# scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

# A Hyperspectral Reflectance Database of Plastic Debris with Different Fractional Abundance in River Systems

Mohammadali Olyaei ✉, Ardeshir Ebtehaj & Christopher R. Ellis

**Plastic debris pollution transported by river systems to lakes and oceans has emerged as a significant environmental concern with adverse impacts on ecosystems, food webs, and human health. Remote sensing presents a cost-effective approach to bolster interception and removal efforts. However, unlike marine environments, the optical properties of plastic debris in fresh waters remain poorly understood. This study aims to address this gap by providing an open-access hyperspectral reflectance database of floating weathered and virgin plastic debris found in river systems under controlled laboratory experiments. Utilizing natural waters from the Mississippi River, the database was assembled using a remote sensing data acquisition system deployed over a hydraulic flume operating under subcritical flow conditions and varying suspended sediment concentrations. The measurements encompass hyperspectral diffused light reflectance from ultraviolet (UV, 350 nm) to shortwave infrared (SWIR, 2500 nm) wavelengths. The database archived in Network Common Data Form (NetCDF) and Comma-separated values (CSV), offers valuable insights for better understanding key spectral signatures indicative of floating plastic debris, with different fractional abundance, in freshwater ecosystems.**

## Background & Summary

Plastic pollution in aquatic environments presents significant environmental, economic, and recreational concerns[1–3]. Plastic pollution in the oceans stands out as a pressing environmental issue, garnering recognition from the United Nations Environment Programme (UNEP) as one of the foremost emerging pollution challenges[4,5]. This recognition has spurred investments in research, social awareness campaigns, and proactive measures for control and prevention[3,6] with some completed or ongoing projects documented in[7]. The plastic contamination of lakes and rivers has received comparatively less attention[8,9] than oceans, even though a significant portion of plastic waste finds its way to the oceans via river networks, with annual estimates ranging from 0.8 to 2.7 million metric tons[10,11]. Accurate estimation of plastic emissions from rivers to the global oceans necessitates detailed information on the spatiotemporal flux of the floating debris[12,13], which can be intensified during extreme flood events[14]. Traditional methods for detecting plastic litter in aquatic environments rely on labor-intensive in situ sampling[15,16], which proves to be time-consuming and expensive[17,18]. Alternatively, high-resolution optical remote sensing technologies offer cost-effective solutions with extensive space-time coverage[19–22].

Remote sensing of plastic litter requires proper characterization of its optical spectral characteristics especially within the visible to near-infrared (VNIR) and short-wave infrared (SWIR) segments of the electromagnetic spectrum[23–25]. A substantial body of research has been dedicated to measuring the spectral reflectance of marine litter, both in controlled laboratory settings[26] and in field conditions[27,28]. Laboratory investigations have primarily focused on scrutinizing the spectral signatures of various virgin and weathered plastics under dry and wet conditions across the VNIR to SWIR bands[29–35]. It is found that different plastic types display unique spectral signatures[31,32,36] contingent upon their material composition, age, and the optical complexities of their surrounding waters. In contrast, field studies have predominantly concentrated on formulating algorithms for plastic debris detection in marine environments, utilizing airborne and spaceborne platforms such as the multispectral instrument aboard the Sentinel-2 satellite[37–43]. However, research and methodologies concerning

Saint Anthony Falls Laboratory, University of Minnesota, Minneapolis, MN, 55455, USA. ✉e-mail: olyae001@umn.edu

freshwater riverine and lake ecosystems have been relatively limited, with only a handful of studies employing unmanned aerial vehicles (UAVs)[19] and satellites[44] for plastic litter detection using deep learning techniques[45,46], as well as in situ imaging[47,48].

Conducting field-scale experiments in freshwater aquatic environments is costly, primarily due to the limited space-time resolution of remote sensing platforms and the complexities associated with accurate labeling of the observations[38,49–51]. Furthermore, the presence of various other floating materials within the field of view (FOV) can obscure the spectral signature of plastic litter, including different types of *Sargassum*, driftwood, white caps, and water foams[25,52,53]. While discriminating floating algae from other debris is feasible due to the distinct optical properties of chlorophyll pigments, distinguishing between different plastic types and non-plastic debris remains challenging[52]. As previously noted, laboratory experiments[31,33,34,54,55] have primarily focused on the spectral properties of marine litter. These experiments often employed simplified background environments such as static water and a single type of plastic debris within the FOV without explicitly accounting for the heterogeneity of sub-grid debris abundance. These limitations can be important for remote sensing of small patches of different types of debris in freshwater ecosystems, where factors such as flow turbulence and background dynamics, including flow turbidity, waves, and water foams, can alter spectral signatures within the FOV.

Turbid waters with high concentrations of fine-grain suspended sediments tend to exhibit higher reflectance compared to clear water, particularly in the visible and NIR bands[56–59]. At the same time, the presence of turbid waters can also attenuate the magnitude of submerged plastic reflectance[54,60] due to a combination of strong water absorption in the SWIR spectrum and reduced water transparency in the visible region[52]. In lakes and rivers, the presence of colored dissolved organic matter (CDOM) can also affect the reflectance signal in the visible part of the spectrum; however, the peak near 810 nm is proven to be effective in retrieving water quality parameters such as suspended sediments and algal pigments[61]. In marine environments, it has been demonstrated that breaking wind waves or whitecaps can alter the magnitude and spectral shape of surface water reflectance[62,63]. These effects are partially attributed to foam plumes and subsurface bubbles, which enhance the reflectance primarily in the visible and to a lesser extent in NIR and SWIR bands[64] and often manifest themselves as absorption lines centered around 600, 756, 980, 1198, 1448, and 1932 nm[65]. Therefore, we hypothesize that the spectral signatures of plastic debris in river systems can be influenced markedly by their fractional abundance and the optically active background constituents.

To investigate this hypothesis, we designed a data acquisition system at one of the Saint Anthony Falls Laboratory's (SAFL) flumes to collect diffused hyperspectral reflectance of various types of virgin and weathered plastic debris over fresh water from the Mississippi River with the following contributions. (i) The study accounts for the sub-grid fractions of debris by synchronizing a spectroradiometer with a digital single-lens reflex (DSLR) camera. (ii) Various water flow regimes have been designed to replicate different background dynamics, including turbidity and surface water foams. (iii) A collective inference is made regarding important absorption lines across different types of debris, particularly mixed-type plastics. (iv) The collected reflectance database is presented in a machine-independent format containing reflectance spectra of plastic specimens, coincident RGB images, labeled images capturing the debris fractional abundance within the FOV, as well as key measured environmental variables (e.g., sediment concentration, debris polymer, color, background flow status, flow discharge, etc). This open-source dataset will provide valuable insights into the spectral signatures of various types of floating plastic litter in different freshwater conditions. It aims to advance technologies for remotely detecting plastic pollution and estimating its flow from rivers to oceans. Additionally, it will encourage cross-disciplinary collaborations and support the development of remote sensing technologies for accurate detection of floating plastic using airborne and satellite imagery.

## Methods

**Spectral acquisition system.** An indoor data acquisition system (Fig. 1) was designed to measure the diffuse reflectance of floating debris over a hydraulic flume. An Analytical Spectral Devices (ASD) spectroradiometer (350–2500 nm) with a resolution of 3 (10–12) nm in VINR (SWIR) was deployed. The foreoptic FOV was maintained constant at 25° across all measurements. The spectroradiometer was mounted on a mobile carriage system, allowing for adjustment of its position in the flume and its distance from the water surface. Throughout the experiments, the distance from the end of the fiber optic cable to the water surface was set to 110 cm, resulting in a FOV with a diameter of 48 cm. The ASD was set to average five scans per measurement. A digital single-lens reflex (DSLR) camera was mounted on the carriage and synchronized with the spectroradiometer to simultaneously image (Fig. 1a) the sub-grid fractional abundance and type of plastics within the FOV.

