Journal of Sustainable Development of Energy, Water and Environment Systems

http://www.sdewes.org/jsdewes

ISSN 1848-9257
Journal of Sustainable Development of Energy, Water and Environment Systems
https://www.sdewes.org/jsdewes

Year 2025, Volume 13, Issue 2, 1130539

Original Research Article

Turbidity Estimation by Machine Learning Modelling and Remote Sensing Techniques Applied to a Water Treatment Plant

Víctor H. Gauto *[1], *Enid M. Utgés*[1], *Elsa I. Hervot*[1], *María D. Tenev*[1],
Alejandro R. Farías [1]
[1]Research Group on Environmental and Chemical Issues
National Technological University, Resistencia, Argentina
e-mail: victor.gauto@ca.frre.utn.edu.ar

Cite as: Gauto, V., Utges, E., Hervot, E., Tenev, M. D., Farías, A., Turbidity Estimation by Machine Learning Modelling and Remote Sensing Techniques Applied to a Water Treatment Plant, J.sustain. dev. energy water environ. syst., 13(2), 1130539, 2025, DOI: https://doi.org/10.13044/j.sdewes.d13.0539

ABSTRACT

Clean water is a scarce resource, fundamental for human development and well-being. Remote sensing techniques are used to monitor and retrieve quality estimators from water bodies. In situ sampling is an essential and labour-intensive task with high costs. As an alternative, a large water quality dataset from a potabilization plant can be beneficial to this step. Combining laboratory measurements, given by a water treatment plant in North-East Argentina, and spectral data from Sentinel-2 satellite platform, several algorithms were proposed, trained, and compared for turbidity estimation at the plant inlet water, in a local river. The highest performance metrics were from a random forest model with a coefficient of determination close to unit (0.913) and the lowest root-mean squared error (143.9 nephelometric turbidity units). The most influential spectral bands were identified by global feature importance and partial dependencies profiles techniques. Maps and histograms were made to explore the turbidity spatial distribution.

KEYWORDS

Random forest, Remote sensing, Sentinel-2, Turbidity, Water quality.

INTRODUCTION

Ensure water availability is one of the Objectives of 2030 Agenda for Sustainable Development by UN [1]. To achieve this, satellite remote sensing techniques can be applied to study and monitoring of water bodies, since it's possible to retrieve spectral data from large regions of the Earth surface. Applied remote sensing can be used to estimate biophysical water parameters, such as total suspended matter [2], chlorophyl-a [3], Secchi disc depth [4] and turbidity [5]. These regression models, the algorithms, can be relatively simple mathematical expressions [6] or more complex approaches, like in machine learning methods [7], which requires tunning of model specific parameters. Remote sensing techniques can be applied for research in a wide range of environmental topics, such as land pollution [8] and glacier retreat [9], among others.

Sentinel-2 (S2) is a spatial mission developed and operated by the European Space Agency (ESA), consisting of two platforms: S2A and S2B, launched in 2015 and 2017, respectively. The MultiSpectral Instrument (MSI) is the optical sensor mounted in S2, with 10 m of maximum spatial resolution, spectral range of 440 nm to 2200 nm, and 5 days of revisit time

[*] Corresponding author

for the constellation. S2 database is free and open access, available from the Copernicus Open Access Hub. S2-MSI has been used in water monitoring and parameter estimation of physicochemical properties such as colour of water [10], chlorophyll-a concentration [11], coloured dissolved organic matter (CDOM) [12], turbidity by band ratios [13] and to detect microplastic pollution [14]. The generated products from S2-MSI are reliable [15] due to a low radiometric uncertainty [16].

The region of interest is in North-East Argentina, in Chaco Province. This area presents several studies regarding fires [17], floods [18], vegetation cover [19] and biodiversity [20]. Nevertheless, water quality studies with a remote sensing approach are scarce. Paraná river is studied by satellite spectral data, but mainly in the North (inside Brazil borders) or South (Argentina middle region) basins.

Machine learning techniques can find complex relationships between data [21]. The combination of machine learning and remote sensing data is a valuable tool to retrieve water quality indicators and its spatiotemporal distribution [22]. Considering this method, land use classification and its influence in sub-watershed level was obtained by Sentinel-2 imagery and cellular automata Markov chains [23]; river water quality models were developed by MODIS (Moderate Resolution Imaging Spectroradiometer) and long-short term memory network [24]. The advancements in algorithm development, data availability and sensors systems made machine learning popular in water quality estimation, outperforming many other methods [25].

Turbidity is a water property caused by suspended matter producing light scattering, affecting its clarity and colour [26]. This property is a main parameter to define drinking water quality and can alter the water treatment plant functioning, since high values can block microbe disinfection, altering this step. In the overall potabilization process, chemical addition, settling and coagulation are included in the workflow to reduce turbidity and remove sediments [26].

Water treatment plants remove pollutants from raw water to obtained clean potable water to be consumed by regular population. Water turbidity is a sensitive parameter since the plant operation can be stalled if high values are reached. This scenario can put the clean water supply at risk [27].

The treatment plant contacted for the present study needs to adapt its potabilization process to ensure the removal of large amounts of sediments presents in the water. Monitoring and understanding the spatial distribution of water turbidity in the inlet river is a valuable input to the overall system since it can be used to give support in the making decision process.

Remote sensing procedures require regular in situ water sampling to correlate spectral data with physicochemical data. To collect said samples is labour intensive, costly and time consuming [28]. Fiel sampling errors can alter the accuracy and precision of data [29]. An alternative is buoys installation that usually are located in a single site in a water body, with the corresponding maintenance. The internal sensors in buoys require frequent calibration due to accuracy loss and constant cleaning [30]. Anti-vandalism measures are desired to prevent equipment damage. Efforts had been made to develop [31] and deploy [32] low-cost buoys in marine environments. An optimized system design is fundamental to decrease production, operation and maintenance costs [33].

Treatment plants laboratories regularly measure water properties, as part of the usual operation process. These datasets are a valuable tool to complement remote sensing techniques, replacing in situ sampling, as water parameter source for the algorithm development. Potabilization plants databases collect historical measurements, often several times a day, that can be applied to spectral imagery collections to elaborate regression models for water quality estimations. A traditional water sampling program for a large time scale monitoring would represent monetary, time and logistic challenges [34].

Remote sensing has been incorporated into water monitoring in a treatment plant [35], calibrating traditional bands ratio regression models to estimate chlorophyll-a and turbidity, using laboratory sampling data from the plant operation.

In this study, daily water turbidity values were given by the MAGR water plant, replacing conventional in situ water sampling. Using S2-MSI images, processing level L2A, surface reflectance (R_S) was extracted for the water inlet location, at surface level. A database of spectral values and turbidity measurements was built to train several regression models, including traditional single band models, and a sophisticated and advanced machine learning approach by a random forest (RF) algorithm. The model with the best performance metrics was selected and turbidity maps and histograms were made for further study its spatial distribution.

To understand the spectral bands effect in the whole model, two techniques were applied: global feature importance and partial dependencies profiles. The most influential spectral bands were compared with the results from different authors to support the found model.

Water characterization was performed, and several factors are discussed to incorporate context to the obtained results.

MATERIALS AND METHODS

The area of study is described, mentioning the main rivers in the region. Remote sensing and laboratory data, their characteristics, and the mathematical model methodology are included in this section.

Area of study

The Paraná River is the second longest river in South America, running through 4000 km [36]. In Argentina, is the natural boundary of multiple provinces, reaching the Río de la Plata, into its exit in the Atlantic Ocean. Paraguay River, with 2550 km [36], is a tributary of Paraná River in its middle basin. The Bermejo River, an Andean tributary [37], is the main sediment source in the Paraná-Paraguay confluence. Due to the high solids presence in Paraguay River, the discharge made into Paraná River alters the characteristics of its composition, creating two distinct regions of high (West) and low (East) sediment concentration [38].

