Turbidity. The Turbidity model had an R^2 of 0.62, which means that almost 62% of the observed turbidity data variability is captured by the model. An RMSE of 0.07 indicates the average deviation between predicted and observed turbidity values, indicating moderate levels of predictive precision (ref. **Fig. 4a**).

Model of Chlorophyll. The Chl model had an R^2 equal to 0.65, indicating that approximately 64.84% of the chlorophyll variability was explained by the model. The RMSE of 0.0095 indicates the average difference between actual chlorophyll values and their respective predicted values.There is a strong p-value for the model's significance of 0.001 or less, (ref. **Fig. 4b**).

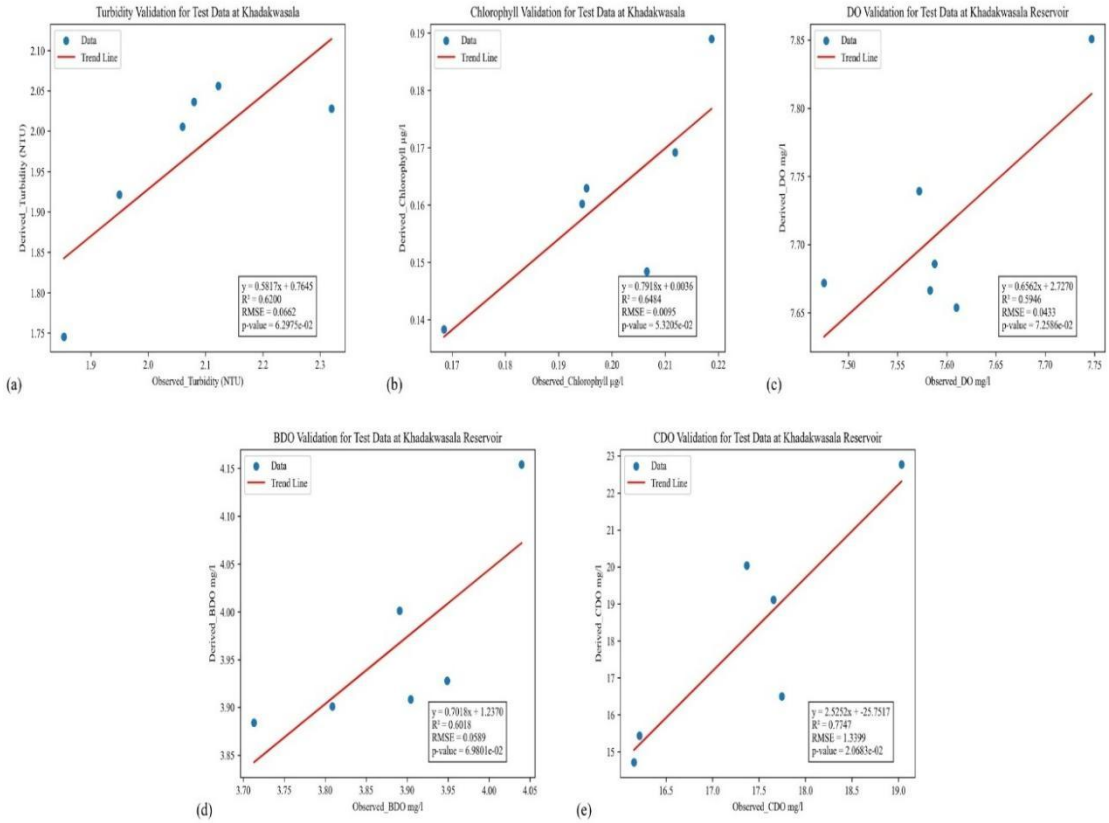

Fig. 4. Test-set validation of stepwise-regression models for Landsat 8/9 OLI water-quality retrievals sampled on 28-10-2022 (Observed vs Modelled), Khadakwasala Reservoir: Turbidity, Chlorophyll a, Dissolved Oxygen, Biochemical Oxygen Demand and Chemical Oxygen Deman.