The light source was provided by a set of six Tungesten-Halogen lights operating at temperature 3000 K offering stable illumination across the complete spectral range of the spectroradiometer for accurate indoor diffuse reflectance measurements (Fig. 1b). A Direct Current (DC) power supply was employed to provide precise control over the illumination intensity by eliminating the periodic fluctuations in the SWIR region of the spectrum – caused by the interference between the light source powered by alternating current and the ASD scan rate. The lights were positioned off-nadir to prevent multiple (specular) reflections within the FOV of ASD and the DSLR camera while providing adequate radiance.

It is well understood that reflectance $R = \frac{\Gamma_{up}}{\Gamma_{dn}}$ is defined as the ratio of upwelling $\Gamma_{up}$ to the downwelling irradiance $\Gamma_{dn}$[24], which can represent both diffused and directional reflection mechanisms[31]. Since the ASD is a radiometer, a reference Lambertian reflectance value must be acquired to calibrate the subsequent measurements. To that end, a white GORE-TEX Polytetrafluoroethylene disk, with a wide-band reflectance of close to 1 and a diameter of $D = 70$ cm, was used to cover the entire FOV as an ideal reflector. The disk material was chosen based on personal communications with the spectroradiometer design team. The spectroradiometer was recalibrated for each experiment round, especially when the background settings (e.g., sediment concentration,
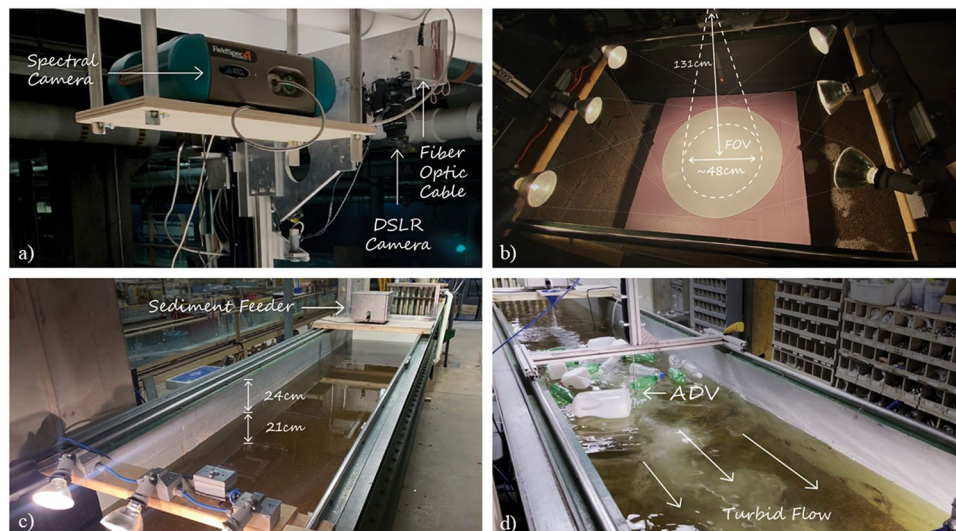
**Fig. 1** (**a**) The ASD FieldSpec-4 spectroradiometer along with the DSLR camera, (**b**) Illumination system and the reference panel showing that the walls were painted black, (**c**) Sediment feeder placement in the flume, (**d**) floating plastic debris in a turbid flow showing the position of the Acoustic Doppler Velocimeter (ADV).

illumination intensity) were changed. The following key steps are considered to systematize the data collection process. (1) A representative number of plastic-type specimens is used. (2) A wide range of background water turbidity and foamy conditions is simulated. (3) Multiple measurements per sample are collected to minimize noise and variability. (4) The measurements are categorized into relevant labels and subgroups for organized analyses. (5) Quality assurance is applied before the data are organized in a structured format. (6) The data are validated and compared against other published data sets.

**Flume and water flow characteristics.** Spectral measurements were conducted over the tilted flume at SAFL. The flume, with dimensions $14.6 \times 0.9 \times 0.6$ m, can transport up to $0.17$ m$^3$s$^{-1}$ of discharge. The flume has a pivot that allows it to tilt up to 6° and reproduce a wide range of river flows and sediment transport regimes. The bottom of the flume was covered by 16 cm coarse sand (Fig. 1c) to resemble natural sediments in riverine systems. The perimeter of the flume was painted black at the location of the spectral acquisition system to alleviate multiple reflections at the boundaries and resemble an optically deep region. The discharge was set to $0.01$ m$^3$s$^{-1}$, resulting in a Froude number of 0.045, implying a subcritical flow regime. The streamwise velocity was measured using an Acoustic Doppler velocimetry (ADV), at the channel center (Fig. 1d), and the flume slope was set to zero.

Three distinct background conditions were considered: naturally clear, turbid, and foamy waters. Mississippi water served as the clear water reference as the bottom of the channel was visible at a depth of around 20 cm during the experiments in the winter. The concentration of suspended sediments was changed to increase the turbidity as it impacts the surface reflectance across VNIR to SWIR[59,66], particularly between 700 and 900 nm wavelength[67]. Well-mixed suspended sediments require a shear velocity $u_*$ considerably greater than the settling velocity $v_s$, a condition that can be assessed through the Rouse number $R_o = v_s/\kappa u_*$, where $\kappa = 0.41$ is the Von Kármán constant[68]. To fulfill this condition, fine silt, with $d_{50} = 0.0075$ mm, was used resulting in $R_o \in [0.02 - 0.07] \ll 1$, assuring a well-mixed suspended sediment load[69].

The silt particles were released into the water via an in-house built sediment feeder (Fig. 1c), positioned 6 m upstream of the reflectance acquisition system. To that end, the feeder uses a peristaltic pump with an adjustable angular velocity. For measuring the total suspended material (TSM), we developed a mass flux curve representing the weight of released sediments in time as a function of the angular velocity to produce five different well-mixed TSM values at 58, 116, 180, 242, and 313 mgl$^{-1}$. The selected TSM range of 58 to 313 mgl$^{-1}$ encompasses various natural turbidity levels, including concentrations above 100 mgl$^{-1}$, which are indicative of high turbid conditions typically found in certain river systems and estuaries[66,70]. A propeller attached to an electric motor was properly positioned to generate upstream bubbles and make the water surface foamy for obtaining measurements over foamy waters. In a marine environment, whitecap reflectance varies based on the foam layers and the concentration of submerged bubbles that appear white to human eyes[65]. Inspired by this observation, we employed an image segmentation technique to quantify the extent of foam within FOV, as explained later.

**Plastic specimens.** In this study, we examine three primary floating polymers: polyethylene, polypropylene (PP), and expanded polystyrene (EPS)[21]. For polyethylene specimens, the variants are Polyethylene terephthalate (PET), High-density Polyethylene (HDPE), and Low-density polyethylene (LDPE). Examples for each group include water bottles, detergent or milk containers, and cups. The polypropylene plastic specimens are mainly ropes and bags and foams are used for EPS polymers. Three distinct settings for plastics were defined. i) Single virgin: Plastic specimens from one specific type of polymer were used. ii) Mix virgin: Plastics from multiple polymer types are used. iii) Weathered plastics: A collection of mixed weathered plastics is gathered from the Mississippi
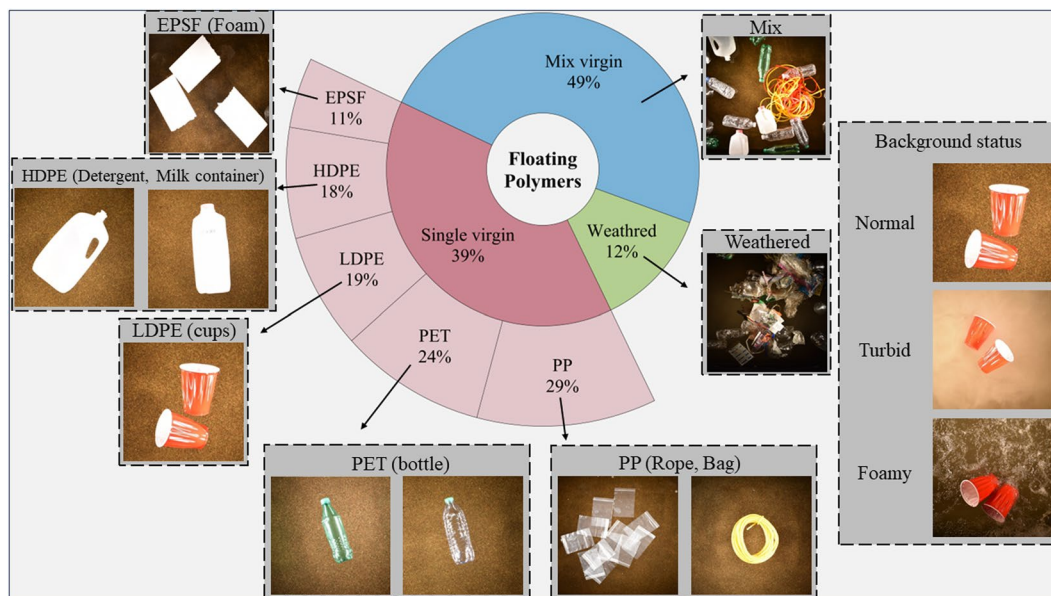
**Fig. 2** Floating plastic specimens in different settings (Single virgin, Mix virgin, and Weathered) against various background conditions (Normal, Turbid, and Foamy). The numbers are the percentage of experiments for each setting.