The Metropolitan Area of Gran Resistencia (MAGR) is an urban region in Chaco Province, North-East Argentina. It's composed of four cities, including Resistencia, the capital city of Chaco. MAGR has a population of 423000 inhabitants, according to the last census [39]. Paraná River has a large impact in its society: fishing industry, tourism, recreational activities of the local communities, and transportation route [40]. The water source for the MAGR potabilization plant is located in an arm of Paraná River, Barranqueras River, which is connected to two main rivers in the metropolitan area, Black and Tragadero Rivers.

A map of the region of interest is shown in Figure 1. The inset image corresponds to Argentina, with Chaco province (pink), MAGR location (white dot) and Paraná River extension (blue line). The main image is a real colour satellite view of the study area, with the potabilization treatment plant (yellow triangle) in Barranqueras city and main rivers.

The sample point (red star), located in 58°54'23"W 27°28'20"S, was selected over the Barranqueras River, at the inlet position.

From the inlet point, the water is pumped into a chamber from which its distributed to the different plant sections. In this chamber, samples are collected and delivered to the in-site laboratory to measure a series of parameters, mainly turbidity, pH, electrical conductivity, and alkalinity.

Laboratory data

Daily measurements were given by the in-site laboratory at the water treatment plant [41], located in Barranqueras city, from 2017-01-01 to 2021-09-03. In this time span, 1732 observations were recorded. The parameters and their units were: pH; electrical conductivity, in micro siemens per centimetre [µS/cm], alkalinity, as parts per million of calcium carbonate [ppm $CaCO_3$]; and turbidity, in nephelometric turbidity units [NTU]. Alongside these data, supplementary water samples were taken to assess more sediments related parameters, such as

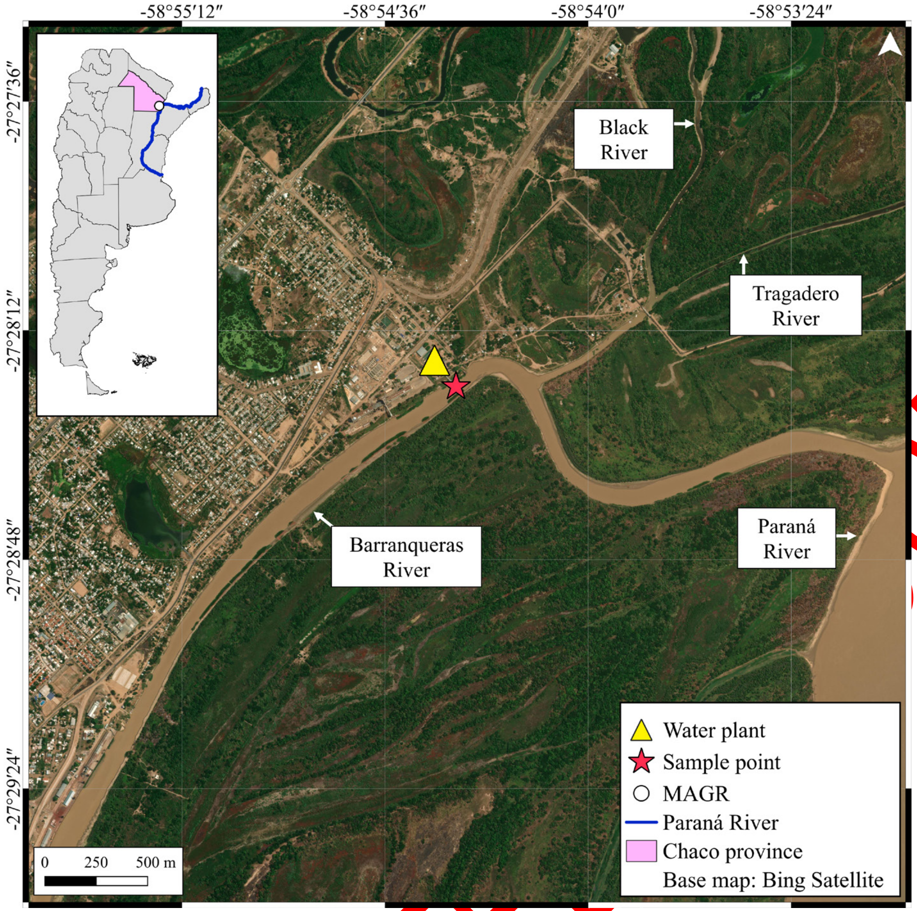

Figure 1. Region of study, indicating main rivers, water plant location and sample site. Inset: Chaco Province relative to Argentina, MAGR location and Paraná River extension.

total suspended matter (TSM), total dissolved matter (TDM), and total matter (TM). These parameters, measured in parts per million [ppm], are related since TM is the sum of TSM and TDM. In total, 28 complementary samples were collected, from 2021-08-24 to 2022-12-07. The sampling was made at the distribution chamber, collecting 1 litter of water, in a dark glass bottle.

The physicochemical methods applied to measure pH, conductivity, alkalinity, turbidity, TSM and TM were 4500 H^+, 2510-B, 2320-B, 2130 B, 2540 D and 2540 B, respectively, according to Standard Methods techniques [42]. TDM was calculated as the difference between TM and TSM.

Remote sensing data

Satellite spectral R_S data was obtained from S2-MSI. Table 1 resumes the characteristics of both sensors since products of platform S2A and S2B were used. Maximum spatial resolution of 10 m (when available), 5 days revisit time, and 11 spectral bands were used. Bands B09 and B10, at 945 nm and 1373 nm, respectively, were discarded since no surface measurement is done at those wavelengths.

Copernicus Data Space Ecosystem provides complete, open, and free access to S2 products. For the same period, 382 images were acquired. S2-MSI, processing level L2A, are atmospherically corrected by Sen2Cor processor [43]. Using the quality assessment band, QA60, the images with present clouds were discarded. This simple method was preferred over more complex approaches [44], since QA60 band is a coded bit mask detecting clear sky, dark clouds, and cirrus clouds. This band was acquired from the Sentinel-2 dataset in Google Earth Engine platform [45].

Table 1. Sentinel-2 spatial and spectral resolutions of platforms S2A and S2B.

		S2A		S2B	
Band	Spatial resolution [m]	Central wavelength [nm]	Bandwidth [nm]	Central wavelength [nm]	Bandwidth [nm]
1	60	442.7	21	442.3	21
2	10	492.4	66	492.1	66
3	10	559.8	36	559.0	36
4	10	664.6	31	665.0	31
5	20	704.1	15	703.8	16
6	20	740.5	15	739.1	15
7	20	782.8	20	779.7	20
8	10	832.8	106	833.0	106
8A	20	864.7	21	864.0	22
11	20	1613.7	91	1610.4	94
12	20	2202.4	175	2185.7	185

After removing the dates with clouds over the study area, 181 products remained to continue the analysis.

The satellite products were cropped around the area of interest, then resampled to the same spatial resolution of 10 m. The R_S values extraction was made using a 3x3 pixel window around the point near the plant water entrance, on Barranqueras River (Figure 1). The final pixel value was the mean of the individual values in the grid.

Modelling

As a preliminary step, the relationship between turbidity and R_S per band was studied. Thus, the potential impact of individual bands in the turbidity value was evaluated.

The target parameter in the modelling process was turbidity as a mathematical regression problem, with the spectral bands as predictors.

Two main modelling methods were used: linear, with algebraic relationships between the predictors, and a based-tree machine learning RF approach. To perform the linear modelling several spectral bands and normalized difference turbidity index (NDTI) were used. NDTI was obtained by the red and green bands, B04 and B03, respectively [46]. This index was used for water quality assessment and it's proportional to turbidity [47]. The expression is shown in Eq. (1).

$$NDTI = \frac{B04 - B03}{B04 + B03} \tag{1}$$

Machine learning techniques were applied to improve traditional methods for parameter retrieval [48]. RF operates by an ensemble of decision trees [49], each one trained by a subset of the whole data. RF can manage many predictor variables and maintain low levels of over-fitting [50], a negative aspect in modelling.

