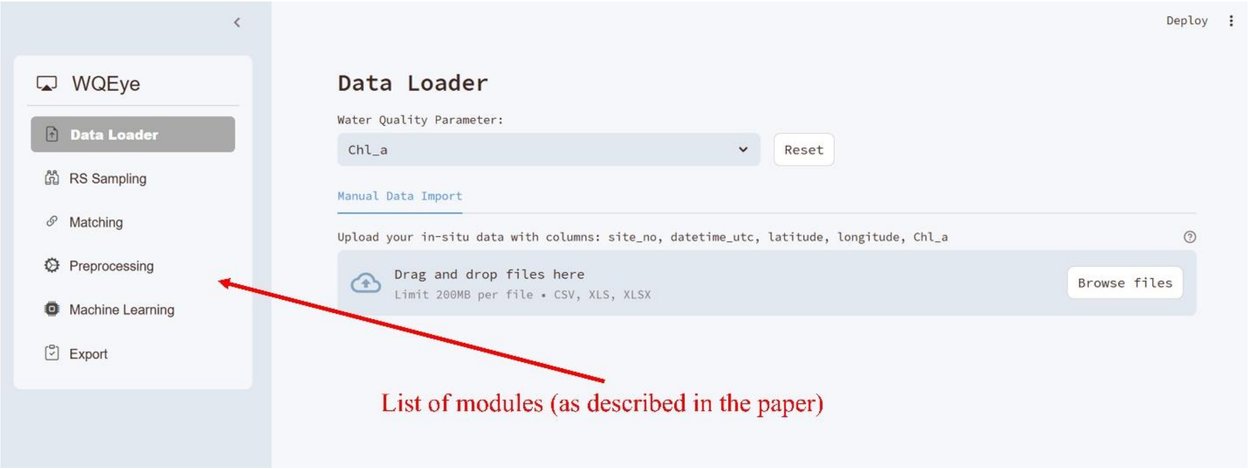
WQEye: A Python-based Software Aided by Google Earth Engine for Machine Learning-based Retrieval of Water Quality Parameters from Sentinel-2 and Landsat-8/9 Remote Sensing Data

1. Step-by-Step Tutorial

After successful installation, WQEye is launched automatically in your default web browser.

1.1. User Interface (Figure 1)

The left-hand panel displays the main modules of WQEye, corresponding to the key steps in the workflow: Data Loader, RS Sampling, Matching, Preprocessing, Machine Learning, and Export. Each module follows the structure described in detail in the paper.



Data Loader

Water Quality Parameter: Chl_a

Upload your in-situ data with columns: site_no, datetime_utc, latitude, longitude, Chl_a

Drag and drop files here
Limit 200MB per file • CSV, XLS, XLSX

List of modules (as described in the paper)

Figure 1. User Interface.

1.2. Data Loader Module

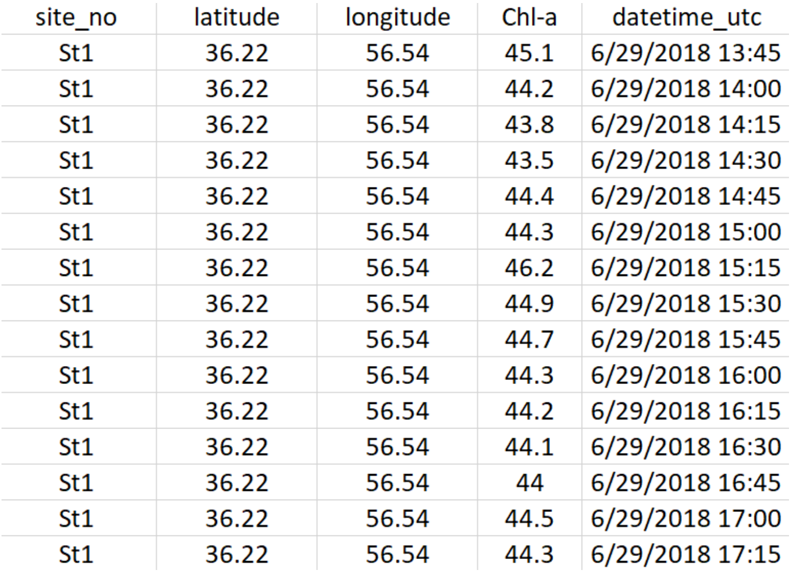
The workflow begins with the Data Loader Module, where users upload in-situ Water Quality Parameter (WQP) data. Uploaded datasets must follow the standardized format outlined below:

Column Name	**Description**
site_no	Unique identifier for each sampling station
latitude	Geographic latitude in decimal degrees (e.g., 36.224578)
longitude	Geographic longitude in decimal degrees (e.g., 56.542136)
WQP	Target water quality parameter
datetime_utc	Date and time of measurement in UTC (format: MM/DD/YYYY HH:MM)

Note:

- Latitude and longitude must be in decimal degrees.
- WQP values must use consistent units across all records.
- datetime_utc must follow the exact format MM/DD/YYYY HH:MM.