| | | Background Waters | | | Sum |
|---|---|---|---|---|---|
| | | Normal | Sediment | Foamy | |
| **SpecimenType** | PET | 54 (2.6%)–10.9% | 119 (5.7%)–19.2% | 19 (0.9%)–20.4% | 192 (9.2%) |
| | HDPE | 30 (1.4%)–13.2% | 89 (4.3%)–21.2% | 26 (1.3%)–16.4% | 145 (7.0%) |
| | LDPE | 18 (0.8%)–10.6% | 82 (3.9%)–16.6% | 52 (2.5%)–15.2% | 152 (7.3%) |
| | PP | 52 (2.6%)–19.9% | 114 (5.5%)–19.1% | 68 (3.3%)–15.8% | 234 (11.3%) |
| | EPSF | 16 (0.8%)–22.6% | 68 (3.3%)–22.3% | 8 (0.4%)–17.1% | 92 (4.4%) |
| | Mix | 445 (21.4%)–28.5% | 332 (16.0%)–28.2% | 231 (11.1%)–29.8% | 1008 (48.5%) |
| | Weathered | 23 (1.1%)–33.1% | 159 (7.7%)–29.0% | 73 (3.5%)–35.9% | 255 (12.3%) |
| | Sum | 638 (30.7%) | 963 (46.3%) | 477 (23.0%) | 2078 (100%) |

**Table 1.** Number of measurements categorized by different plastic types and background waters, with percentages indicated in parentheses followed by the mean value of plastic fraction.

River and Lake Hiawatha in Minneapolis. Figure 2 depicts the plastic and background settings along with the pictures of the specimens within each group. For each set of measurements across all debris subgroups, plastic debris was released upstream in the flume channel. As the debris entered the FOV, we began recording with both the ASD spectrometer and the DSLR camera multiple times until all debris exited the FOV. The synchronization between the ASD spectra and the RGB images was achieved by matching their timestamps, and we discarded any spectra without a corresponding RGB image with a matching timestamp. Since we intended to emulate the natural movement of plastic debris over river flow systems, we did not manually alter the dryness or wetness of the plastic specimens. However, the surface of HDPE, EPSF, and PET samples, whose densities are significantly lighter than water, remain likely dry even though some wetness is expected due to some rotational movements. In foamy waters, the likelihood of surface wetness increases markedly because of turbulence. However, the PP and LDPE samples are denser than other specimens and more prone to partial submergence and surface wetness. In total, 2078 measurements of spectral reflectance were obtained for 21 subgroups: seven specimen types (i.e., PET, HDPE, LDPE, PP, EPSF, Mixed, and Weathered) and three background water conditions (i.e., normal, sediment, and foamy). Table 1 provides an overview of sample distributes within these subgroups.

**Labeling DSLR images.** All coincident DSLR RGB images are cropped to display the circular FOV of the ASD spectroradiometer. The plastic fractions within the FOV are estimated using an image segmentation approach based on Otsu's method[71] of color thresholding[72]. To that end, the RGB color model is first transformed into the YCbCr color code, separating the luminance (Y) from chrominance (Cb and Cr). Unlike the RGB domain, the YCbCr color space separates the brightness from colors, allowing improved segmentation[73]. Otsu's method is among variance-based techniques[74] that seek a threshold that minimizes the within-class variance between the background and foreground values of an image[75]. Figure 3 shows examples of the image segmentation approach for different plastic specimens.
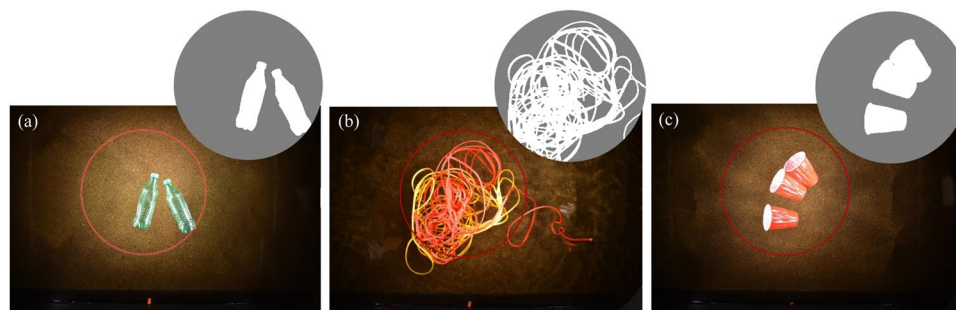
**Fig. 3** Examples of image segmentation approach using color thresholding for (**a**) green PET bottle, (**b**) red and yellow polypropylene rope, and (**c**) red LDPE cups.
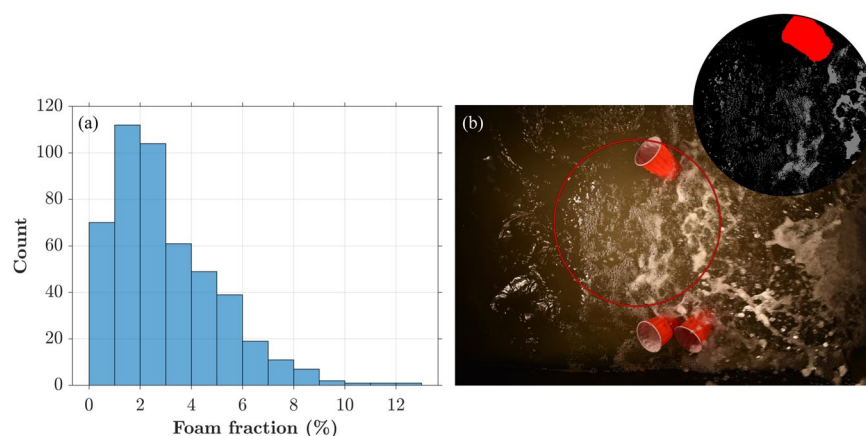


**Fig. 4** (**a**) Histogram of foam fractions, based on thresholding the grayscale imageries, with foamy pixels exhibiting intensities greater than 160. (**b**) Example of image segmentation with LDPE cup specimens in a foamy flow with three labels (water, foam, and plastics). The foam fraction is equal to 12.4% of FOV.

An additional step is taken to estimate the foam fractions. Initially, the RGB image is converted into a grayscale image by forming a weighted sum of the three R (red), G (green), and B (blue) channels using the coefficient from the National Television Standards Committee (NTSC)[76]. Subsequently, pixels above a specified threshold are selected to separate white foams from the background, after removing the plastic labels. We found that setting the threshold as 160 out of 255 could effectively capture the foamy pixels by trial and error. Figure 4(a) represents the histogram of foam fraction showing a range of 0–12% with a mode of 1.5%. An example of segmenting the foams from the background water and the floating debris is shown in Fig. 4(b).

## Data Records

The data are at Zenodo research repository[77]. The entire set of data records is stored in a machine-independent Network Common Data Form (NetCDF) and Comma-separated values (CSV) format. Each file follows the labeling structure in the format of ON_PT_WSC_C_SC_F_QF, where the observation number (ON), specimen polymer type (PT), background water condition followed by sediment concentration (WSC), specimen color (C), plastic fraction within the FOV (F), and a quality flag (QF)[78] are stored. The flag ranges from 1 to 5 representing low to high quality values of the provided segmentation based on visual inspections.