RF modelling used all spectral bands available (Table 1), since this method is appropriated to find non-linear relationships between multiple predictors. To improve the performance of RF, a tuning step was applied to obtain the best arguments, the hyperparameters, required in this model. The tuned hyperparameters were the minimum number of samples taken from the dataset to form a node in a decision tree (min_n) and the number of predictors that will be sampled (mtry). The 'trees' hyperparameter was fixed at 1000 units. Both steps of sample observations and predictors selection are random, across all trees. The final turbidity estimation was an average value of all tree's individual estimations.

The tuning process applied was made by the racing technique [51]. This technique evaluated the model in a subset of resamples obtaining the performance metrics, continuing only with the hyperparameters that showed good results. Usually, racing techniques are faster to compute than traditional methods, such as grid search [52].

To measure the model's performance the metrics calculated were: Pearson's coefficient of determination (R^2) and root mean squared error (RMSE). The following equations show the mathematical expressions for these metrics:

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{3}$$

Where i represents each measurement, from a total of n samples. y_i are the real turbidity values, and x_i the estimated values from the correspondent model. For R^2, Eq. (2), $\bar{y}_i$ is the mean value of y_i. A preferred model consists of a high value of R^2, closer to 1, and low RSME.

To produce a model, linear or RF, the dataset was split in two parts: 75% of the samples was used for training and tuning of the corresponding model; the remaining 25% was used only for testing and to finalizing the model, that is, to get the last version of the model specification. This methodology followed the best practices for data modelling [53].Since the original dataset corresponded to a time series of turbidity values, the testing split corresponded to the most recent dates. The training dataset was resampled by a 10-fold cross-validation

The performance metrics were calculated in the training step only for the selection of the best model. Once the selection was made, the model was evaluated using the testing dataset, with new and later observations, to calculate the final performance metrics. The estimated turbidity values were compared with the validation values in a time series plot.

Several maps showcasing the turbidity spatial distribution were made applying the selected model to S2-MSI products, in the Barranqueras River, for four different dates. The spectral index MNDWI (modified normalized difference water index) was used to mask the water from the scene [54]. An automatic method was used to identify the MNDWI threshold value [55].

RESULTS AND DISCUSSION

Water characterization results are summarized and discussed as a parameter time series. Anthropogenic and environmental factors are mentioned to explain water quality. Model selection and hyperparameters tuning are described. To evaluate spectral bands effect, two techniques were applied to the best model. Water turbidity spatial distribution was assessed by maps and histograms.

Water characterization

The parameters measured by the water treatment plant are shown in Figure 2 as a time series plot. The number of samples (n) is shown in the top right corner of each panel.

In 2019, Paraná River scarce rains and drought caused an historic low level of water [56]. This can be seen in the steady increase in turbidity (Figure 2a) and conductivity (Figure 2d). Conductivity from 2019 and forward started to be more disperse than previous years. Turbidity presented yearly cycles, with high values at the beginning of each year, between January and April-May, then followed by a low-turbidity period.

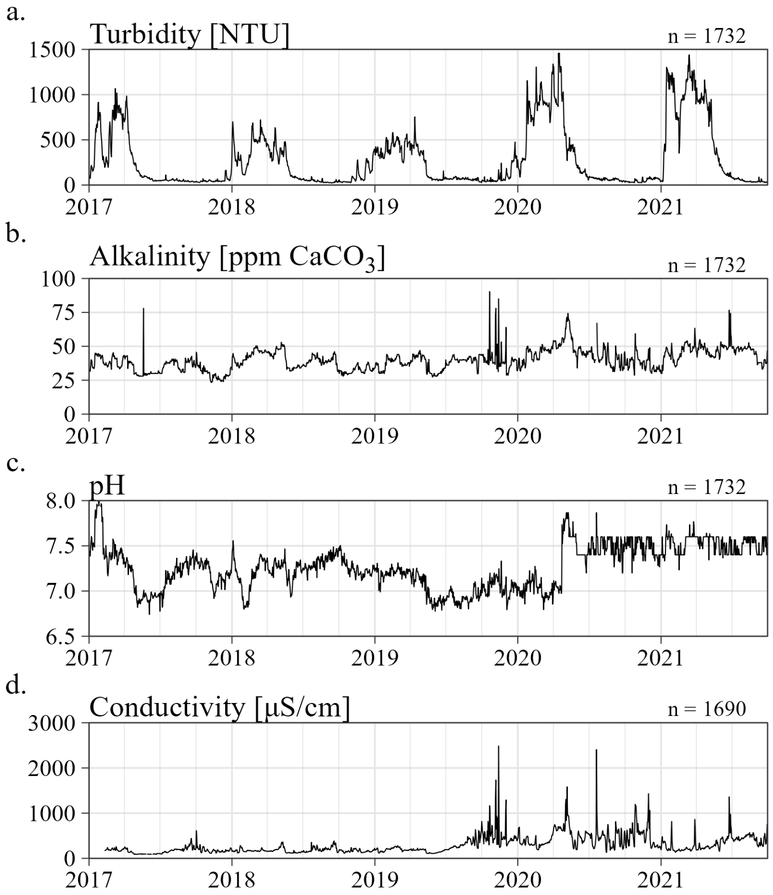

Figure 2. Time series of measured water parameters: turbidity, alkalinity, pH, and conductivity.

Main statistical values per parameter are summarised in Table 2. Mean, median, standard deviation (SD), initial and final sampling date and number of samples (n).

Water properties are heavily influenced by dams' operations [57] in the North basin (south Brazil) being Itaipú dam (Paraguay-Brazil) and Yaciretá reservoir (Paraguay-Argentina), the closest to the study area.

The main source of sediments in Paraná river is due Bermejo River, a Paraguay River tributary, creating a high turbidity imbalance [58]. Due to heavy rains in Bermejo River headwaters, between October and April, a high sediment concentration is reached in Paraná River between December and May [57]. Said period corresponded with the turbidity cycles observed in Figure 2.a.

In a regional scope, Black River is in a meander area with low surface slope causing a low soil erosion that carries sediments to Tragadero River (Figure 1). MAGR has flood risk and intense rains (1500 mm a year [59]) can cause hydric emergency [60], altering the water properties in the treatment plant inlet and increasing Paraná River flow rate.

Table 2. Statistical summary of measured water properties.

Parameter	Mean	Median	SD	Initial date	Last date	n
Turbidity [NTU]	280.7	89.8	328.2	2017-01-01	2021-09-30	1732
Alkalinity [ppm $CaCO_3$]	40.0	40.0	7.9	2017-01-01	2021-09-30	1732
pH	7.3	7.3	0.2	2017-01-01	2021-09-30	1732
Conductivity [µS/cm]	293.6	220.2	205.6	2017-02-10	2021-09-30	1690
TSM [ppm]	84.2	30.0	171.6	2021-08-24	2022-12-07	25
TDM [ppm]	195.9	166.8	100.8	2021-08-24	2022-12-07	26

TM [ppm]	281.9	209.5	212.1	2021-08-24	2022-12-07	25

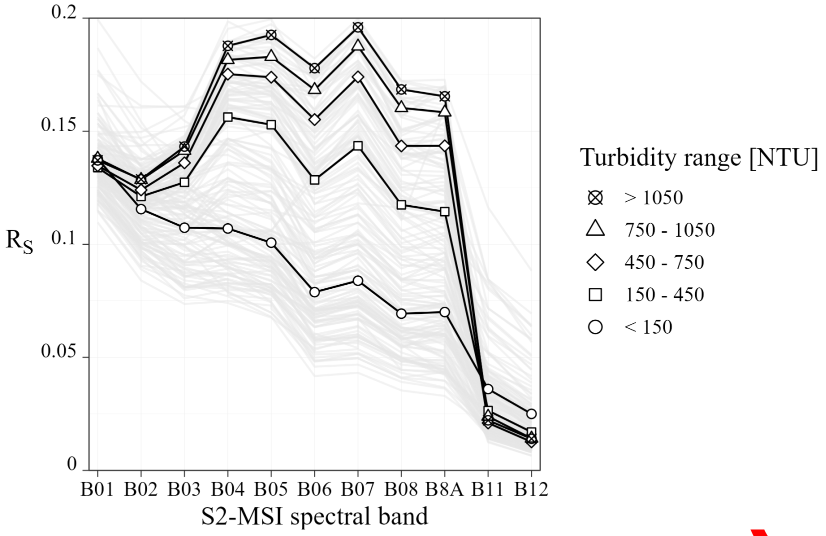

Figure 3. Mean spectral signatures per turbidity range, for S2-MSI bands.