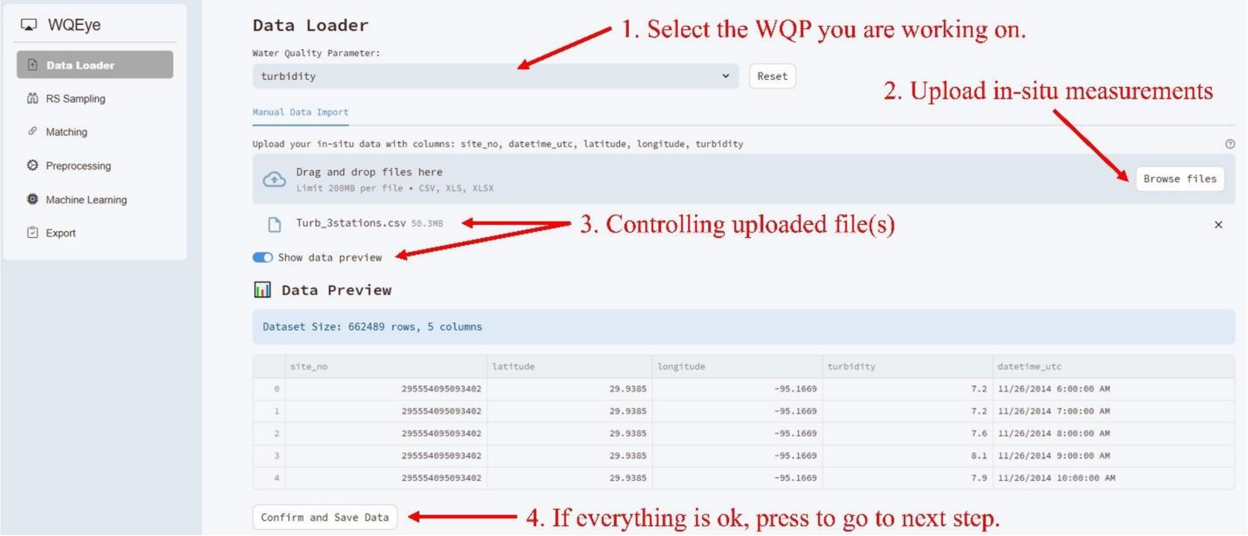
Figure 2 illustrates correctly formatted in-situ data for the Chl-a parameter at a single station St1. Users may provide a single file containing all stations (ensuring site_no is unique per station) or multiple files, each corresponding to a different station. Accepted file formats are: .CSV, .XLS, .XLSX. A sample dataset is available in the software repository for reference.



site_no	latitude	longitude	Chl-a	datetime_utc
St1	36.22	56.54	45.1	6/29/2018 13:45
St1	36.22	56.54	44.2	6/29/2018 14:00
St1	36.22	56.54	43.8	6/29/2018 14:15
St1	36.22	56.54	43.5	6/29/2018 14:30
St1	36.22	56.54	44.4	6/29/2018 14:45
St1	36.22	56.54	44.3	6/29/2018 15:00
St1	36.22	56.54	46.2	6/29/2018 15:15
St1	36.22	56.54	44.9	6/29/2018 15:30
St1	36.22	56.54	44.7	6/29/2018 15:45
St1	36.22	56.54	44.3	6/29/2018 16:00
St1	36.22	56.54	44.2	6/29/2018 16:15
St1	36.22	56.54	44.1	6/29/2018 16:30
St1	36.22	56.54	44	6/29/2018 16:45
St1	36.22	56.54	44.5	6/29/2018 17:00
St1	36.22	56.54	44.3	6/29/2018 17:15

Figure 2. Correctly formatted in-situ data for the Chl-a parameter.

Figure 3 demonstrates how to upload and manage in-situ WQP data within the Data Loader Module of WQEye. The first step in the process is to select the specific WQP the user intends to work on. Then, the user proceeds to upload the corresponding in-situ measurement data. This is accomplished by clicking the Browse files button, which allows the user to select one or more data files from their local system. After uploading the data, WQEye provides the option to preview the contents of the uploaded file(s). This is done by enabling the Show data preview toggle. When activated, a table is displayed, showing an overview of the dataset, including its total size and a sample of the records. This preview step allows the user to quickly verify that the data format and content are correct. If the uploaded data appears satisfactory, the user can click the Confirm and Save Data button to finalize the upload and proceed to the next module in the WQEye workflow. In case of any issues with the uploaded files or if the user wishes to restart the process, the Reset button can be used. This option removes all previously uploaded files and allows the process to begin from the start. In Figure 3, a unified file containing data from three monitoring stations has been uploaded for turbidity mapping. The dataset is reviewed using the preview option, and once confirmed to be correct, the user proceeds to the next step.



Data Loader

1. Select the WQP you are working on.

Water Quality Parameter: turbidity

2. Upload in-situ measurements

Upload your in-situ data with columns: site_no, datetime_utc, latitude, longitude, turbidity