Specifically, each NetCDF file contains the following data fields: i) hyperspectral reflectance spectrum from 350–2500 nm with the aforementioned spectral resolutions; ii) coincident RGB images associated with the time of spectral measurements within the FOV; iii) labeled images segmenting the plastics and foams from the background water; and iv) general information related to polymer type, color, the categories describing water background (clear, turbid, and foamy), the areal fraction of the debris, sediment contention, flow discharge, and the quality flag – as reported in Table 2. The CSV files only contain the spectral measurements. It is worth noting that after recording each reflectance spectrum, a splice correction was applied to remove potential discontinuity in spectral measurements at 1000 and 1800 nm, where the type of scanner changes in the ASD spectroradiometer. To that end, we used the technical guides[79] that offer splice corrections based on the calculated biases at the spectral interfaces between the scanners.

**Plastic fraction signatures.** Figure 5 represents some statistics of the collected reflectance spectra in the clear water for single virgin specimens (i.e., PET, HDPE, LDPE, PP, and EPSF), mixed and weathered plastics. The statistics are calculated based on the number of sample spectra provided in Table 1. The mean spectra for plastic

| Term | Definition |
|------|-----------|
| Reflectance | Hyperspectral reflectance of the plastic specimen measured across the spectrum from 350 to 2500 nm. |
| RgbImage | Digital image of the FOV comprising three color channels: red, green, and blue. |
| LableledImage | Pixel-wise image indicating water, plastics, and foams within the FOV. |
| Debris Polymer | The constitute polymer (e.g., PET, HDPE,...) of the plastic specimens. |
| Debris Color | The colors of the plastic specimen (T: Transparent, G: Green, W: White, R: Red, Y: Yellow, M: Mix). |
| Background flow status | Three categories describing the water background conditions: clear, turbid, and foamy. |
| Plastic fraction (%) | Percentage of sub-grid fraction of the plastic litter. |
| Foam fraction (%) | Percentage of sub-grid fraction of foamy waters. |
| Sediment concentration | Suspended sediment concentration $mgl^{-1}$. |
| Flow discharge | Discharge of the flume in $m^3s^{-1}$. |
| Quality Flag | A numeric value indicating the quality of the pixel-wise labeled image. |

**Table 2.** Description of attributes used in the NETCDF library.

fractions above and below the median are added to assess the impacts. As is evident, higher plastic fractions lead to increased reflectance across all spectra, while the absorption lines remain relatively unchanged. Notably, certain absorption lines are shared among single plastics such as wavelengths around 1145, 1680, 2165 nm for LDPE and EPSF, as well as 1215, 1420, and 1730 nm for HDPE and PP. The reflectance corresponding to maximum and minimum plastic fractions is also shown in Fig. 5. The spectra for the minimum plastic fraction are nearly zero in the SWIR range, indicating that water is the dominant component. The spectrum for the maximum plastic fraction shares similar absorption features with the mean spectrum, except for the mixed and weathered types. This is expected due to the varying composition of debris in these categories.

**Sediment signatures.** The impacts of suspended sediments and foams on the spectrum of clear water are investigated using the spectral angular mapper (SAM) score for each plastic type[20]. The SAM score is an angular similarity metric between two vectors. Given the two spectra $\mathbf{t} = \{t_1, t_2, …, t_b\}$ and $\mathbf{s} = \{s_1, s_2, …, s_b\}$ at $b$ wavebands, the SAM score is calculated as[80] $SAM = \cos^{-1}\left(\frac{\sum_{i=1}^{b} t_i s_i}{\sqrt{\sum_{i=1}^{b} t_i^2}\sqrt{\sum_{i=1}^{b} s_i^2}}\right)$. It should be noted that the SAM score is invariant to shifts in signal magnitude, which primarily occur as plastic coverage increases. Therefore, any increase in within-type SAM scores reflects changes in spectral shapes rather than magnitude, capturing the effects of suspended sediments and foams on the reflectance spectra due to the background influences. The mean SAM scores (Fig. 6) for each debris group between clear and turbid waters are as follows: 2.4° (HDPE), 5.5° (LDPE), 8.3° (PP), 9.3° (EPSF), and 9.9° (PET).

To offer additional insights, Fig. 7 shows the mean of plastic spectra for HDPE and PET samples with different sediment concentrations and sub-scale fractions. In these cases, it is evident that increased sediment concentration sharpens the main absorption lines but does not drastically change their positions, especially across the regions where the background reflectance rises markedly. Although prior literature[66,70] has documented that in the highly reflective segment of the clear water spectrum, the red region, the precise location of the absorption lines differ[81] due to the possible variations in dissolved optically active constituents. Nevertheless, two distinct reflectance peaks around 810 nm and 1070 nm, shown by red arrows in Fig. 7, are found to be attributed to turbid waters consistent with the previous findings[54,60]. Figure 8(a) compares only the clear and turbid water spectra – showing an increase in reflectance with rising sediment concentrations and highlighting the known major and minor signatures at 810 and 1070 nm, respectively.

The SAM score between PET (HDPE) samples in clear and highly turbid waters (i.e., $C = 313$ $mgl^{-1}$) is around 17.3° (12.6°). Thus even for HDPE with the lowest mean SAM score (Fig. 6), a high sediment concentration increases the score by more than five times. We observe that the rise in SAM scores in VNIR (i.e., $< 1000$ nm) is higher than SWIR for PET (HDPE) by 90% (3)% as the sediment concentration rises from 0 to 313 $mgl^{-1}$. This implies that higher turbidity can variably affect the spectral response of floating plastic debris across VNIR and SWIR bands depending on the plastic types and subgrid fractions.

**Foam signatures.** The mean SAM scores between foamy and clear waters are as follows: 4.2° (HDPE), 10° (PET), 12.1° (PP), 14.6° (LDPE), and 26.6° (EPSF). All the SAM scores are higher than those obtained for turbid waters counterparts. Among different plastic types, the increase in SAM scores is pronounced for PP and LDPE specimens and is the highest for EPSF samples. Figure 9 compares the mean of the collected spectra for PP and EPSF specimens in clear and foamy waters. Unlike the impacts of suspended sediments, the increase in SAM scores in the SWIR region is higher than VNIR as the score increases by 36 (700)% for PP (EPSF) due to the presence of foam.