The highest water flow in Paraná River occurs between February and March, with values over 30000 m^3/s, with a mean of 17000 m^3/s [61]. The change in water flow, land use and hydroelectric development (by dam constructions) alters the hydrologic characteristics of Paraná waters, thus affecting water treatment operation in the last decades [61].

To estimate turbidity from R_S was necessary to inspect the relationship between them. Figure 3 shows the spectral signatures of all the observations in light grey lines. Grouping the data observations by turbidity ranges, mean spectral signatures were obtained, in black lines. Lower turbidity (<150 NTU) presented the lowest R_S. As the turbidity range increased, the spectral signature response raised until values higher than 1050 NTU.

Bands B01, B11 and B12 (Figure 3) are not sensitive to turbidity change, since the points for different turbidity ranges remained in the same position. Bands B05, B06 and B07 presented the highest changes. These bands were related to algorithms for turbidity estimation [5].

Model selection

Several models were tested to estimate the inlet water turbidity for the treatment plant. The predictors variables were selected according to the model. For RF, S2-MSI bands shown in Table 1 were used as predictors. Traditional linear models were produced by the following variables: an interaction between B06 and B07; individual bands B05, B06 and B08; and the spectral index NDTI. The aforementioned spectral bands were selected according to the results from Figure 3.

The characteristics and performance metrics for all proposed models are resumed in Table 3. These metrics were used to select the model, and, in a later step, the model is finalized using the preserved testing dataset, with observations not used in the training. The best results were achieved by the RF model.

R^2 for individual bands B05, B06 and B08, presented 0.693, 0.732 and 0.736, respectively. In a similar work [5], also in turbid lakes, R^2 for the same bands were 0.83, 0.66 and 0.63, respectively.

RF model was the selected model in the following analyses, since it presented the best performance metrics, with the lowest deviations (RMSE) and highest correlation (R^2). This machine learning algorithm could capture the complex relationships between satellite spectral data and turbidity values, more than the proposed usual regression models. A RF model with proper modifications performed better than others machine learning options for turbidity estimation [62].

The interaction model between B06 and B07 was the second-best model, combining bands in the red edge, related to sediments in water [63]. In comparison, using single band linear performed poorly. The NDTI index presented the lowest R^2 and the highest RMSE, unlike other studies [64].

The tuned hyperparameters values and the main characteristics of the final RF model are shown in Table 4.

Table 4. RF model main characteristics and hyperparameters.

RF type	Regression
Training observations	116
Variables	11
Trees	1000
min_n	14
mtry	2

After the model selection the last fitting was performed. For the RF model, the final performance metrics were obtained by the testing dataset. These observations were kept apart so they have no influence on the modelling. The metrics were RMSE = 143.9 NTU and R^2 = 0.913, with 39 data points. Noted that these values are different from Table 3, since those were obtained from the training data set, and are used only for model selection.

Table 3. Regression models candidates and training performance metrics

Model characteristics		Performance metrics	
Specification	Expression	RMSE [NTU]	R^2
RF	Turbidity ~ all bands	111.5	0.841
Linear model	Turbidity ~ B06 + B07 + B06×B07	121.9	0.802
Linear model	Turbidity ~ B08	142.9	0.736
Linear model	Turbidity ~ B06	145.7	0.732
Linear model	Turbidity ~ B05	155.8	0.693
Linear model	Turbidity ~ NDTI	218.7	0.296

The comparison between measured and estimated observations in the testing split is shown in Figure 4

The solid line represents the linear relationship between estimated and measured turbidity, with a dashed line at 45°, for comparison reasons. Lower estimated turbidity values are closer to the real values. For higher turbidity the deviations increased, with estimates being lower than measured values, with the solid line below the dashed line. The outcome variable presented a wide range, with many observations under 100 NTU, and measurements as high as 1100 NTU.

The measured and estimated turbidity values, for the validation dataset (Figure 4), are shown as a time series plot in Figure 5. The crosses represented the estimations made by the RF model; the turbidity measures were plotted as a solid line. The number of samples in the testing dataset is shown in the top right corner.

The biggest differences between estimated and measured turbidity in Figure 5 are within the larger values, equivalent to Figure 4. The estimations followed the same trend seen in the turbidity time series in Figure 2, with high turbidity in the beginning of the year, continued by lower values in the middle and end year.

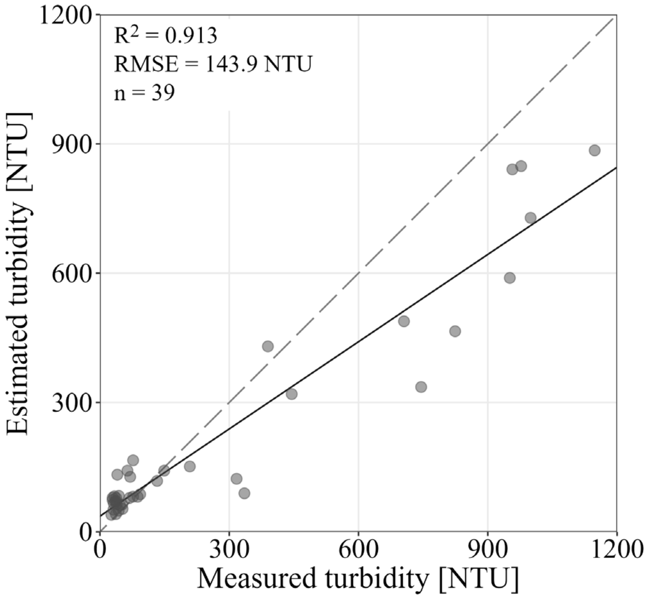

Figure 4. Estimated and measured turbidity using the validation dataset.

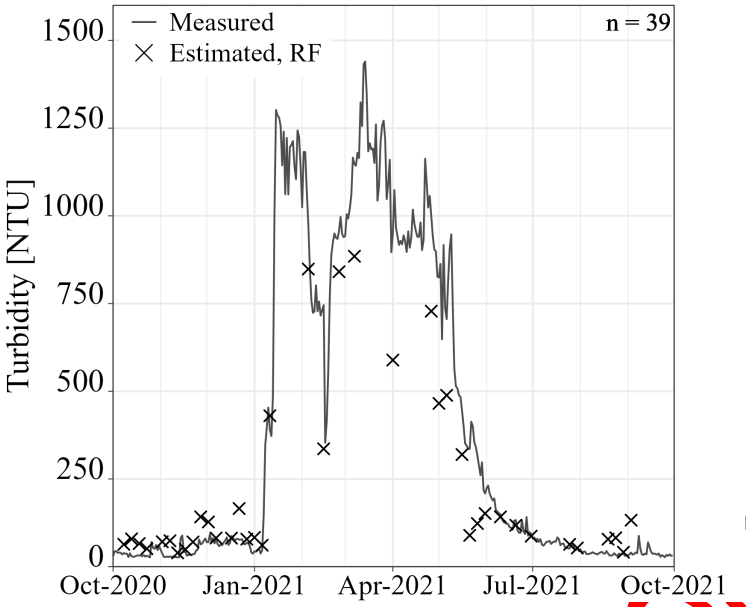

Figure 5. Time series of estimated and measured turbidity in validation dataset.