Turb_3stations.csv 50.3MB

3. Controlling uploaded file(s)

Show data preview

Data Preview

Dataset Size: 662489 rows, 5 columns

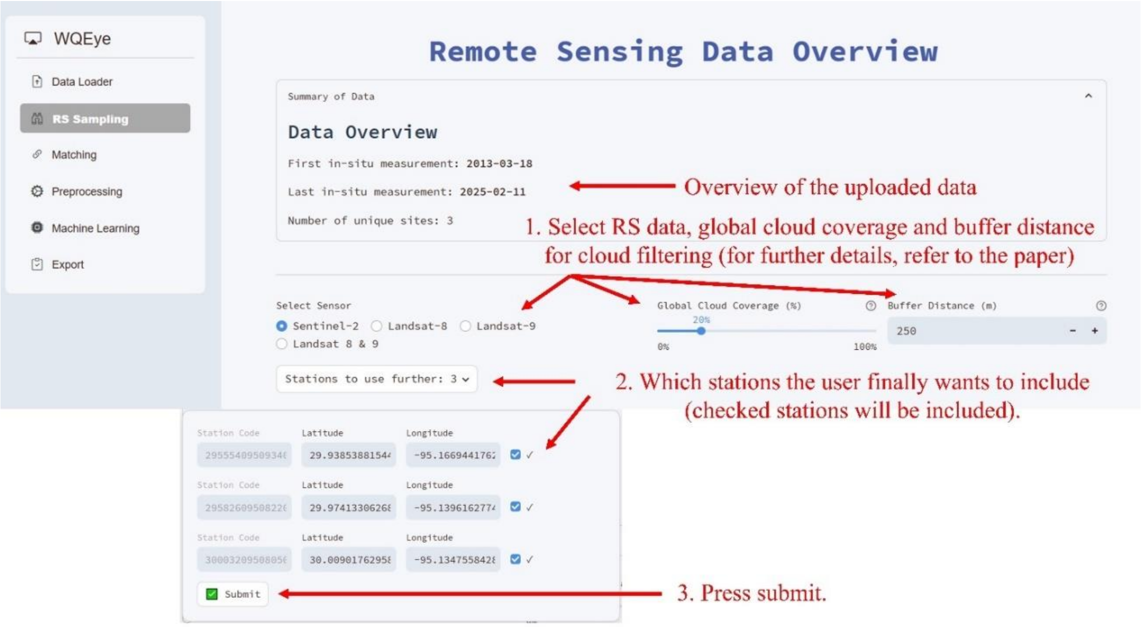
	site_no	latitude	longitude	turbidity	datetime_utc
0	295554095093402	29.9385	-95.1669	7.2	11/26/2014 6:00:00 AM
1	295554095093402	29.9385	-95.1669	7.2	11/26/2014 7:00:00 AM
2	295554095093402	29.9385	-95.1669	7.6	11/26/2014 8:00:00 AM
3	295554095093402	29.9385	-95.1669	8.1	11/26/2014 9:00:00 AM
4	295554095093402	29.9385	-95.1669	7.9	11/26/2014 10:00:00 AM

4. If everything is ok, press to go to next step.

Figure 3. Data Loader module.

1.3. RS sampling Module

Once the in-situ data has been successfully uploaded, the user proceeds to the RS Sampling Module (Figure 4), where remote sensing data are prepared for the analysis. At the top of this interface, WQEye provides a summary of the uploaded in-situ data. This includes the date of the first and last in-situ measurements, as well as the total number of unique monitoring stations identified in the uploaded dataset. This overview allows the user to verify that the correct dataset is being used for subsequent remote sensing data extraction. The next step is to select the remote sensing dataset to be used for sampling. WQEye currently supports Sentinel-2, Landsat-8, Landsat-9, and a combined Landsat-8 and Landsat-9 dataset. In Figure 4, Sentinel-2 has been selected as the source of satellite data. In addition to selecting the dataset, users have the option to define a global cloud coverage threshold and a buffer distance (please refer to the paper for further details). Based on the latitude and longitude information provided during the data upload, WQEye automatically identifies all unique monitoring stations. These are displayed along with their corresponding geographic coordinates. The user can then select which stations to include in the analysis by checking the relevant boxes. In Figure 4, all stations have been selected for further processing. Once the selections are complete, the user clicks the Submit button to confirm their choices. The software will then automatically extract satellite reflectance values for selected locations.



Remote Sensing Data Overview

Summary of Data

Data Overview

First in-situ measurement: 2013-03-18

Last in-situ measurement: 2025-02-11

Number of unique sites: 3

Overview of the uploaded data

1. Select RS data, global cloud coverage and buffer distance for cloud filtering (for further details, refer to the paper)

Select Sensor

● Sentinel-2 ○ Landsat-8 ○ Landsat-9
○ Landsat 8 & 9

Global Cloud Coverage (%): 20% (0% – 100%)

Buffer Distance (m): 250

Stations to use further: 3

2. Which stations the user finally wants to include (checked stations will be included).

Station Code	Latitude	Longitude	
2955540950934	29.93853881544	-95.1669441762	☑ ✓
2958260950822	29.9741330626	-95.1396162774	☑ ✓
3000320950805	30.0090176295	-95.1347558428	☑ ✓

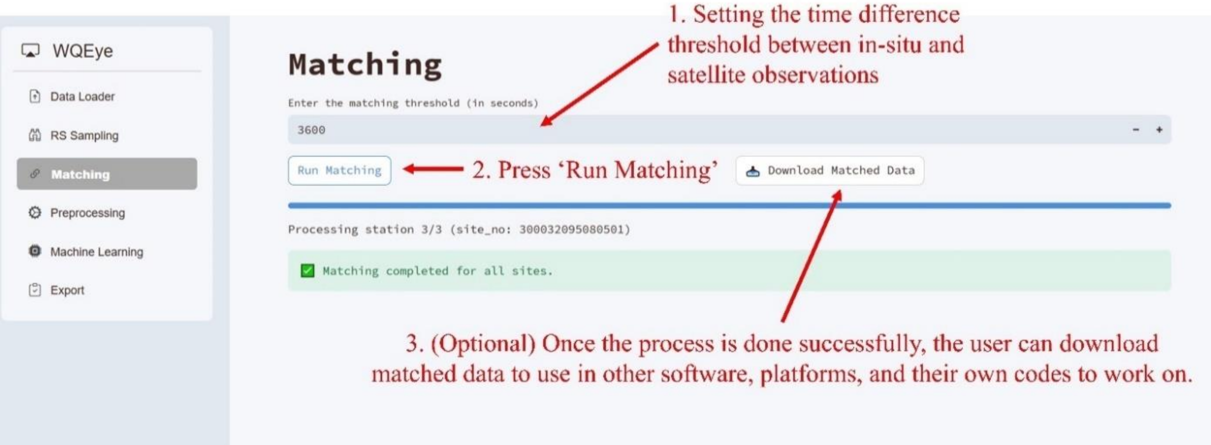
☑ Submit

3. Press submit.

Figure 4. RS sampling module.