For the PP rope samples, the background foam decreases the reflectance and flattens the absorption lines, especially over the SWIR region. This can be partly attributed to the fact that those samples were prone to be partially submerged and their reflectance can be masked by high water absorption. Experiments identify a few significant absorption lines at 756 nm, 980 nm, 1198, and 1448 nm, shown with blue arrows, which are consistent with those reported in the literature[65]. For EPSF, an additional absorption line appears at 1932 nm, resulting in the largest SAM score changes, which also manifested themselves in Fig. 8(b) that compares the clear and
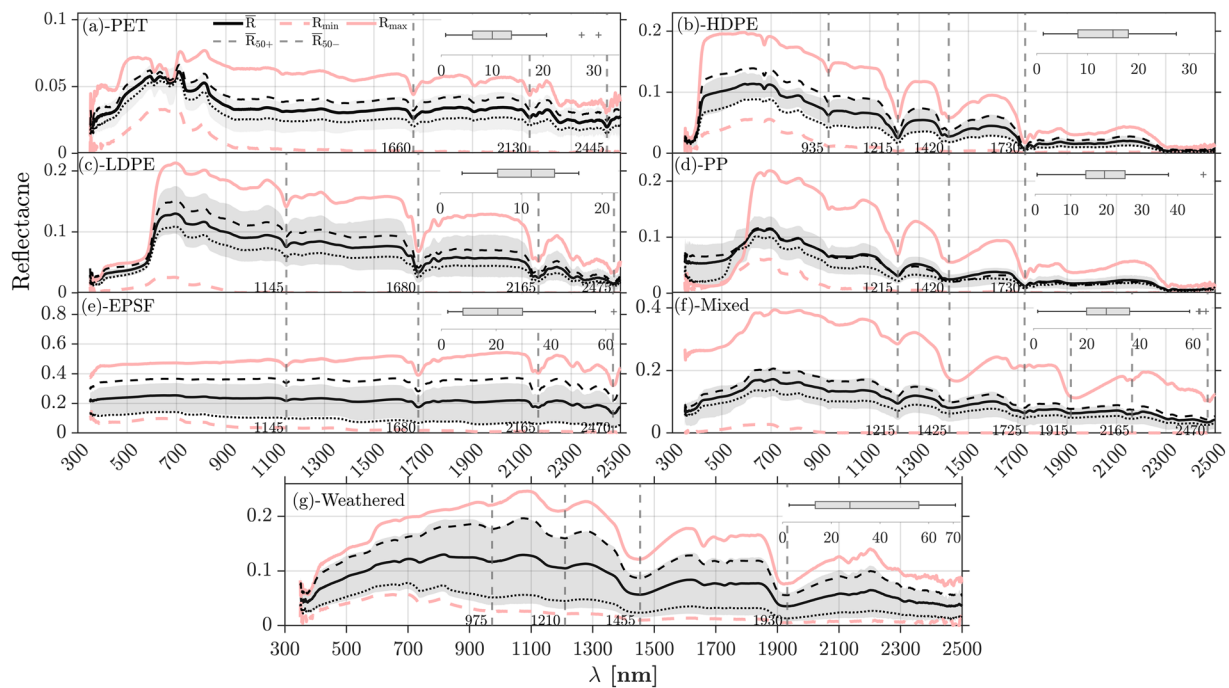
**Fig. 5** Mean reflectance $\overline{R}$ and 75% confidence bound of plastic litter spectra in the clear water along with box plot of plastic fractions for (**a**) PET, (**b**) HDPE, (**c**) LDPE, (**d**) PP, (**e**) EPSF, (**f**) Mixed, and (**g**) Weathered specimens. The mean reflectance for those samples with plastic fractions greater ($\overline{R}_{50+}$) and smaller ($\overline{R}_{50-}$) than 50% of FOV are also shown with vertical lines marking the main absorption lines. The reflectance corresponding to maximum ($R_{max}$) and minimum ($R_{min}$) plastic fractions are shown with solid and dashed lines, respectively, in pale red. The box plots show the interquartile range around the median with whiskers extending to 1.5 times the interquartile range of plastic fractions.
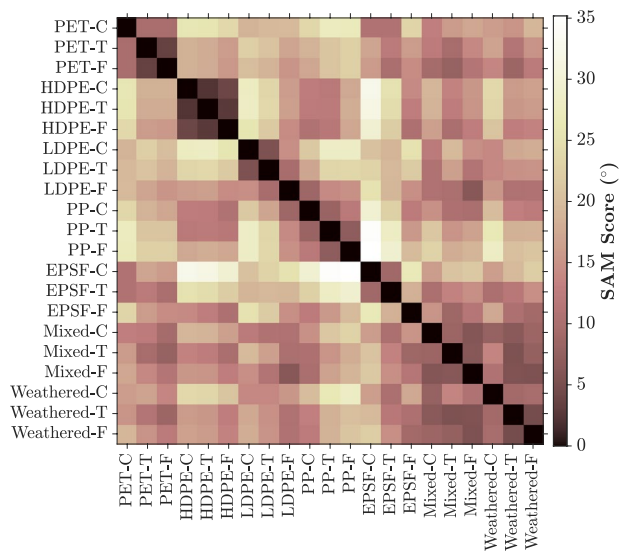


**Fig. 6** Heatmap illustrating the Spectral Angle Mapper (SAM) scores between the measured spectra for all 21 debris subgroups, comprising seven plastic types and three background water conditions including clear (C), turbid (T), and foamy (F).

foamy waters spectra. The reduction in the mean spectra in foamy water compared to clear water for both PP and EPSF specimens (Fig. 9) can be chiefly attributed to higher level wetness and submergence of plastic debris due to the induced turbulence. At the same time, the plastic fractions in foamy waters were lower than the clear water counterparts in both cases, which can cause a drop in the mean spectra.
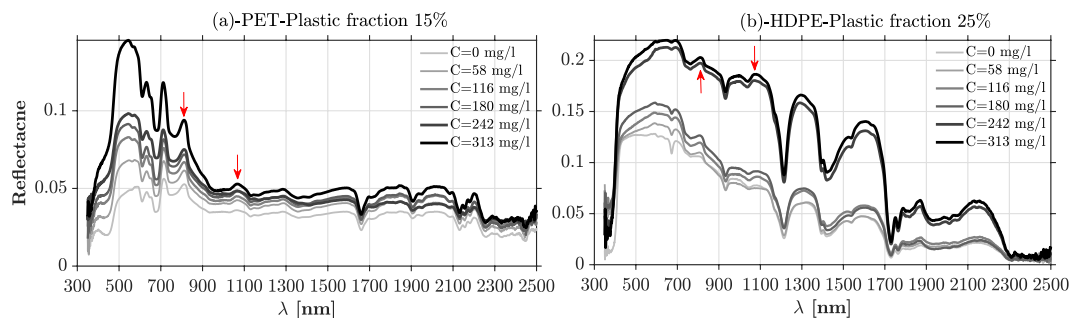
**Fig. 7** Mean of plastic litter spectra at different levels of sediment concentrations (i.e., 0, 58, 22, 180, 242 and 313 mgl⁻¹) for (**a**) PET with plastic fraction 15% and (**b**) HDPE with plastic fraction 25%. Red arrows indicate notable peak reflectance values attributed to water turbidity.
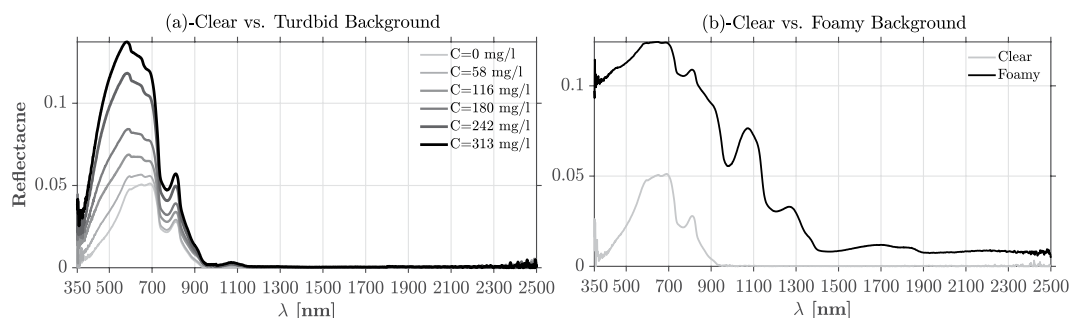


**Fig. 8** Comparison of spectra for (**a**) clear versus turbid waters with varying sediment levels and (**b**) clear versus foamy water.
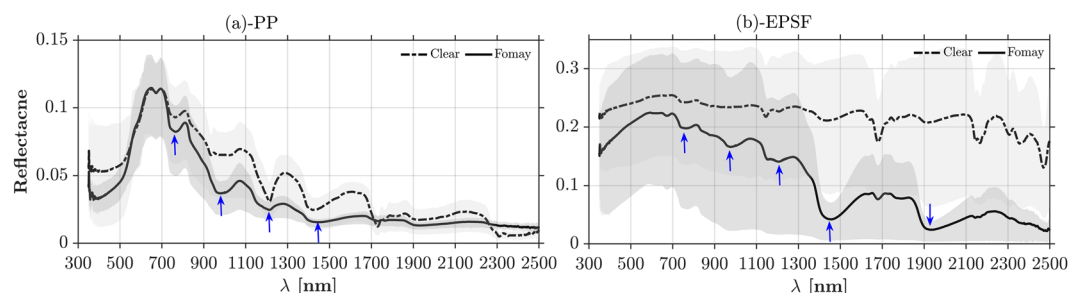


**Fig. 9** Mean and 75% confidence bound of plastic litter spectra in the foamy (solid line) and clear water (dashed line) for (**a**) PP and (**b**) EPSF. Blue arrows indicate notable absorption lines attributed to the presence of water foams.