Understanding the random forest model

The complexity of a RF model is difficult to explain since the explicit form is not as clean as a simpler model, i.e., linear model. The global feature importance is an explanatory technique that assists in understanding the driving predictors of a RF, aggregated in the entire training observations.

The results of the global feature importance for the obtained RF model, analysing each spectral band, is shown in Figure 6. This technique is based on the notion of the overall change in the model due to the perturbation of a specific variable [65]. A permutation-based approach is a valuable tool for model explanation, since it is expected that after the permutation of said variable the model performance will decrease [49]. Spectral band B07 presents the most effect in the model, according to Figure 6, since the boxplot had the highest RMSE (104.9 NTU). Close to B07 was B06, B08 and B05. Spectral bands B05 [66] and B08 [13] have been reported to be related to turbidity. The least effects were given by B01, B02 and B03, since the perturbation of these bands had a much lesser impact on the overall model. The vertical dashed line represents the base RMSE.

The most influential bands were in the range of 704 nm to 830 nm, with the least influential between 440 nm and 500 nm. A similar result, was found applying the same method in the North Tyrrhenian Sea [67], but the importance order was B05 (the highest), followed by B07 and B06. Turbidity estimations were most successful when the wavelength was between 700 nm and 800 nm for surface water [68]; this range including B05, B06 and B07.

Partial dependencies profiles allowed to show the change in the expected value of a model estimate alongside a single explanatory variable [65]. According to the global feature importance technique, B07 was the spectral band that had the highest effect on the RF model. For this band the partial dependencies profile was obtained and is shown in Figure 7a.

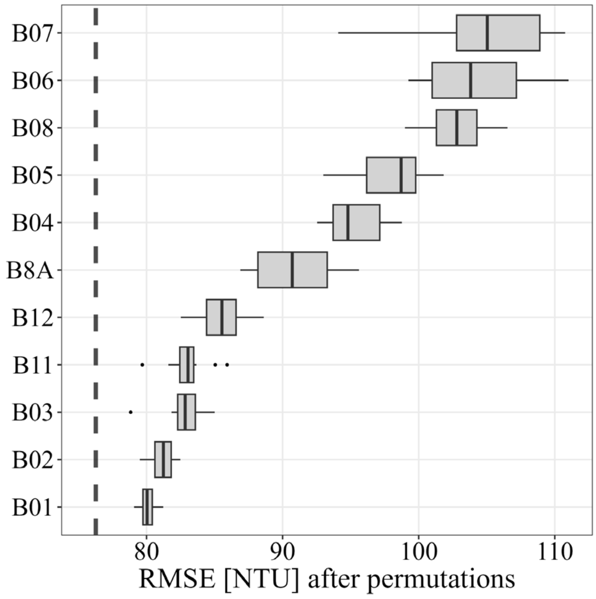

Figure 6. Global feature importance of S2-MSI spectral bands in RF model.

The thin grey lines in Figure 7 correspond to 100 randomly selected observations from the training dataset. The black line indicates the mean. The effect of B07 (Figure 7a) on turbidity estimates was constant until $R_S = 0.12$, then started to increase until its highest effect at $R_S = 0.2$. In this range of surface reflectance, the change in turbidity was from 218.7 NTU to 353.8 NTU. For comparison, Figure 7b corresponds to an identical analysis for B01, the band with the lowest feature importance as mentioned previously (Figure 6). The partial dependency profile of this band was constant, that is, the turbidity presented no change in the entire range of R_S from B01 values. This result was consistent with Figure 3 and Figure 6.

Turbidity spatial distribution

The obtained RF model was applied to the spectral values from Barranqueras River to evaluate the spatial turbidity distribution. Maps for four different dates from 2020 are shown in Figure 8.

The yellow triangle on the top-centre of each panel represents the water plant location. A water mask was applied to the region of interest to only extract pixel values from the Barranqueras River.

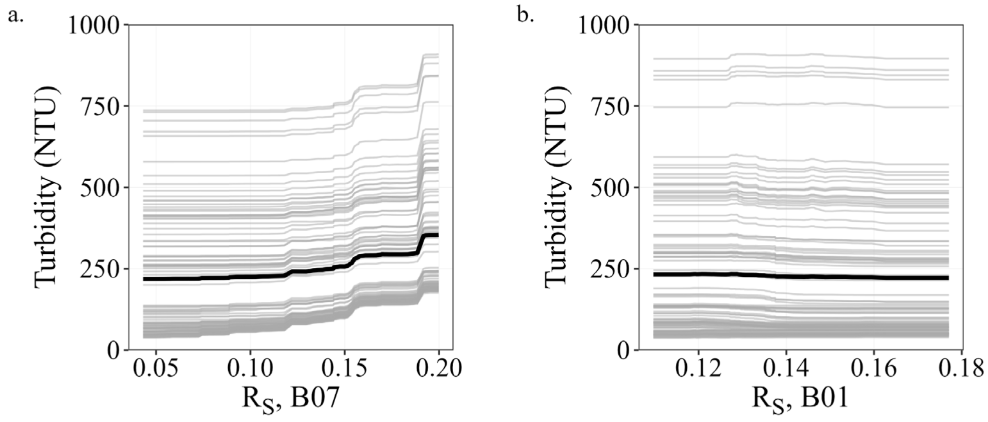

Figure 7. Partial dependencies profiles for spectral bands B07 and B01.

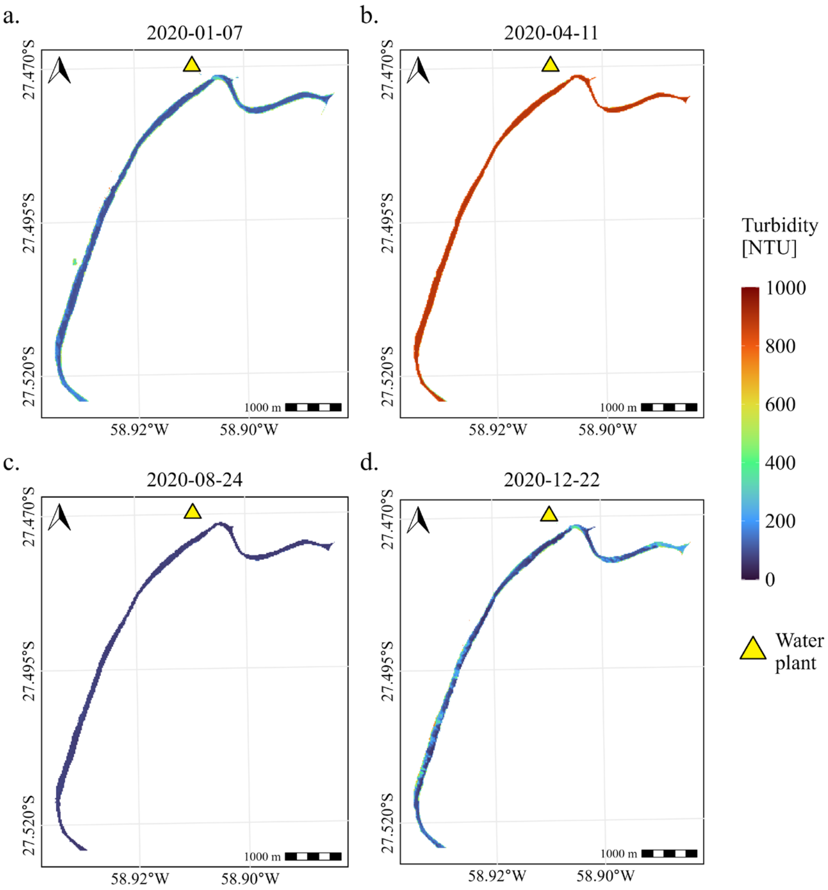

Figure 8. Barranqueras River turbidity maps for four different dates.

