

## Article

# Framework for Regional to Global Extension of Optical Water Types for Remote Sensing of Optically Complex Transitional Water Bodies

Elizabeth C. Atwood <sup>1,\*</sup> , Thomas Jackson <sup>1,2</sup> , Angus Laurenson <sup>1</sup> , Bror F. Jönsson <sup>1,3</sup> , Evangelos Spyarakos <sup>4</sup> , Dalin Jiang <sup>4</sup> , Giulia Sent <sup>5</sup> , Nick Selmes <sup>1</sup> , Stefan Simis <sup>1</sup> , Olaf Danne <sup>6</sup> , Andrew Tyler <sup>4</sup>  and Steve Groom <sup>1</sup>

<sup>1</sup> Earth Observation Science and Applications, Plymouth Marine Laboratory, Plymouth PL1 3DH, UK; thomas.jackson@eumetsat.int (T.J.); anla@pml.ac.uk (A.L.); bror.jonsson@unh.edu (B.F.J.); nse@pml.ac.uk (N.S.); stsi@pml.ac.uk (S.S.); sbg@pml.ac.uk (S.G.)

<sup>2</sup> Climate Services Group, European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), 64295 Darmstadt, Germany

<sup>3</sup> Ocean Process Analysis Lab, University of New Hampshire, Durham, NH 03824, USA

<sup>4</sup> Earth and Planetary Observation Sciences (EPOS), Department of Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, UK; evangelos.spyarakos@stir.ac.uk (E.S.); dalin.jiang@stir.ac.uk (D.J.); a.n.tyler@stir.ac.uk (A.T.)

<sup>5</sup> MARE—Marine and Environmental Science Centre, ARNET—Aquatic Research Network, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

<sup>6</sup> Brockmann Consult GmbH, 21029 Hamburg, Germany; olaf.danne@brockmann-consult.de

\* Correspondence: liat@pml.ac.uk; Tel.: +44-(0)1752-633110

**Abstract:** Water quality indicator algorithms often separate marine and freshwater systems, introducing artificial boundaries and artifacts in the freshwater to ocean continuum. Building upon the Ocean Colour- (OC) and Lakes Climate Change Initiative (CCI) projects, we propose an improved tool to assess the interactions across river–sea transition zones. Fuzzy clustering methods are used to generate optical water types (OWT) representing spectrally distinct water reflectance classes, occurring within a given region and period (here 2016–2021), which are then utilized to assign membership values to every OWT class for each pixel and seamlessly blend optimal in-water algorithms across the region. This allows a more flexible representation of water provinces across transition zones than classic hard clustering techniques. Improvements deal with expanded sensor spectral band-sets, such as Sentinel-3 OLCI, and increased spatial resolution with Sentinel-2 MSI high-resolution data. Regional clustering was found to be necessary to capture site-specific characteristics, and a method was developed to compare and merge regional cluster sets into a pan-regional representative OWT set. Fuzzy clustering OWT timeseries data allow unique insights into optical regime changes within a lagoon, estuary, or delta system, and can be used as a basis to improve WQ algorithm performance.

**Keywords:** c-means; Tagus and Sado Estuaries; Plymouth Sound; Danube Razelm–Sinoe Lagoon System; Venice Lagoon; Curonian Lagoon; Elbe Estuary; water quality monitoring; water-leaving reflectance; multispectral



**Citation:** Atwood, E.C.; Jackson, T.; Laurenson, A.; Jönsson, B.F.; Spyarakos, E.; Jiang, D.; Sent, G.; Selmes, N.; Simis, S.; Danne, O.; et al. Framework for Regional to Global Extension of Optical Water Types for Remote Sensing of Optically Complex Transitional Water Bodies. *Remote Sens.* **2024**, *16*, 3267. <https://doi.org/10.3390/rs16173267>

Academic Editor: SeungHyun Son

Received: 12 June 2024

Revised: 19 August 2024

Accepted: 27 August 2024

Published: 3 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Optically complex coastal and inshore waters represent some of the most productive aquatic environments worldwide [1] and contribute significantly to global biogeochemical cycles [2], despite their relatively small area compared with the open oceans. Around 37% of the world's population, circa 3 billion people, live within 100 km of the coast, relying on resources therefrom for human consumption, food production, industry including terrestrial and coastal fisheries or aquaculture, as well as nature and recreation services. Coastal systems are also one of the marine areas most highly impacted by human activities [3,4]; thus, there emerges a natural interest to monitor these regions, over large areas, under