## Technical Validation

Two complementary approaches are pursued for the technical validation of the presented dataset. First, we conduct a cross-comparison with existing public datasets obtained under different background conditions. Second, a classification machine learning algorithm is used to demonstrate that the collected data sets capture cohesively the spectral signatures of plastic debris and background waters.

**Cross comparisons and assessments.** Figure 10 compares the mean of the collected spectra in the clear waters for single virgin plastics (i.e., PET, LDPE, and PP) with analogous re-scaled reflectance spectra of ocean dry plastics, obtained from a publicly available dataset[33,36,54]. Since the mean values of spectra are highly dependent on the distance to the reference panel and illumination intensity values at the time of calibration, we rescaled the previously published spectra to match the same mean and variance through z-score normalization. The published spectra for PET, LDEP, and PP were collected over a water tank with no natural water flows, where debris samples were placed at different depths[33,54]. The spectra were measured through an 8° foreoptic FOV and lacked any information about the impacts of water foams and the sub-grid fractions of plastic debris. For HDPE, the published spectra[36] were also collected in a tank but with different nadir angles. We selected those measurements with a nadir viewing angle of $\theta = 0$.
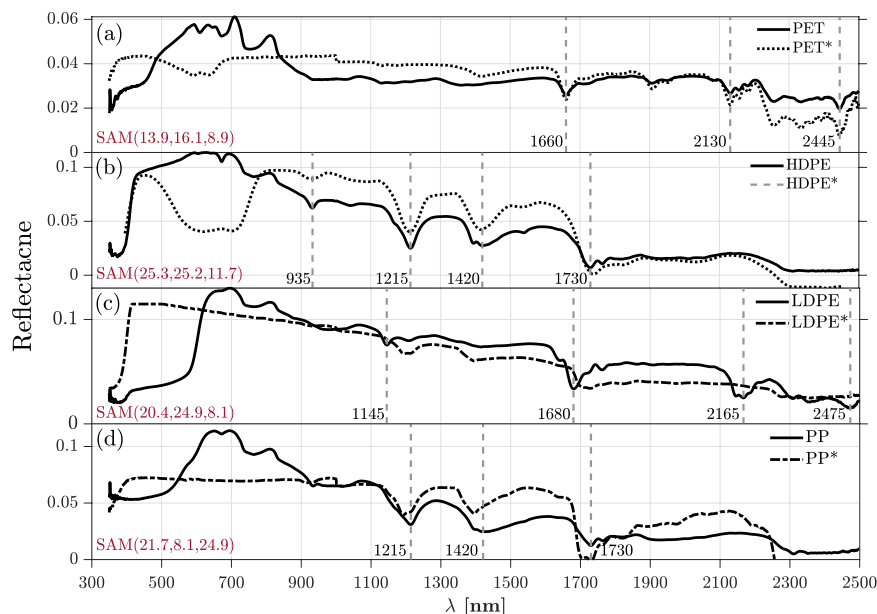
**Fig. 10** Comparison of the mean spectra of plastic litter in clear water with analogous dry spectra from[33] and[36] shown with dashed lines and asterisks, for (**a**) PET, (**b**) HDPE, (**c**) LDPE, and (**d**) PP. The red-ordered triples represent the SAM scores between the new and previously published spectra across the entire spectrum, VNIR, and SWIR regions, respectively.

In the SWIR region, comparisons demonstrate the consistency of the new data set with the published one in terms of the identified absorption lines with discrepancies of less than 50 nm. The lowest and the highest discrepancies are between the PET and LDPE samples. These differences are expected concerning differences in acquisition systems, composition of background waters, sample characteristics, and preparation techniques across the two datasets. The SAM scores between the measured spectrum and their counterparts from the literature are also computed and reported in Fig. 10. The computations are conducted for the entire spectrum including VNIR and SWIR. As expected, on average the SAM scores in SWIR are lower than VNIR by 50%. For EPSF samples, the presented results are also consistent with those provided in Fig. 4(a) in[31].

However, the newly provided data diverges from those in[33,54] over the VIS and NIR regions. The difference in the VIS region is expected due to the varying specimens' colors. However, the bed materials, variability of the optically active constituents of the background water, and random distribution of the plastic debris with various area fractions explain the differences in the NIR bands. As shown in Fig. 5, the median of subgrid plastic fractions in the provided data set varies from 10 to 28% in the FOV, implying that the water signatures can be a primary factor influencing the differences in the shape of the spectra. This is evident in Fig. 9, where, unlike the previously published data, the spectra in the new dataset peak in the red region 600–700 nm. The produced data also show peak reflectance values around 810 and 1070 nm, capturing the turbid water signatures, as explained previously.

**Cross validation using machine learning.** The second validation approach tests whether the collected dataset provides coherent reflectance signatures and contains generalized predictive information. To that end, the dataset of single virgin plastics is split into training (70%) and testing (30%) sets. The eXtreme Gradient Boosting (XGboost)[82,83] is used to train a sequential ensemble of decision trees as a classifier to learn the signatures associated with the occurrence and type of the debris. For the training, the learning rate and maximum depth are set to 0.05 and 13, respectively. Through grid search, it was determined that applying regularization does not enhance the classification performance.

The confusion matrix (Fig. 11) for the test data set shows that the recall or probability of detection (last row) and the precision (last column) exceed 90% for all plastic types in the test set with an accuracy (lower diagonal element) greater than 94%. In most plastic types, except PP and PET samples, the recall is smaller than precision, indicating that the classifier is more vulnerable to higher false negative than false positive rates. The low precision in PP detection can be attributed to the fact that ropes could be partially submerged within the FOV and thus their reflectance signals can be easily masked due to high water absorption.

The largest (smallest) false negative rate (i.e., type-II error) is around 8 (1.8)% for EPSF (PET) specimens. A low false negative rate for a class often indicates that the within-class spectra are similar but remain highly dissimilar to the other class spectra, which is evident for PET samples. On the other hand, as previously discussed, water foams affect the EPSF spectra markedly resulting in a high within-class dissimilarity with a SAM score of 26.6°. It's important to highlight that most false negative rates for EPSF samples were found within the PP classes. This is primarily because the SAM score between the EPSF and PP types is smaller than the with-class SAM score for the EPSF samples (Fig. 6).

**Fig. 11** Confusion matrix for classification of single virgin plastics using XGBoost decision tree for a test data set as explained in the text. The blue values in the last row (column) represent the recall (precision). The accuracy is reported in the lower diagonal element of the matrix.

## Usage Notes

We have included the scripts for data reading and visualization in the GitHub shared repository, organized into two folders: MATLAB_Codes and Python_Codes, as outlined below:

1. *MatlabDemo.m*: This MATLAB script loops through all the NetCDF files stored in the data_NetCDF folder and enables visualization of all 21 subgroups. Based on the user identification of the observation number (ON), the software shows the RGB image, labeled image, the spectra, and ancillary information related to measurement conditions (e.g., background condition, plastic fraction).
2. *PythonDemo.ipynb*: A Python script similar to the explained MATLAB script is also provided.
3. *Statistical_analysis_Xgboost.ipynb*: This Jupyter Notebook Python script offers additional visualizations including a bar plot illustrating the percentage of measurements categorized by material and background waters as well as the distributions and box plots of reflectance across 10 random wavelengths. Additionally, it presents the spectrum of each sampled group in the clear, turbid, and foamy waters together with the SAM scores. Finally, it provides codes utilizing the XGBoost library in Python for supervised classification of the measured spectra and computation of the confusion matrix.

We additionally provided a tabular presentation of the data containing all 2078 measured spectra, stored in both *.mat (for MATLAB) and *.feather (for Python). The feather format in Python allows efficient reading and writing of the data frame. This tabular data includes the plastic fractions and identifiers representing the 21 subgroups, explained in the Readme.txt file. The library can be deployed for exploring the spectral behavior of different plastic litter with various background conditions and applying different machine-learning techniques.

It must be emphasized that the presented dataset does not account for other floating materials, such as macroalgae, cyanobacteria, and CDOM, which can be found in large river systems. Future research could investigate the complexities related to the presence of these optically active constituents. Further characterization of the effects of various types and ages of weathered plastics will also enhance our understanding of plastic debris signature.