Figure 8a (2020-01-07) and Figure 8d (2020-12-22) presented relatively low turbidity, with a wide dispersion. Figure 8b (2020-04-11) and Figure 8c (2020-08-24) presented a narrower turbidity dispersion, thus the colour homogeneity. For the former date, high values were estimated; for the latter, low turbidity values were obtained.

For a better understanding of the turbidity spatial distribution, histograms were plotted, as seen in Figure 9, to showcase the estimations dispersion alongside Barranqueras River. The bin width was set in 10 units.

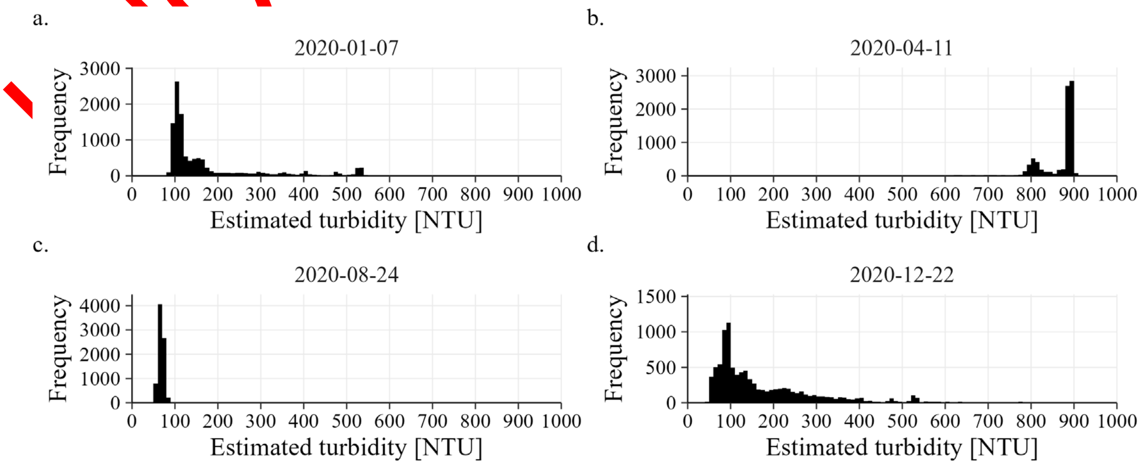

Figure 9. Turbidity histograms per selected dates, as seen in Figure 8.

Figure 9a (2020-01-07) presented relatively lower values, with a wide dispersion. The median turbidity was 117 NTU. High values and a narrow dispersion were observed in Figure 9b (2020-04-11), with an 889 NTU median. The lowest turbidity distribution was obtained in Figure 9c (2020-08-24), presenting a single peak at 69 NTU. Finally, in Figure 9d (2020-12-22), the values increased until a 128 NTU median and a wider dispersion.

At extreme values, the turbidity dispersion was low, as seen in Figure 9b y Figure 9c.

The estimations observed in Figure 8 maps and Figure 9 histograms followed the same measured turbidity trends as Figure 2a for the year 2020.

CONCLUSION

Inlet water properties are an important input in a water treatment plant, to set the filtration operation and the reagents needed for the flocculation step. Water turbidity is a valuable parameter in the decision-making process.

In Resistencia, Chaco province in Argentina, the inlet water turbidity of the local water treatment plant was studied as a time series. Annual turbidity cycles were observed, with high values between January to April-May, and lower values for the rest of the year.

Anthropogenic and environmental factors are discussed as reasons for water quality, mainly dam operation, floods, rain and tributaries in Paraná River. This complex hydrological system modifies water parameters, affecting treatment plant management.

Several linear and machine learning models were tested for turbidity estimation, with the spectral response at different bands as predictors. A tuned RF model outperformed the proposed traditional linear models presenting the highest performance metrics, with $R^2 = 0.913$ and RMSE = 143.9 NTU. The machine learning method allowed to create a sophisticated model to obtain an accurate turbidity estimation from S2-MSI spectral data. The largest turbidity values presented the biggest differences between measured and estimated turbidity. Applying global feature importance technique to the RF model, band B07 (780 nm) was established as the most important variable, followed by B06 (740 nm). The partial dependence profile for B07 indicated the highest change in the outcome variable.

The maps generated from the RF applied to S2-MSI products follow the same trend as the observed turbidity for the same period. Extreme turbidity values presented low dispersion, according to the histograms.

Using water treatment plant laboratory data, replacing traditional in situ water sampling, with remote sensing techniques and combined with machine learning modelling, allowed the development of a validated random forest model with high performance metrics. Turbidity estimation by this study was a relevant contribution in the vital process of water potabilization, in a region with scarce studies regarding satellite data and water quality.

NOMENCLATURE

B1	band 1	[-]
B10	band 10	[-]
B11	band 11	[-]
B12	band 12	[-]
B2	band 2	[-]
B3	band 3	[-]
B4	band 4	[-]
B5	band 5	[-]
B6	band 6	[-]
B7	band 7	[-]
B8	band 8	[-]
B9	band 9	[-]

min_n	minimum number of samples	[-]
mtry	number of predictors	[-]
n	number of samples	[-]
R^2	Pearson's coefficient of determination	[-]
R_S	reflectance remote sensing	[-]
x	turbidity estimated value	NTU
y	real turbidity value	NTU
$\bar{y}$	mean turbidity value	NTU

Subscripts

i	measurement	[-]

Abbreviations

CDOM	Coloured Dissolved Organic Matter	
ESA	European Space Agency	
MAGR	Metropolitan Area of Gran Resistencia	
MNDWI	Modified Normalized Difference Water Index	
MODIS	Moderate Resolution Imaging Spectroradiometer	
MSI	MultiSpectral Instrument	
NDTI	Normalized Difference Turbidity Index	
NTU	Nephelometric Turbidity Units	
pH	Hydrogen Potential	
ppm	Particles Per Million	
QA60	Quality Assurance 60	
RF	Random Forest	
RMSE	Root Mean Squared Error	
S2	Sentinel-2	
S2A	Sentinel-2 platform A	
S2B	Sentinel-2 platform B	
TDM	Total Dissolved Matter	[ppm]
TM	Total Matter	[ppm]
TSM	Total Suspended Matter	[ppm]
UN	United Nations	

REFERENCES

1. United Nations, "2030 Agenda for Sustainable Development," vol. 11371, no. July, pp. 1–13, 2017, https://doi.org/10.1109/TNSRE.2015.2480755.
2. Y. Du *et al.*, "Total suspended solids characterization and management implications for lakes in East China," *Science of the Total Environment*, vol. 806, Feb. 2022, https://doi.org/10.1016/j.scitotenv.2021.151374.
3. W. G. Buma and S. Il Lee, "Evaluation of Sentinel-2 and Landsat 8 images for estimating Chlorophyll-a concentrations in Lake Chad, Africa," *Remote Sens (Basel)*, vol. 12, no. 15, Aug. 2020, https://doi.org/10.3390/RS12152437.
4. G. Rodrigues *et al.*, "Temporal and Spatial Variations of Secchi Depth and Diffuse Attenuation Coefficient from Sentinel-2 MSI over a Large Reservoir," *Remote Sens (Basel)*, vol. 12, no. 5, p. 768, Feb. 2020, https://doi.org/10.3390/rs12050768.
5. Y. Ma *et al.*, "Remote sensing of turbidity for lakes in Northeast China using sentinel-2 images with machine learning algorithms," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 9132–9146, 2021, https://doi.org/10.1109/JSTARS.2021.3109292.