Model of Dissolved Oxygen. The Dissolved Oxygen (DO) model attained a R^2 score of 0.60 for the observed DO data, where about 59.46% variance could be explained. Low RMSE values (0.043) signify that predictive accuracy is not very high in the case of DO concentration. The statistical validity of this model has been attested by the p-value being less than 0.001 (ref. **Fig. 4c**).

Model of Biological Oxygen Demand. The Biological Oxygen Demand (BOD) model displayed an R^2 value equal to 0.60, implying that it covers approximately 60.18% of all BOD data variances. RMSE was estimated at 0.06, thus showing an average difference between the projected and actual BOD values. A P-value less than 0.001 confirms the significance of the relation between turbidity and BOD (ref. **Fig. 4d**).

Model of Chemical Oxygen Demand. In the COD model, an R^2 of 0.78 was obtained, indicating that the model could justify almost 77.47% of COD variability. A RMSE of 1.34 is used as an average

comprehension between the estimated and actual values of COD. A P-value below 0.001 for this model also adds strength to its predictivity (ref. **Fig. 4e**).

The models exhibited statistically significant correlations (p-value < 0.001) between the predictor variables and the corresponding water quality parameters; however, the R^2 values suggest a moderate degree of predictive accuracy. The validation outcomes of the Landsat-derived SWR models for Khadakwasala Reservoir demonstrate a moderate degree of precision in assessing water quality parameters. The moderate R^2 for DO and BOD (≈0.60–0.65) reflects on certain limitations induced in predicting non-optical water quality parameters from optically sensitive parameters. In this study, turbidity was the only retained predictor for both DO and BOD, while other external parameters (e.g., temperature, microbial activity, etc.) are not considered in the model which could be one reason for moderate accuracy.

The limited dataset comprising 23 samples (17 for training and 6 for testing), all collected from a single surface location (0.25 m), restricts the ability to map both horizontal and vertical concentration variations Although field work was aligned with satellite passes, short-term variability and residual correction uncertainties (radiometric/atmospheric) introduce additional noise in reflectance–chemistry relationships. These factors together account for the observed R^2 while the fits remain statistically robust; denser, multi-site/depth sampling and added covariates are expected to narrow uncertainties in future work.

Improving model performance can be achieved by rectifying the constraints of the training dataset, integrating supplementary data points, and investigating alternative modeling methodologies. Such enhancements will ultimately support more accurate and reliable estimation of surface water quality parameters using remote sensing data, contributing to better water resource management and environmental monitoring efforts.

The influence of seasonal variability, especially significant events like monsoons, on water quality parameters and subsequently on model performance, is a critical consideration. Seasonal changes, particularly monsoon events, can drastically alter the physical, chemical, and biological characteristics of a water body. Increased rainfall during monsoons typically leads to higher runoff from the surrounding catchment area, introducing larger quantities of suspended sediments, organic matter, and nutrients into the reservoir. This influx can lead to significant increases in turbidity, total suspended solids, and potentially influence chlorophyll concentrations due to nutrient enrichment. The models developed in this study were calibrated on data from a specific period specially during non-monsoon period, and their performance may vary when applied to different seasons with distinct environmental conditions. Models developed using limited, localized datasets tend to lack generalizability across different seasons.

5. CONCLUSIONS

This study successfully integrates In situ water quality measurements have been integrated with Landsat 8/9 Operational Land Imager (OLI) data for assessing water quality in the Khadakwasala Reservoir, Pune, India. Using stepwise regression analysis, predictive models have been developed that relate Landsat-derived spectral indices to key parameters, including turbidity, chlorophyll-a, dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD). The models demonstrated predictive accuracies with R2 values of 0.85 for turbidity, 0.90 for chlorophyll-a, 0.65 for DO, 0.64 for BOD, and 0.84 for COD, confirming the feasibility of using remote sensing for water quality assessment.