## Code availability

The software codes for working with the database in MATLAB and Python are accessible on GitHub at https://github.com/olyae001/Hyperspectral_reflectance_library or https://github.com/aebtehaj/Hyperspectral_reflectance_library.

## References

1. Mihai, F.-C. *et al.* Plastic pollution in marine and freshwater environments: abundance, sources, and mitigation. In *Emerging Contaminants in the Environment*, 241–274 (Elsevier, 2022).
2. Van Emmerik, T. & Schwarz, A. Plastic debris in rivers. *Wiley Interdisciplinary Reviews: Water* **7**, e1398 (2020).

10

3. McIlgorm, A., Campbell, H. F. & Rule, M. J. The economic cost and control of marine debris damage in the asia-pacific region. *Ocean & coastal management* **54**, 643–651 (2011).

4. UNEP, U. Year book 2014 emerging issues update. *United Nations Environment Programme, Nairobi, Kenya* (2014).

5. Imhof, H. K. *et al.* Variation in plastic abundance at different lake beach zones-a case study. *Science of the Total Environment* **613**, 530–537 (2018).

6. Wendt-Potthoff, K. *et al. Monitoring plastics in rivers and lakes: Guidelines for the harmonization of methodologies* (United Nations Environment Programme, 2020).

7. Blume, S. *et al.* Advances in remote sensing of plastic waste. Tech. Rep., Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH https://www.giz.de/en/worldwide/93799.html (2023).

8. Dris, R. *et al.* Beyond the ocean: contamination of freshwater ecosystems with (micro-) plastic particles. *Environmental chemistry* **12**, 539–550 (2015).

9. Blettler, M. C., Abrial, E., Khan, F. R., Sivri, N. & Espinola, L. A. Freshwater plastic pollution: Recognizing research biases and identifying knowledge gaps. *Water research* **143**, 416–424 (2018).

10. Lebreton, L. C. *et al.* River plastic emissions to the world's oceans. *Nature communications* **8**, 15611 (2017).

11. Jia, T. *et al.* Deep learning for detecting macroplastic litter in water bodies: a review. *Water Research* **231**, 119632 (2023).

12. Van Emmerik, T. *et al.* A methodology to characterize riverine macroplastic emission into the ocean. *Frontiers in Marine Science* **5**, 372 (2018).

13. González-Fernández, D. *et al.* Floating macrolitter leaked from europe into the ocean. *Nature Sustainability* **4**, 474–483 (2021).

14. Van Emmerik, T. H. *et al.* River plastic transport and deposition amplified by extreme flood. *Nature Water* **1**, 514–522 (2023).

15. Hardesty, B. D., Lawson, T., van der Velde, T., Lansdell, M. & Wilcox, C. Estimating quantities and sources of marine debris at a continental scale. *Frontiers in Ecology and the Environment* **15**, 18–25 (2017).

16. Grøsvik, B. E. *et al.* Assessment of marine litter in the barents sea, a part of the joint norwegian–russian ecosystem survey. *Frontiers in Marine Science* **5**, 72 (2018).

17. Ryan, P. G., Moore, C. J., Van Franeker, J. A. & Moloney, C. L. Monitoring the abundance of plastic debris in the marine environment. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 1999–2012 (2009).

18. Salgado-Hernanz, P. M. *et al.* Assessment of marine litter through remote sensing: recent approaches and future goals. *Marine Pollution Bulletin* **168**, 112347 (2021).

19. Geraeds, M., van Emmerik, T., de Vries, R. & bin Ab Razak, M. S. Riverine plastic litter monitoring using unmanned aerial vehicles (uavs). *Remote Sensing* **11**, 2045 (2019).

20. Tasseron, P. F., Schreyers, L., Peller, J., Biermann, L. & van Emmerik, T. Toward robust river plastic detection: Combining lab and field-based hyperspectral imagery. *Earth and Space Science* (2022).

21. Veettil, B. K., Quan, N. H., Hauser, L. T., Van, D. D. & Quang, N. X. Coastal and marine plastic litter monitoring using remote sensing: A review. *Estuarine, Coastal and Shelf Science* 108160 (2022).

22. Park, Y.-J., Garaba, S. P. & Sainte-Rose, B. Detecting the great pacific garbage patch floating plastic litter using worldview-3 satellite imagery. *Optics Express* **29**, 35288–35298 (2021).

23. Topouzelis, K., Papageorgiou, D., Suaria, G. & Aliani, S. Floating marine litter detection algorithms and techniques using optical remote sensing data: A review. *Marine Pollution Bulletin* **170**, 112675 (2021).

24. Goddijn-Murphy, L., Peters, S., Van Sebille, E., James, N. A. & Gibb, S. Concept for a hyperspectral remote sensing algorithm for floating marine macro plastics. *Marine pollution bulletin* **126**, 255–262 (2018).

25. Karakuş, O. On advances, challenges and potentials of remote sensing image analysis in marine debris and suspected plastics monitoring. *Frontiers in Remote Sensing* **4**, 1302384 (2023).

26. Garaba, S. P. & Harmel, T. Top-of-atmosphere hyper and multispectral signatures of submerged plastic litter with changing water clarity and depth. *Optics Express* **30**, 16553–16571 (2022).

27. Ciappa, A. C. Marine plastic litter detection offshore hawai'i by sentinel-2. *Marine Pollution Bulletin* **168**, 112457 (2021).

28. Garaba, S. P. *et al.* Sensing ocean plastics with an airborne hyperspectral shortwave infrared imager. *Environmental science & technology* **52**, 11699–11707 (2018).

29. Bonifazi, G., Capobianco, G. & Serranti, S. A hierarchical classification approach for recognition of low-density (ldpe) and high-density polyethylene (hdpe) in mixed plastic waste based on short-wave infrared (swir) hyperspectral imaging. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **198**, 115–122 (2018).

30. Serranti, S., Palmieri, R., Bonifazi, G. & Cózar, A. Characterization of microplastic litter from oceans by an innovative approach based on hyperspectral imaging. *Waste Management* **76**, 117–125 (2018).

31. Goddijn-Murphy, L. & Dufaur, J. Proof of concept for a model of light reflectance of plastics floating on natural waters. *Marine pollution bulletin* **135**, 1145–1157 (2018).

32. Garaba, S. P., Acuña-Ruz, T. & Mattar, C. B. Hyperspectral longwave infrared reflectance spectra of naturally dried algae, anthropogenic plastics, sands and shells. *Earth System Science Data* **12**, 2665–2678 (2020).

33. Knaeps, E. *et al.* Hyperspectral-reflectance dataset of dry, wet and submerged marine litter. *Earth System Science Data* **13**, 713–730, https://doi.org/10.5194/ESSD-13-713-2021 (2021).

34. de Vries, R. V., Garaba, S. P. & Royer, S.-J. Hyperspectral reflectance of pristine, ocean weathered and biofouled plastics from dry to wet and submerged state. *Earth System Science Data Discussions* **2023**, 1–29 (2023).

35. Leone, G. *et al.* Hyperspectral reflectance dataset of pristine, weathered and biofouled plastics. *Earth System Science Data Discussions* **2022**, 1–24 (2022).

36. Garaba, S. P. *et al.* Concentration, anisotropic and apparent colour effects on optical reflectance properties of virgin and ocean-harvested plastics. *Journal of Hazardous Materials* **406**, 124290 (2021).

37. Biermann, L., Clewley, D., Martinez-Vicente, V. & Topouzelis, K. finding plastic patches in coastal Waters using optical Satellite Data. *Scientific RepoRtS* | 10, https://doi.org/10.1038/s41598-020-62298-z (2020).

38. Investigating Detection of Floating Plastic Litter from Space Using Sentinel-2 Imagery https://doi.org/10.3390/rs12162648 (2020).

39. Basu, B., Sannigrahi, S., Basu, A. S. & Pilla, F. remote sensing Development of Novel Classification Algorithms for Detection of Floating Plastic Debris in Coastal Waterbodies Using Multispectral Sentinel-2 Remote Sensing Imagery. https://doi.org/10.3390/rs13081598 (2021).