6. M. Pereira-Sandoval *et al.*, "Evaluation of atmospheric correction algorithms over spanish inland waters for sentinel-2 multi spectral imagery data," *Remote Sens (Basel)*, vol. 11, no. 12, pp. 1–23, 2019, https://doi.org/10.3390/rs11121469.
7. Z. Cao *et al.*, "A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes," *Remote Sens Environ*, vol. 248, no. July, p. 111974, Oct. 2020, https://doi.org/10.1016/j.rse.2020.111974.
8. I. Barut, H. Keskin-Citiroglu, M. Oruc, and A. M. Marangoz, "Determination by Landsat Satellite Imagery to Local Scales in Land and Pollution Monitoring: a Case of Buyuk Melen Watershed (Turkey)," *Journal of Sustainable Development of Energy, Water and Environment Systems*, vol. 3, no. 4, pp. 389–404, Dec. 2015, https://doi.org/10.13044/j.sdewes.2015.03.0029.
9. C. S. Carrión *et al.*, "Multi-Temporal Analysis of the Glacier Retreat Using Landsat Satellite Images in the Nevado of the Ampay National Sanctuary, Peru," *Journal of Sustainable Development of Energy, Water and Environment Systems*, vol. 10, no. 1, Mar. 2022, https://doi.org/10.13044/j.sdewes.d8.0380.
10. C. Giardino *et al.*, "The Color of Water from Space: A Case Study for Italian Lakes from Sentinel-2," in *Earth Observation and Geospatial Analyses [Working Title]*, IntechOpen, 2019.
11. M. H. Tavares, R. C. Lins, T. Harmel, C. R. Fragoso, J. M. Martínez, and D. Motta-Marques, "Atmospheric and sunglint correction for retrieving chlorophyll-a in a productive tropical estuarine-lagoon system using Sentinel-2 MSI imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, no. April 2020, pp. 215–236, 2021, https://doi.org/10.1016/j.isprsjprs.2021.01.021.
12. K. Toming, T. Kutser, A. Laas, M. Sepp, B. Paavel, and T. Nõges, "First experiences in mapping lakewater quality parameters with sentinel-2 MSI imagery," *Remote Sens (Basel)*, vol. 8, no. 8, pp. 1–14, 2016, https://doi.org/10.3390/rs8080640.
13. N. M. Hussein, M. N. Assaf, and S. S. Abohussein, "Sentinel 2 analysis of turbidity retrieval models in inland water bodies: The case study of Jordanian dams," *Can J Chem Eng*, vol. 101, no. 3, pp. 1171–1184, Mar. 2023, https://doi.org/10.1002/cjce.24526.
14. T. Ali *et al.*, "Evaluating Microplastic Pollution Along the Dubai Coast: An Empirical Model Combining On-Site Sampling and Sentinel-2 Remote Sensing Data," *Journal of Sustainable Development of Energy, Water and Environment Systems*, vol. 12, no. 1, pp. 1–20, Mar. 2024, https://doi.org/10.13044/j.sdewes.d11.0482.
15. D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sensing*, vol. 12, no. 14. MDPI AG, Jul. 01, 2020, https://doi.org/10.3390/rs12142291.
16. J. Gorroño, A. C. Banks, N. P. Fox, and C. Underwood, "Radiometric inter-sensor cross-calibration uncertainty using a traceable high accuracy reference hyperspectral imager," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 393–417, Aug. 2017, https://doi.org/10.1016/j.isprsjprs.2017.07.002.
17. C. M. Di Bella, M. A. Fischer, and E. G. Jobbágy, "Fire patterns in north-eastern Argentina: influences of climate and land use/cover," *Int J Remote Sens*, vol. 32, no. 17, pp. 4961–4971, Sep. 2011, https://doi.org/10.1080/01431161.2010.494167.
18. J. J. Neiff, E. M. Mendiondo, and C. A. Depettris, "ENSO floods on river ecosystems: from catastrophes to myths," in *River flood defence, kassel reports of hydraulic engineering*, vol. 9, Herkules Verlag, 2000.
19. R. Stanimirova, J. Graesser, P. Olofsson, and M. A. Friedl, "Widespread changes in 21st century vegetation cover in Argentina, Paraguay, and Uruguay," *Remote Sens Environ*, vol. 282, p. 113277, Dec. 2022, https://doi.org/10.1016/j.rse.2022.113277.
20. E. M. O. Silveira *et al.*, "Spatio-temporal remotely sensed indices identify hotspots of biodiversity conservation concern," *Remote Sens Environ*, vol. 258, p. 112368, Jun. 2021, https://doi.org/10.1016/j.rse.2021.112368.

21. A. Najah Ahmed *et al.*, “Machine learning methods for better water quality prediction,” *J Hydrol (Amst)*, vol. 578, Nov. 2019, https://doi.org/10.1016/j.jhydrol.2019.124084.
22. N. Wagle, T. D. Acharya, and D. H. Lee, “Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data,” *Sensors and Materials*, vol. 32, no. 11. M Y U Scientific Publishing Division, pp. 3879–3892, Nov. 30, 2020, https://doi.org/10.18494/SAM.2020.2953.
23. A. N. Putra, S. K. Paimin, S. F. Alfaani, I. Nita, S. Arifin, and M. Munir, “A Machine Learning Approach to Estimating Land Use Change and Scenario Influence in Soil Infiltration at the Sub-Watershed Level,” *Journal of Sustainable Development of Energy, Water and Environment Systems*, vol. 12, no. 1, Mar. 2024, https://doi.org/10.13044/J.SDEWES.D11.0477.
24. S. H. Rahat *et al.*, “Remote sensing-enabled machine learning for river water quality modeling under multidimensional uncertainty,” *Science of the Total Environment*, vol. 898, Nov. 2023, https://doi.org/10.1016/j.scitotenv.2023.165504.
25. V. Sagan *et al.*, “Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing,” *Earth Sci Rev*, vol. 205, no. August 2019, p. 103187, 2020, https://doi.org/10.1016/j.earscirev.2020.103187.
26. F. R. Spellman, *Handbook of Water and Wastewater Treatment Plant Operations*. CRC Press, 2004.
27. C. L. Chang and C. S. Liao, “Assessing the risk posed by high-turbidity water to water supplies,” *Environ Monit Assess*, vol. 184, no. 5, pp. 3127–3132, May 2012, https://doi.org/10.1007/s10661-011-2176-6.
28. A. Lausch *et al.*, “Monitoring Water Diversity and Water Quality with Remote Sensing and Traits,” *Remote Sensing*, vol. 16, no. 13. Multidisciplinary Digital Publishing Institute (MDPI), Jul. 01, 2024, https://doi.org/10.3390/rs16132425.
29. F. Bianchi, V. Pineiro, F. Weinstein, R. Terra, and C. Colacce, *Remote Sensing of Water Quality in Laguna del Sauce, Uruguay*. World Bank, 2020.
30. T. M. Balakrishnan Nair *et al.*, “An integrated buoy-satellite based coastal water quality nowcasting system: India’s pioneering efforts towards addressing UN ocean decade challenges,” *J Environ Manage*, vol. 354, Mar. 2024, https://doi.org/10.1016/j.jenvman.2024.120477.
31. S. Sendra, L. Parra, J. Lloret, and J. M. Jiménez, “Oceanographic multisensor buoy based on low cost sensors for posidonia meadows monitoring in mediterranean sea,” *J Sens*, vol. 2015, 2015, https://doi.org/10.1155/2015/920168.
32. C. Albaladejo, F. Soto, R. Torres, P. Sánchez, and J. A. López, “A low-cost sensor buoy system for monitoring shallow marine environments,” *Sensors (Switzerland)*, vol. 12, no. 7, pp. 9613–9634, Jul. 2012, https://doi.org/10.3390/s120709613.
33. A. Shukla, P. S. Matharu, and B. Bhattacharya, “Design and development of a continuous water quality monitoring buoy for health monitoring of river Ganga,” *Engineering Research Express*, vol. 5, no. 4, Dec. 2023, https://doi.org/10.1088/2631-8695/ad0d40.
34. M. Ramadas and A. K. Samantaray, “Applications of Remote Sensing and GIS in Water Quality Monitoring and Remediation: A State-of-the-Art Review,” in *Energy, Environment, and Sustainability*, Springer Nature, 2018, pp. 225–246.
35. J. Lioumbas *et al.*, “Satellite remote sensing to improve source water quality monitoring: A water utility’s perspective,” *Remote Sens Appl*, vol. 32, Nov. 2023, https://doi.org/10.1016/j.rsase.2023.101042.
36. A. A. Bonetto, “1. The Parana River system,” 1986.
37. E. Abrial, R. E. Lorenzón, A. P. Rabuffetti, M. C. M. Blettler, and L. A. Espínola, “Hydroecological implication of long-term flow variations in the middle Paraná river

floodplain,” *J Hydrol (Amst)*, vol. 603, p. 126957, Dec. 2021, https://doi.org/10.1016/j.jhydrol.2021.126957.