1.4. Matching module

In this module (Figure 5), in-situ and satellite observations are paired based on their temporal proximity. The user begins by entering a time difference threshold, expressed in seconds. This threshold defines the maximum allowable time gap between an in-situ observation and the corresponding satellite image acquisition. Once the threshold is set, the user initiates the matching process by clicking the Run Matching button. When the matching process is completed successfully, the software displays a confirmation message indicating that all sites have been processed. At this stage, the user has the option to download the matched dataset by clicking the Download Matched Data button. The matched data can be used in other software platforms or integrated into the user's own custom code for further analysis if desired.



Enter the matching threshold (in seconds)
3600

Processing station 3/3 (site_no: 300032095080501)

✓ Matching completed for all sites.

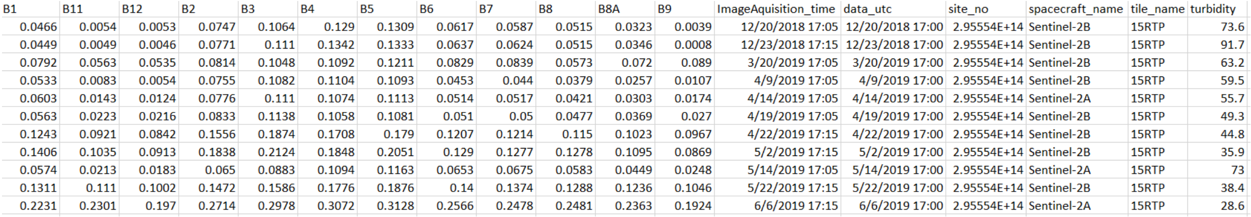
1. Setting the time difference threshold between in-situ and satellite observations

2. Press 'Run Matching'

3. (Optional) Once the process is done successfully, the user can download matched data to use in other software, platforms, and their own codes to work on.

Figure 5. Matching module.

WQEye generates a matched dataset, which is provided in a .csv format for convenient use in further analyses or model development (Figure 6). Each record in the matched dataset contains a comprehensive set of information that links satellite observations to the corresponding in-situ water quality measurements. Specifically, the file includes the spectral band values extracted from the selected remote sensing dataset (Sentinel-2 bands are shown in Figure 6). In addition to the spectral information, the file contains both the image acquisition time and the in-situ measurement time, expressed in Coordinated Universal Time (UTC). The dataset also includes the site number (site_no), which corresponds to the station identifiers provided by the user during the initial data upload. Further metadata are provided in the form of the spacecraft platform name, indicating the specific satellite from which the data were acquired (e.g., Sentinel-2A, Sentinel-2B, or Sentinel-2C). The tile_name column specifies the image scene (tile) from which the reflectance values were extracted, based on the Military Grid Reference System (MGRS) for Sentinel-2 data. Finally, the file contains the measured in-situ water quality parameter value for each record. In Figure 6, turbidity is the parameter of interest, and its corresponding values are listed in the final column. This matched dataset provides all the essential information required for model development, statistical analysis, or integration into the user's own workflows. The structured and transparent format ensures that users can readily apply the data for machine learning tasks or for validation purposes in other platforms and software environments.