40. Kako, S., Morita, S. & Taneda, T. Estimation of plastic marine debris volumes on beaches using unmanned aerial vehicles and image processing based on deep learning. *Marine Pollution Bulletin* **155**, 111127 (2020).

41. Topouzelis, K., Papakonstantinou, A. & Garaba, S. P. Detection of floating plastics from satellite and unmanned aerial systems (plastic litter project 2018). *International Journal of Applied Earth Observation and Geoinformation* **79**, 175–183 (2019).

42. MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data https://doi.org/10.1371/journal.pone.0262247.g001 (2022).

43. Olyaei, M., Ebtehaj, A. & Hong, J. Optical Detection of Marine Debris using Deep Knockoff. *IEEE Trans. on Geosci. and Remote Sens. Accepted* (2022).

44. Garaba, S. P. & Park, Y.-J. Riverine litter monitoring from multispectral fine pixel satellite images. *Environmental Advances* **15**, 100451 (2024).

45. Jakovljevic, G., Govedarica, M. & Alvarez-Taboada, F. A deep learning model for automatic plastic mapping using unmanned aerial vehicle (uav) data. *Remote Sensing* **12**, 1515 (2020).

46. Maharjan, N. *et al*. Detection of river plastic using uav sensor data and deep learning. *Remote Sensing* **14**, 3049 (2022).
47. Van Lieshout, C., van Oeveren, K., van Emmerik, T. & Postma, E. Automated river plastic monitoring using deep learning and cameras. *Earth and space science* **7**, e2019EA000960 (2020).
48. Tharani, M. *et al*. Trash detection on water channels. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I* **28**, 379–389 (Springer, 2021).
49. Topouzelis, K., Papageorgiou, D., Karagaitanakis, A., Papakonstantinou, A. & Ballesteros, M. A. Plastic litter project 2019: Exploring the detection of floating plastic litter using drones and sentinel 2 satellite images. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 6329–6332 (IEEE, 2020).
50. Papageorgiou, D., Topouzelis, K., Suaria, G., Aliani, S. & Corradi, P. Sentinel-2 detection of floating marine litter targets with partial spectral unmixing and spectral comparison with other floating materials (plastic litter project 2021). *Remote Sensing* **14**, 5997 (2022).
51. Waqas, M. *et al*. Marine plastic pollution detection and identification by using remote sensing-meta analysis. *Marine Pollution Bulletin* **197**, 115746 (2023).
52. Hu, C. Remote detection of marine debris using satellite observations in the visible and near infrared spectral range: Challenges and potentials. *Remote Sensing of Environment* **259**, 112414, https://doi.org/10.1016/J.RSE.2021.112414 (2021).
53. Hu, C. Remote detection of marine debris using sentinel-2 imagery: A cautious note on spectral interpretations. *Marine Pollution Bulletin* **183**, 114082 (2022).
54. Moshtaghi, M., Knaeps, E., Sterckx, S., Garaba, S. & Meire, D. Spectral reflectance of marine macroplastics in the vnir and swir measured in a controlled environment. *Scientific Reports* **11**, 5436 (2021).
55. Garaba, S. P. & Dierssen, H. M. Hyperspectral ultraviolet to shortwave infrared characteristics of marine-harvested, washed-ashore and virgin plastics. *Earth System Science Data* **12**, 77–86, https://doi.org/10.5194/ESSD-12-77-2020 (2020).
56. Jain, S. K. & Singh, V. P. *Water resources systems planning and management* (Elsevier, 2003).
57. Ekercin, S. Water quality retrievals from high resolution ikonos multispectral imagery: A case study in istanbul, turkey. *Water, Air, and Soil Pollution* **183**, 239–251 (2007).
58. Thiemann, S. & Kaufmann, H. Lake water quality monitoring using hyperspectral airborne data—a semiempirical multisensor and multitemporal approach for the mecklenburg lake district, germany. *Remote sensing of Environment* **81**, 228–237 (2002).
59. Gholizadeh, M. H., Melesse, A. M. & Reddi, L. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors* **16**, 1298 (2016).
60. Olyaei, M. & Ebtehaj, A. Uncovering plastic litter spectral signatures: A comparative study of hyperspectral band selection algorithms. *Remote Sensing* **16**, 172 (2023).
61. Kutser, T. *et al*. Remote sensing of black lakes and using 810 nm reflectance peak for retrieving water quality parameters of optically complex waters. *Remote Sensing* **8**, 497 (2016).
62. Nicolas, J.-M., Deschamps, P.-Y. & Frouin, R. Spectral reflectance of oceanic whitecaps in the visible and near infrared: Aircraft measurements over open ocean. *Geophysical research letters* **28**, 4445–4448 (2001).
63. Kokhanovsky, A. Spectral reflectance of whitecaps. *Journal of Geophysical Research: Oceans* **109** (2004).
64. Bartlett, D., Gurganus, E. & Whitlock, C. Sea foam reflectance and influence on optimum wavelength for remote sensing of ocean aerosols (1982).
65. Dierssen, H. M. Hyperspectral measurements, parameterizations, and atmospheric correction of whitecaps and foam from visible to shortwave infrared for ocean color remote sensing. *Frontiers in Earth Science* **7**, 14 (2019).
66. Doxaran, D., Froidefond, J.-M. & Castaing, P. Remote-sensing reflectance of turbid sediment-dominated waters. reduction of sediment type variations and changing illumination conditions effects by use of reflectance ratios. *Applied Optics* **42**, 2623–2634 (2003).
67. Ritchie, J. C., Schiebe, F. R. & McHenry, J. R. Remote sensing of suspended sediments in surface waters. *Photogrammetric Engineering and Remote Sensing* **42**, 1539–1545 (1976).
68. Yang, C.-Y. & Julien, P. Y. The ratio of measured to total sediment discharge. *International Journal of Sediment Research* **34**, 262–269 (2019).
69. Dorrell, R. M. & Hogg, A. J. Length and time scales of response of sediment suspensions to changing flow conditions. *Journal of Hydraulic Engineering* **138**, 430–439 (2012).
70. Knaeps, E. *et al*. A swir based algorithm to retrieve total suspended matter in extremely turbid waters. *Remote Sensing of Environment* **168**, 66–79 (2015).
71. Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* **9**, 62–66 (1979).
72. Arifin, A. Z. & Asano, A. Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern recognition letters* **27**, 1515–1521 (2006).
73. Kaur, A. & Kranthi, B. Comparison between ycbcr color space and cielab color space for skin color segmentation. *International Journal of Applied Information Systems* **3**, 30–33 (2012).
74. Nie, F., Wang, Y., Pan, M., Peng, G. & Zhang, P. Two-dimensional extension of variance-based thresholding for image segmentation. *Multidimensional systems and signal processing* **24**, 485–501 (2013).
75. Sezgin, M. & Sankur, BL Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging* **13**, 146–168 (2004).
76. Güneş, A., Kalkan, H. & Durmuş, E. Optimizing the color-to-grayscale conversion for image classification. *Signal, Image and Video Processing* **10**, 853–860 (2016).
77. Olyaei, M., Ebtehaj, A. & Ellis, C. R. A hyperspectral reflectance database of plastic debris for river ecosystems https://doi.org/10.5281/zenodo.13377060 (2024).
78. Urquhart, E. A. & Schaeffer, B. A. Envisat meris and sentinel-3 olci satellite lake biophysical water quality flag dataset for the contiguous united states. *Data in brief* **28**, 104826 (2020).
79. Hatchell, D. Analytical spectral devices, inc.(asd) technical guide (1999).
80. Kruse, F. A. *et al*. The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment* **44**, 145–163 (1993).
81. Knaeps, E. *et al*. The seaswir dataset. *Earth System Science Data* **10**, 1439–1449 (2018).
82. Chen, T. *et al*. Xgboost: extreme gradient boosting. *R package version 0.4-2* **1**, 1–4 (2015).
83. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

## Acknowledgements

## Author contributions
M.O. designed the study, collected the specimens, managed and analyzed the data files, and wrote the manuscript. A.E. designed the study and edited the manuscript. C.E. designed and built the data acquisition system.

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to M.O.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.