These results highlight the utility of satellite imagery, particularly Landsat 8/9 data, for monitoring water quality over large geographic areas, providing an efficient and cost-effective alternative to traditional in situ monitoring methods. This study do not observe the major fluctuation in the water quality in both dataset in-situ and by Landsat. This means the 8 days revisit temporal resolution monitoring frequency of alternate Landsat 8 and 9 does not affect lake and reservoir monitoring. However, it may make the issue of concern in case of rivers. Furthermore, turbidity was found to be a significant predictor for several non-optically sensitive parameters (DO, BOD, COD),

enhancing the robustness of the models. This study provides a scalable method for routine water quality monitoring that can be applied to other water bodies with similar environmental conditions.

The Integrating remote sensing data with physical water quality measurements presents a comprehensive approach to effective water resources management. This approach not only aids in the monitoring and management of water bodies like the Khadakwasala Reservoir but also holds promise for broader applications in environmental monitoring, supporting the sustainable management of water resources across regions. Future research could focus on increasing the accuracy of the models by incorporating additional data sources or exploring advanced machine learning techniques to improve the predictive ability for water quality parameters. Also, we recommend to consider other water quality parameters like temperature, CDOM, Secchi Depth etc. for betterment and improvement of the model.

ANNEX CONTAINING DATA AVAILABILITY

A1) In-situ Water Quality Data

The surface water quality parameters data, including Tur, Chl, DO, BOD, and COD for the Khadakwasala Reservoir in Pune, is available for public access. This dataset can be freely downloaded and used from the Zenodo Open Access Repository at https://zenodo.org/records/10901935 , (Kulkarni & Khare, 2024a ; Kulkarni & Khare, 2024b). The data is hosted at Zenodo and is licensed for open access, ensuring its availability for researchers and practitioners interested in water quality assessment and related studies.

A2) Landsat 8/9 OLI Derived Water Quality Data

The publicly available dataset "Water Quality Parameters for Khadakwasala Dam derived using Landsat 8 OLI Surface Reflectance" (Mendeley Data, v1; DOI: 10.17632/nwp835npgk.1) provides GeoTIFF raster layers for chlorophyll-a, turbidity, COD, BOD, and DO derived from Landsat-8 OLI surface-reflectance bands and band ratios for 20 Oct 2022–22 Apr 2023 (post-monsoon to pre-monsoon period). The accompanying description details the regression equations used to produce the layers (OACs from reflectance; NOACs from turbidity/chlorophyll), the file naming convention (Parameter_YYYYMMDD_Khadakwasala.tiff), and a caution that the relationships are site-specific. The dataset is released under CC BY 4.0, enabling reuse with attribution. All image data generated in this study are stored in the Mendeley Data repository and can be accessed via the following DOI: https://data.mendeley.com/datasets/nwp835npgk/1. This dataset is publicly available without any access restrictions, for distribution and reproduction in any medium, provided the original work is properly cited.

A3) Code availability

The Python code used for the generation and processing of datasets in this study is available at https://github.com/rpkulkarni/stepwise-regression-SWQP-khadakwasala.git. This repository contains the exact version of the code used in the study, along with the relevant documentation. The primary dependencies for this code include Python 3.8, along with libraries such as NumPy (v1.21.0), Pandas (v1.3.0), and Scikit-learn (v0.24.2). The specific variables and parameters used for dataset generation and processing are documented within the repository's README file. Access to the code is unrestricted and available for public use.

ACKNOWLEDGEMENTS

We thank the Water Resources Department, Government of Maharashtra, Pune Division, for granting access to collect the water samples from the Khadakwasala Reservoir and providing assistance for the same. We also thank the Staff of the Civil Engineering Department, Symbiosis Institute of Technology, for supporting us in conducting the experiments in the laboratory.