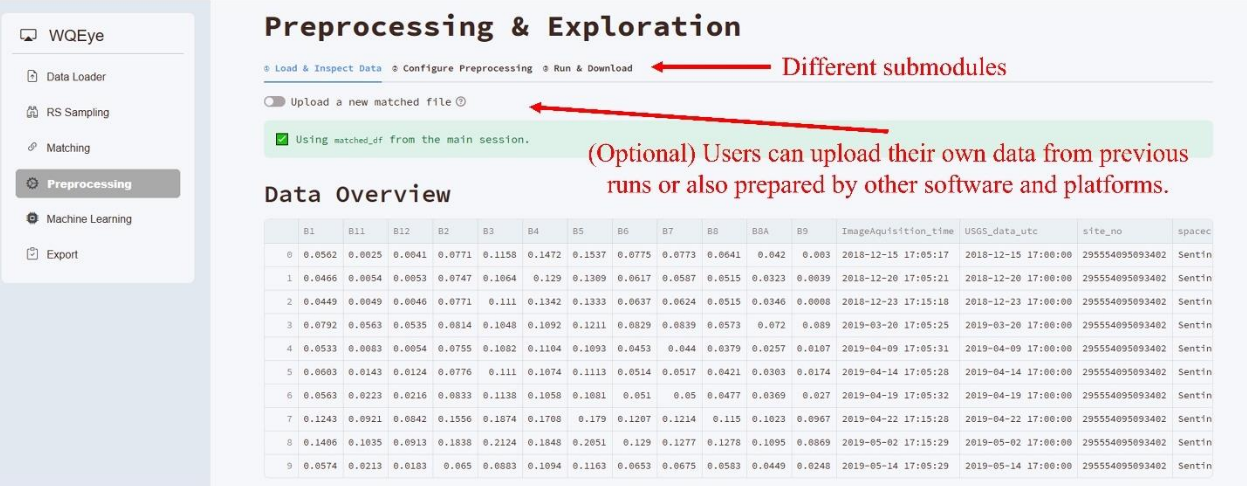


B1	B11	B12	B2	B3	B4	B5	B6	B7	B8	B8A	B9	ImageAquisition_time	data_utc	site_no	spacecraft_name	tile_name	turbidity
0.0466	0.0054	0.0053	0.0747	0.1064	0.129	0.1309	0.0617	0.0587	0.0515	0.0323	0.0039	12/20/2018 17:05	12/20/2018 17:00	2.95554E+14	Sentinel-2B	15RTP	73.6
0.0449	0.0049	0.0046	0.0771	0.111	0.1342	0.1333	0.0637	0.0624	0.0515	0.0346	0.0008	12/23/2018 17:15	12/23/2018 17:00	2.95554E+14	Sentinel-2B	15RTP	91.7
0.0792	0.0563	0.0535	0.0814	0.1048	0.1092	0.1211	0.0829	0.0839	0.0573	0.072	0.089	3/20/2019 17:05	3/20/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	63.2
0.0533	0.0083	0.0054	0.0755	0.1082	0.1104	0.1093	0.0453	0.044	0.0379	0.0257	0.0107	4/9/2019 17:05	4/9/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	59.5
0.0603	0.0143	0.0124	0.0776	0.111	0.1074	0.1113	0.0514	0.0517	0.0421	0.0303	0.0174	4/14/2019 17:05	4/14/2019 17:00	2.95554E+14	Sentinel-2A	15RTP	55.7
0.0563	0.0223	0.0216	0.0833	0.1138	0.1058	0.1081	0.051	0.05	0.0477	0.0369	0.027	4/19/2019 17:05	4/19/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	49.3
0.1243	0.0921	0.0842	0.1556	0.1874	0.1708	0.179	0.1207	0.1214	0.115	0.1023	0.0967	4/22/2019 17:15	4/22/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	44.8
0.1406	0.1035	0.0913	0.1838	0.2124	0.1848	0.2051	0.129	0.1277	0.1278	0.1095	0.0869	5/2/2019 17:15	5/2/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	35.9
0.0574	0.0213	0.0183	0.065	0.0883	0.1094	0.1163	0.0653	0.0675	0.0583	0.0449	0.0248	5/14/2019 17:05	5/14/2019 17:00	2.95554E+14	Sentinel-2A	15RTP	73
0.1311	0.111	0.1002	0.1472	0.1586	0.1776	0.1876	0.14	0.1374	0.1288	0.1236	0.1046	5/22/2019 17:15	5/22/2019 17:00	2.95554E+14	Sentinel-2B	15RTP	38.4
0.2231	0.2301	0.197	0.2714	0.2978	0.3072	0.3128	0.2566	0.2478	0.2481	0.2363	0.1924	6/6/2019 17:15	6/6/2019 17:00	2.95554E+14	Sentinel-2A	15RTP	28.6

Figure 6. An example of matched data.

1.5. Preprocessing module

The Preprocessing module (Figure 7) contains three submodules for managing and preparing data for model development. In the first submodule, Load & Inspect Data, the matched dataset from the previous step is automatically displayed. If the user already has matched data from previous runs or external sources, they can also upload it here, provided it follows the same format as the generated matched file.



Preprocessing & Exploration

Load & Inspect Data · Configure Preprocessing · Run & Download

Different submodules

Upload a new matched file

☑ Using matched_df from the main session.

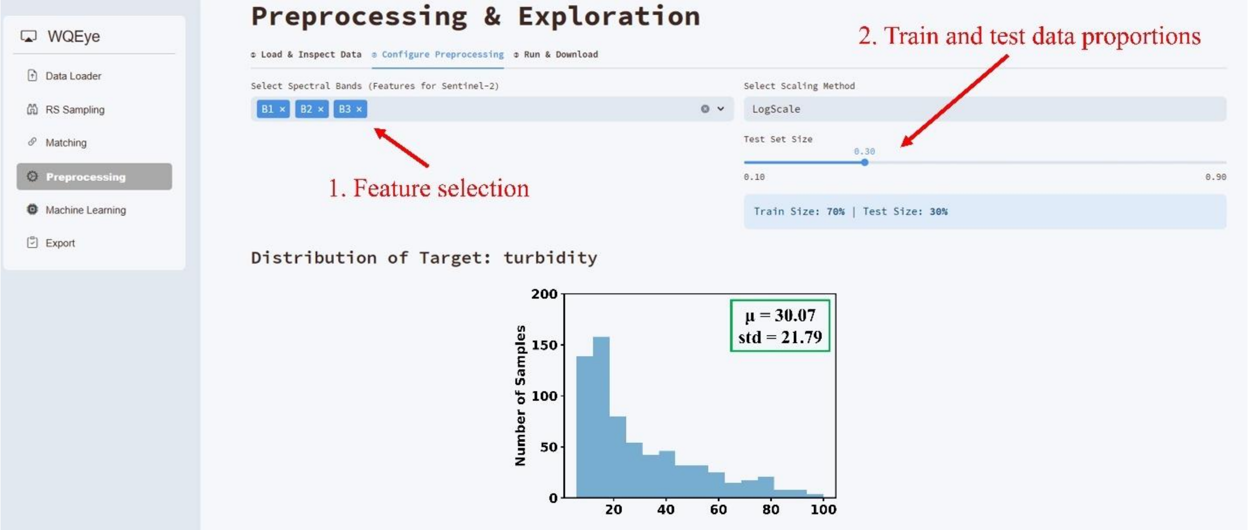
(Optional) Users can upload their own data from previous runs or also prepared by other software and platforms.