38. S. N. Lane, D. R. Parsons, J. L. Best, O. Orfeo, R. A. Kostaschuk, and R. J. Hardy, “Causes of rapid mixing at a junction of two large rivers: Río Paraná and Río Paraguay, Argentina,” *J Geophys Res Earth Surf*, vol. 113, no. 2, Jun. 2008, https://doi.org/10.1029/2006JF000745.
39. Instituto Nacional de Estadística y Censos (Argentina), *National Population, Households and Housing Census 2022. Provisional results, in Spanish*. 2022.
40. J. J. Neiff, A. S. G. Poi De Neiff, and S. L. Casco, “Ecological importance of the Paraguay-Paraná River Corridor as a context for sustainable management, in Spanish,” *Humedales fluviales de América del Sur*, pp. 193–210, 2005.
41. SAMEEP - Departamento de Calidad Laboratorio Central, “Potable Water Quality Control - Procedures Manual, in Spanish.”
42. R. B. Baird, C. E. W. Rice, and A. D. Eaton, *Standard Methods for the Examination of Water and Wastewater, 23rd*, no. 1. Water Environment Federation, American Public Health Association, American Water Works Association, 2017.
43. M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, “Sen2Cor for Sentinel-2,” in *Image and Signal Processing for Remote Sensing XXIII*, Oct. 2017, p. 3, https://doi.org/10.1117/12.2278218.
44. M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, “AGGREGATING CLOUD-FREE SENTINEL-2 IMAGES WITH GOOGLE EARTH ENGINE,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, no. 2/W7, pp. 145–152, Sep. 2019, https://doi.org/10.5194/isprs-annals-IV-2-W7-145-2019.
45. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, “Google Earth Engine: Planetary-scale geospatial analysis for everyone,” *Remote Sens Environ*, vol. 202, no. 2016, pp. 18–27, 2017, https://doi.org/10.1016/j.rse.2017.06.031.
46. J. P. Lacaux, Y. M. Tourre, C. Vignolles, J. A. Ndione, and M. Lafaye, “Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal,” *Remote Sens Environ*, vol. 106, no. 1, pp. 66–74, Jan. 2007, https://doi.org/10.1016/j.rse.2006.07.012.
47. X. Chen, W. Chen, Y. Bai, and X. Wen, “Changes in turbidity and human activities along Haihe River Basin during lockdown of COVID-19 using satellite data,” *Environmental Science and Pollution Research*, vol. 29, no. 3, pp. 3702–3717, Jan. 2022, https://doi.org/10.1007/s11356-021-15928-6.
48. S. Magri, E. Ottaviani, E. Prampolini, B. Federici, G. Besio, and B. Fabiano, “Application of machine learning techniques to derive sea water turbidity from Sentinel-2 imagery,” *Remote Sens Appl*, p. 100951, Apr. 2023, https://doi.org/10.1016/j.rsase.2023.100951.
49. L. Breiman, “Random Forests,” 2001. https://doi.org/https://doi.org/10.1023/A:1010933404324.
50. A. B. Ruescas, M. Hieronymi, G. Mateo-Garcia, S. Koponen, K. Kallio, and G. Camps-Valls, “Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data,” *Remote Sens (Basel)*, vol. 10, no. 5, May 2018, https://doi.org/10.3390/rs10050786.
51. O. Maron and A. W. Moore, “Hoeffding Races: Accelerating Model Selection Search for Classification and Function Approximation.”
52. M. Kuhn, “Futility Analysis in the Cross-Validation of Machine Learning Models,” May 2014, [Online]. Available: http://arxiv.org/abs/1405.6974.
53. M. Kuhn and J. Silge, *Tidy modeling with R*, 1st ed. O’Reilly Media, Inc., 2022.

54. H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int J Remote Sens*, vol. 27, no. 14, pp. 3025–3033, Jul. 2006, https://doi.org/10.1080/01431160600589179.

55. N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans Syst Man Cybern*, vol. 9, no. 1, pp. 62–66, Jan. 1979, https://doi.org/10.1109/TSMC.1979.4310076.

56. S. Montico, N. C. Di Leo, and J. A. Berardi, "Drought, low water levels and effects of fires on soils in the Paraná Delta, Argentina, in Spanish," *Cuadernos del CURIHAM*, May 2023, https://doi.org/10.35305/curiham.vi.199.

57. M. L. Amsler and E. C. Drago, "A review of the suspended sediment budget at the confluence of the Paraná and Paraguay Rivers," *Hydrol Process*, vol. 23, no. 22, pp. 3230–3235, Oct. 2009, https://doi.org/10.1002/hyp.7390.

58. M. L. Amsler *et al.*, *The Paraná River in its middle section: contribution to hydrological, geomorphological and sedimentological knowledge, in Spanish*. Santa Fe: Ediciones UNL, 2020.

59. L. I. Alcalá and M. F. Rus, "Critical Deficient Urban Areas in Territories with Water Risk. Comparative analysis of situations in the cities of Resistencia and Corrientes, in Spanish," 2017.

60. I. Hugo Rohrmann, I. Patricia Parini, I. Andrea Rolón, and T. Laura Noguera, "Urban water risk zoning by rainfall, in Spanish," Oct. 2013.

61. S. Ambrosino *et al.*, *Urban flooding in Argentina, in Spanish*. 2004.

62. K. Gu, Y. Zhang, and J. Qiao, "Random Forest Ensemble for River Turbidity Measurement from Space Remote Sensing Data," *IEEE Trans Instrum Meas*, vol. 69, no. 11, pp. 9028–9036, Nov. 2020, https://doi.org/10.1109/TIM.2020.2998615.

63. T. S. Rahul, J. Brema, and G. J. J. Wessley, "Evaluation of surface water quality of Ukkadam lake in Coimbatore using UAV and Sentinel-2 multispectral data," *International Journal of Environmental Science and Technology*, vol. 20, no. 3, pp. 3205–3220, Mar. 2023, https://doi.org/10.1007/s13762-022-04029-7.

64. M. Elhag, I. Gitas, A. Othman, J. Bahrawi, and P. Gikas, "Assessment of water quality parameters using temporal remote sensing spectral reflectance in arid environments, Saudi Arabia," *Water (Switzerland)*, vol. 11, no. 3, Mar. 2019, https://doi.org/10.3390/w11030556.

65. P. Biecek and T. Burzykowski, *Explanatory Model Analysis*. New York: Chapman and Hall/CRC, 2021.

66. M. Chowdhury, C. Vilas, S. van Bergeijk, G. Navarro, I. Laiz, and I. Caballero, "Monitoring turbidity in a highly variable estuary using Sentinel 2-A/B for ecosystem management applications," *Front Mar Sci*, vol. 10, Jul. 2023, https://doi.org/10.3389/fmars.2023.1186441.

67. S. Magri, E. Ottaviani, E. Prampolini, B. Federici, G. Besio, and B. Fabiano, "Application of machine learning techniques to derive sea water turbidity from Sentinel-2 imagery," *Remote Sens Appl*, p. 100951, Apr. 2023, https://doi.org/10.1016/j.rsase.2023.100951.

68. J. C. Ritchie, P. V Zimba, and J. H. Everitt, "Remote Sensing Techniques to Assess Water Quality," 2003.