REFERENCES

Akhtar, N., Ishak, M. I. S., Bhawani, S. A., & Umar, K. (2021). Various natural and anthropogenic factors responsible for water quality degradation: A review. Water, 13(19), 2660. https://doi.org/10.3390/w13192660

Chang, N.-B., Imen, S., & Vannah, B. (2014). Remote sensing for monitoring surface water quality status and ecosystem state in relation to the nutrient cycle: A 40-year perspective. *Critical Reviews in Environmental Science and Technology, 44*(19), 2209-2250. https://doi.org/10.1080/10643389.2013.829981

Chawla, I., Karthikeyan, L., & Mishra, A. K. (2020). A review of remote sensing applications for water security: Quantity, quality, and extremes. *Journal of Hydrology, 585*, 124826. https://doi.org/10.1016/j.jhydrol.2020.124826

Chen, J., Zhang, M., Cui, T., & Wen, Z. (2013). A review of some important technical problems in respect of satellite remote sensing of chlorophyll-a concentration in coastal waters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(5), 2275–2289. https://doi.org/10.1109/JSTARS.2013.2242845

Kirk, J. T. O. (2010). Light and photosynthesis in aquatic ecosystems (3rd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139168212

Kuhwald, K., & Oppelt, N. (2016). Remote sensing for lake research and monitoring – Recent advances. *Ecological Indicators, 64*, 105–122. https://doi.org/10.1016/j.ecolind.2015.12.009

Kulkarni, R., & Khare, K. (2024). Water quality parameters for Khadakwasala Dam derived using Landsat 8 OLI surface reflectance [Data set]. *Mendeley Data, V1*. https://data.mendeley.com/datasets/nwp835npgk/1

Kulkarni, R., & Khare, K. (2024). Surface water quality parameters data of Khadakwasala Reservoir Pune, India [Data set]. *Zenodo*. https://zenodo.org/records/10901935

Li, J., et al. (2018). Spatio-temporal variations of CDOM in shallow inland waters from a semi-analytical inversion of Landsat-8. *Remote Sensing of Environment, 218*, 189–200. https://doi.org/10.1016/j.rse.2018.09.014

Olmanson, L. G., Brezonik, P. L., & Bauer, M. E. (2015). Remote sensing for regional lake water quality assessment: Capabilities and limitations of current and upcoming satellite systems. In *Advances in Watershed Science and Assessment* (Vol. 33, pp. 111–140). Springer. https://doi.org/10.1007/978-3-319-14212-8_5

Palmer, S. C. J., Kutser, T., & Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions. *Remote Sensing of Environment, 157*, 1–8. https://doi.org/10.1016/j.rse.2014.09.021

Sagan, V., et al. (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*, 103187. https://doi.org/10.1016/j.earscirev.2020.103187

Sharaf El Din, E., Zhang, Y., & Anctil, F. (2017). Mapping concentrations of surface water quality parameters using a novel remote sensing and artificial intelligence framework. *International Journal of Remote Sensing, 38*(4), 1023–1042. https://doi.org/10.1080/01431161.2016.1275056

Sun, D., Qiu, Z., Shi, K., & Gao, S. (2014). Detection of total phosphorus concentrations of turbid inland waters using a remote sensing method. *Water, Air, & Soil Pollution, 225*(3), 1953. https://doi.org/10.1007/s11270-014-1953-6

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. *Journal of Cleaner Production, 247*, 119134. https://doi.org/10.1016/j.jclepro.2019.119134

Wang, C., et al. (2017). The spatial and temporal variation of total suspended solid concentration in Pearl River Estuary during 1987–2015 based on remote sensing. *Science of the Total Environment, 618–619*, 1125–1138. https://doi.org/10.1016/j.scitotenv.2017.09.196

Wang, S., et al. (2019). High spatial resolution monitoring of land surface energy, water, and CO_2 fluxes from an unmanned aerial system. *Remote Sensing of Environment, 229*, 14–31. https://doi.org/10.1016/j.rse.2019.03.040

Wang, X., & Yang, W. (2019). Water quality monitoring and evaluation using remote-sensing techniques in China: A systematic review. *Ecosystem Health and Sustainability, 5*(1), 1571443. https://doi.org/10.1080/20964129.2019.1571443