Data Overview

	B1	B11	B12	B2	B3	B4	B5	B6	B7	B8	B8A	B9	ImageAquisition_time	USGS_data_utc	site_no	spacec
0	0.0562	0.0025	0.0041	0.0771	0.1158	0.1472	0.1537	0.0775	0.0773	0.0641	0.042	0.003	2018-12-15 17:05:17	2018-12-15 17:00:00	295554095093402	Sentin
1	0.0466	0.0054	0.0053	0.0747	0.1064	0.129	0.1309	0.0617	0.0587	0.0515	0.0323	0.0039	2018-12-20 17:05:21	2018-12-20 17:00:00	295554095093402	Sentin
2	0.0449	0.0049	0.0046	0.0771	0.111	0.1342	0.1333	0.0637	0.0624	0.0515	0.0346	0.0008	2018-12-23 17:15:18	2018-12-23 17:00:00	295554095093402	Sentin
3	0.0792	0.0563	0.0535	0.0814	0.1048	0.1092	0.1211	0.0829	0.0839	0.0573	0.072	0.089	2019-03-20 17:05:25	2019-03-20 17:00:00	295554095093402	Sentin
4	0.0533	0.0083	0.0054	0.0755	0.1082	0.1104	0.1093	0.0453	0.044	0.0379	0.0257	0.0107	2019-04-09 17:05:31	2019-04-09 17:00:00	295554095093402	Sentin
5	0.0603	0.0143	0.0124	0.0776	0.111	0.1074	0.1113	0.0514	0.0517	0.0421	0.0303	0.0174	2019-04-14 17:05:28	2019-04-14 17:00:00	295554095093402	Sentin
6	0.0563	0.0223	0.0216	0.0833	0.1138	0.1058	0.1081	0.051	0.05	0.0477	0.0369	0.027	2019-04-19 17:05:32	2019-04-19 17:00:00	295554095093402	Sentin
7	0.1243	0.0921	0.0842	0.1556	0.1874	0.1708	0.179	0.1207	0.1214	0.115	0.1023	0.0967	2019-04-22 17:15:28	2019-04-22 17:00:00	295554095093402	Sentin
8	0.1406	0.1035	0.0913	0.1838	0.2124	0.1848	0.2051	0.129	0.1277	0.1278	0.1095	0.0869	2019-05-02 17:15:29	2019-05-02 17:00:00	295554095093402	Sentin
9	0.0574	0.0213	0.0183	0.065	0.0883	0.1094	0.1163	0.0653	0.0675	0.0583	0.0449	0.0248	2019-05-14 17:05:29	2019-05-14 17:00:00	295554095093402	Sentin

Figure 7. Preprocessing module (overview).

In the Configure Preprocessing submodule (Figure 8), the user selects which remote sensing spectral bands to include in the analysis and defines the proportion of training and testing data. During this step, the data are also preprocessed using the LogScale transformation, as described in detail in the paper. Additionally, the user can view the distribution of the matched water quality parameter values to assess data variability before proceeding.



Preprocessing & Exploration

Load & Inspect Data · Configure Preprocessing · Run & Download

Select Spectral Bands (Features for Sentinel-2)

B1 × B2 × B3 ×

1. Feature selection

Select Scaling Method

LogScale

2. Train and test data proportions

Test Set Size

0.30

0.10 0.90

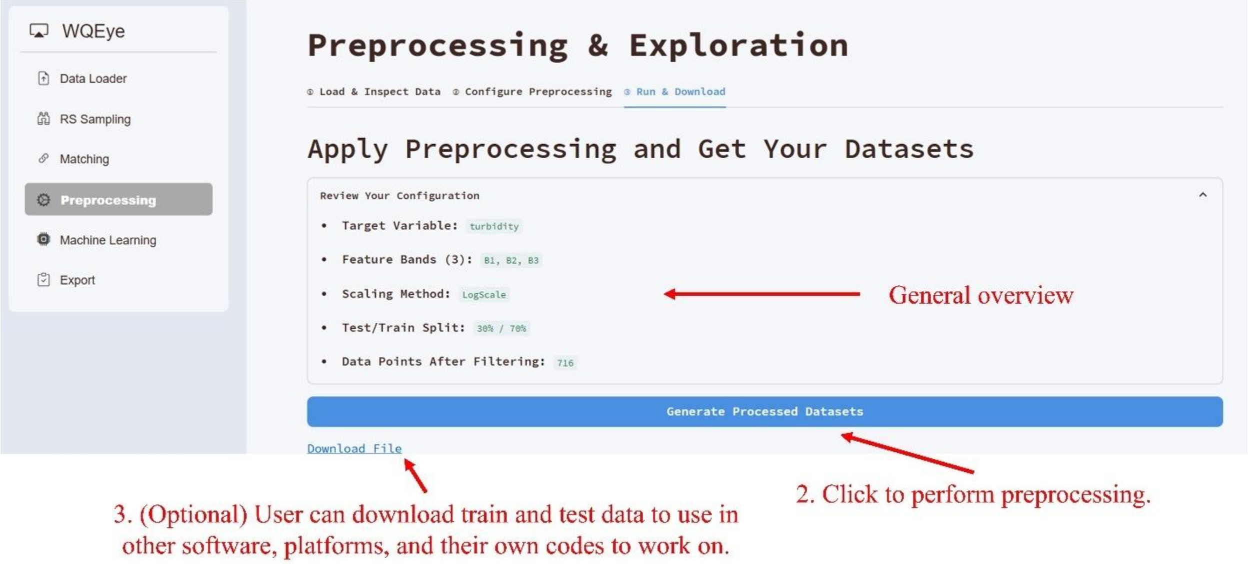
Train Size: 70% | Test Size: 30%

Distribution of Target: turbidity

μ = 30.07
std = 21.79

Figure 8. Preprocessing module (Configure Preprocessing).

In the Run & Download submodule (Figure 9), the user is first shown an overview of the selected preprocessing configuration, including the target variable, chosen spectral bands, scaling method, and the train-test split. After reviewing this information, the user clicks the Generate Processed Datasets button to apply the transformations. Once completed, a download option becomes available, allowing the user to export training and testing datasets for use in other software, platforms, or their own custom code.



Preprocessing & Exploration

Apply Preprocessing and Get Your Datasets

Review Your Configuration

- Target Variable: turbidity
- Feature Bands (3): B1, B2, B3
- Scaling Method: LogScale
- Test/Train Split: 30% / 70%
- Data Points After Filtering: 716

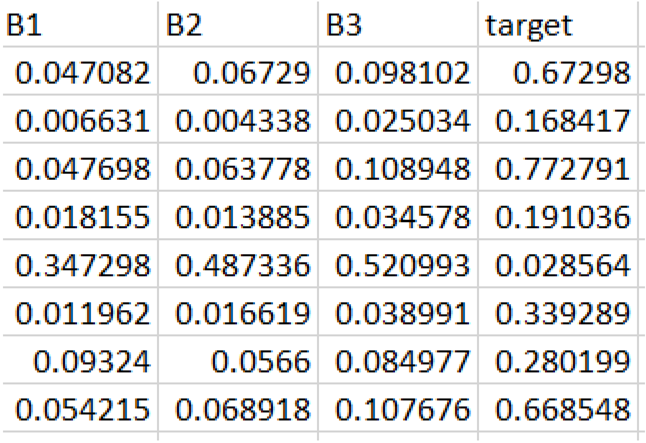
General overview

2. Click to perform preprocessing.

3. (Optional) User can download train and test data to use in other software, platforms, and their own codes to work on.

Figure 9. Preprocessing module (Run and Download).

The downloaded files from the preprocessing step include the training and testing datasets as well as a .pkl file containing the transformation parameters. This .pkl file allows users to reverse the transformations and convert the data back to its original scale if needed. The training and testing datasets contain the selected remote sensing spectral bands, in this example, bands B1, B2, and B3, and the transformed target water quality parameter values, ready for model development.

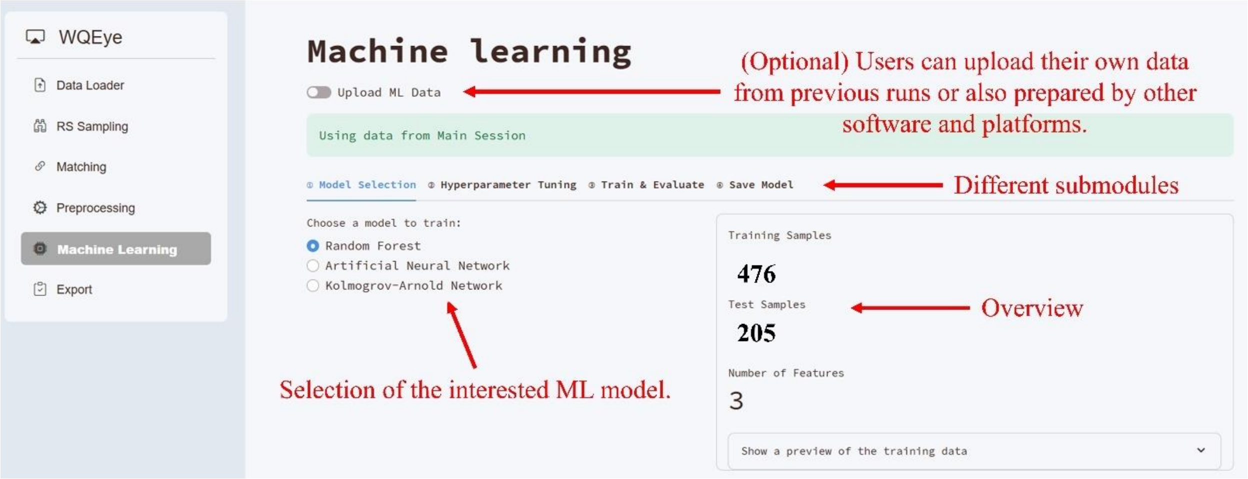


B1	B2	B3	target
0.047082	0.06729	0.098102	0.67298
0.006631	0.004338	0.025034	0.168417
0.047698	0.063778	0.108948	0.772791
0.018155	0.013885	0.034578	0.191036
0.347298	0.487336	0.520993	0.028564
0.011962	0.016619	0.038991	0.339289
0.09324	0.0566	0.084977	0.280199
0.054215	0.068918	0.107676	0.668548

Figure 10. A sample training data.

1.6. ML module

The ML module (Figure 11) contains several submodules that guide the user through the model development. In the Model Selection submodule, the user chooses one of three available machine learning models: Random Forest, Artificial Neural Network, or Kolmogorov-Arnold Network. The module also provides an overview of the dataset, including the number of training and testing samples, as well as the number of selected input features. If users already have their own prepared training and testing datasets from previous runs or external tools, they can upload them directly for use in this module. In the example shown, the Random Forest model has been selected for training.



Machine learning

Upload ML Data

Using data from Main Session

Choose a model to train:

● Random Forest
○ Artificial Neural Network
○ Kolmogorov-Arnold Network

Training Samples
476

Test Samples
205

Number of Features
3

(Optional) Users can upload their own data from previous runs or also prepared by other software and platforms.

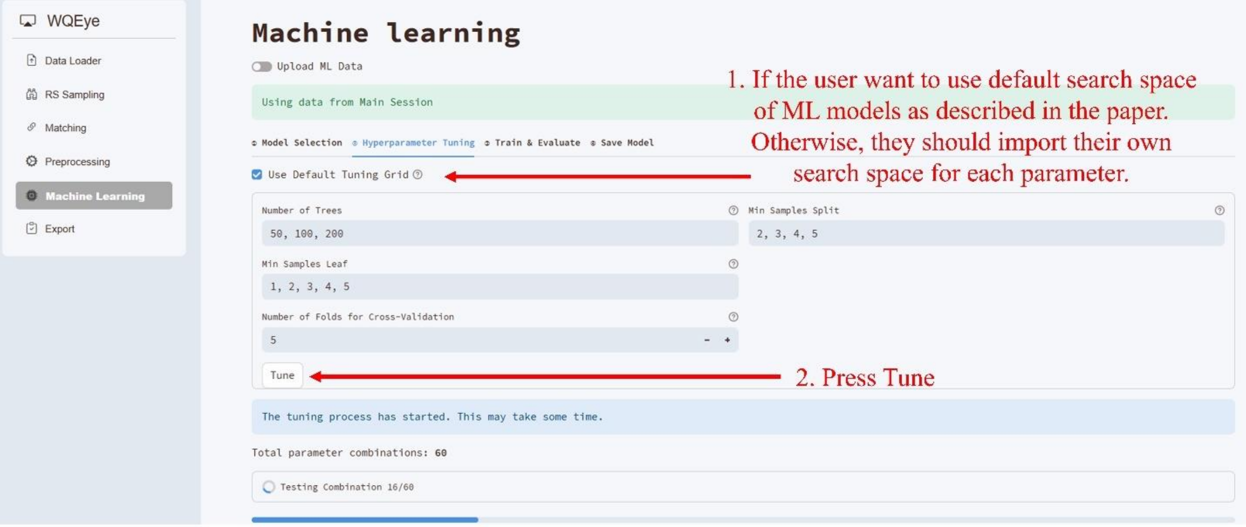
Different submodules

Overview

Selection of the interested ML model.

Figure 11. ML module (model selection).

In the Hyperparameter Tuning submodule (Figure 12), the user can adjust the search space for each machine learning model parameter. By default, WQEye provides a recommended search space as described in the paper. Users may also define their own parameter ranges if preferred. Additionally, the number of folds for cross-validation can be specified; in this example, five folds have been selected. Once the settings are confirmed, the user clicks the Tune button to begin hyperparameter tuning. The software tests all possible parameter combinations within the defined grid, and the best set of parameters is selected for model training.



Using data from Main Session

☑ Use Default Tuning Grid

Number of Trees: 50, 100, 200

Min Samples Split: 2, 3, 4, 5

Min Samples Leaf: 1, 2, 3, 4, 5

Number of Folds for Cross-Validation: 5

The tuning process has started. This may take some time.

Total parameter combinations: 60

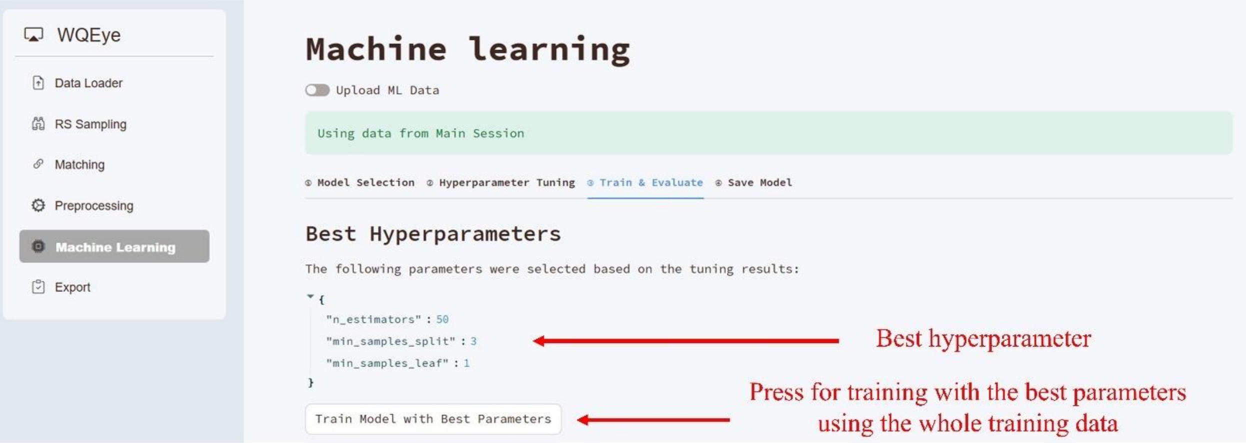
Testing Combination 16/60

1. If the user want to use default search space of ML models as described in the paper. Otherwise, they should import their own search space for each parameter.

2. Press Tune

Figure 12. ML module (Hyperparameter Tuning).

In the Train and Evaluate submodule, the model is trained on the entire training dataset using the best hyperparameters identified during the tuning process. Once training is complete, the software automatically computes performance metrics. Finally, in the Save Model submodule, the trained model can be saved for future use, allowing the user to apply it in other analyses or workflows as needed.



Using data from Main Session

Best Hyperparameters

The following parameters were selected based on the tuning results:

```
{
  "n_estimators" : 50
  "min_samples_split" : 3
  "min_samples_leaf" : 1
}
```

Best hyperparameter

Press for training with the best parameters using the whole training data

Figure 13. ML module (Training and Saving).