changing human demographic and climate conditions. This creates a strong motivation to provide operational monitoring over large coastal areas, but efforts to produce a seamless product from inland waters to the open ocean are hindered by algorithm limitations. Transitional waters (defined as estuaries, lagoons, deltas, and river mouths) introduce particular challenges for Earth Observation (EO)-based optical water quality indicator algorithm development. Water reflectance properties are influenced by coastal processes such as river inputs, sediment resuspension (due to forcing from wind or tidal movement), bottom reflectance, adjacency to land effects, and algal or cyanobacterial blooms. The optical complexity of inland waters exceeds that of open marine systems [5] and it is reasonable to assume that transitional waters cover at least as much of the complexity represented within both these environments. Local and regional algorithm development can achieve higher performance for predicting water quality indicators as compared to global algorithms, such as suspended particulate matter (SPM) or chlorophyll-a (Chl-a), but it is difficult to determine the applicability of these algorithms for different areas and/or time periods [6,7]. Some have suggested that it is not feasible to develop a universal bio-optical parameter algorithm that performs optimally in phytoplankton, SPM, and colored dissolved organic matter (CDOM)-dominated waters [8]. Using disparate regional algorithms for these transitional water systems can produce artifacts along mixing boundaries, which hinder intra-system comparison and pan-regional monitoring efforts. Further artifacts can be introduced from errors in atmospheric correction processing, which may be tuned differently for fresh- or saltwater systems. Optical water quality indicators, hereafter referred to simply as water quality (WQ), have largely been developed separately for freshwater, saltwater, or transitional waters systems. Identifying optically similar water types, with different associated optimal WQ algorithms, across these regions provides a pathway to address these difficulties.

A set of optical water types (OWT) can be regarded as a simplified representation of the spectral variability within the training data taken from a particular system. An OWT set can either be used in its own right to determine trends in the variation of the study system or provide a basis for WQ algorithm improvement. It is well-established to utilize water spectral typologies as a basis for the delineation of water parcels with distinct properties [5,7,9,10]. The creation of an OWT set can be broken into four steps: (i) the determination of a representative and balanced training dataset which sufficiently captures the space/time variability of the study system, (ii) focusing cluster formation through the normalization of the training data prior to clustering, (iii) the selection and application of a cluster formation optimization routine, and (iv) the membership assignment of novel data to those clusters. The question of sufficient representative training data can be challenging, often as a result of lacking a priori awareness of short-term or small-scale events within the study system. There are two general schools of thought regarding the appropriate source of training data for clustering, discussed in more detail in Methods Section 2.4. Following training data selection, the normalization of training data impacts which spectral features the cluster formation process optimally separates. Mélin et al. [6] discuss how integral normalization of the training data shifts the focus of the OWT cluster set distribution from the separation of particulate gradient concentrations (mainly impacting spectral amplitude) to absorption parameters (mainly impacting spectral shape). Eleveld et al. [11] compared cluster set performance with and without integral normalization, concluding that very dark, high-CDOM-absorbing lakes were better represented with integral normalization while shallow high-reflecting lakes with high sediment load were better represented without.

OWT cluster sets can be determined using techniques including hard clustering k-means [8,12–15] and the related ISODATA method [6], soft clustering fuzzy c-means [11,16,17] and the hierarchical Ward's algorithm [18–20], the Gordon model [21], max-classification [22], and self-organizing maps [23]. Once clusters have been identified from a training dataset, novel data are assigned memberships to clusters based on a variety of metrics and sometimes irrespective of whether a fuzzy clustering scheme was used in the cluster formation optimization process. In the Ocean Colour-Climate Change Initiative

(OC-CCI), overall error, bias, and relative error were reduced for open-ocean WQ products using a blended algorithm approach based on OWT fuzzy c-means classification [7], which is now part of the operational OC-CCI Chl-a processing chain. Significant improvement in retrieval accuracy (25%) for inland Chl-a products was achieved using retuned algorithms with parameters optimized for each OWT [5,24], and is currently used in the operational Lakes-Climate Change Initiative (Lakes-CCI) processing chain. While there are an ever-increasing number of OWT cluster sets being published, relatively little comparison between sets has occurred to examine the similarity or dissimilarity of classes, for instance, as with a unique OWT class not captured within another set.

The CERTO (Copernicus Evolution—Research for harmonised and Transitional water Observation) project focused on closing remote sensing knowledge gaps in transitional waters and improving coastal water quality monitoring through the harmonization of EO-derived WQ products. In this study, we focus on the following hypotheses within the context of CERTO: (1) fuzzy clustering offers valuable site-specific insights into transitional water systems, (2) it is feasible to compare and merge cluster sets from various sites to generate a pan-regional representative cluster set, and (3) the methods employed in this study can feasibly be extrapolated to other less explored regions. A global OWT cluster set which retains sufficient regional specificity would offer an advancement in water quality data collection and interpretation across diverse aquatic environments, thus helping to remedy the monitoring gap of coastal and transitional marine systems.

## 2. Materials and Methods

### 2.1. Overview

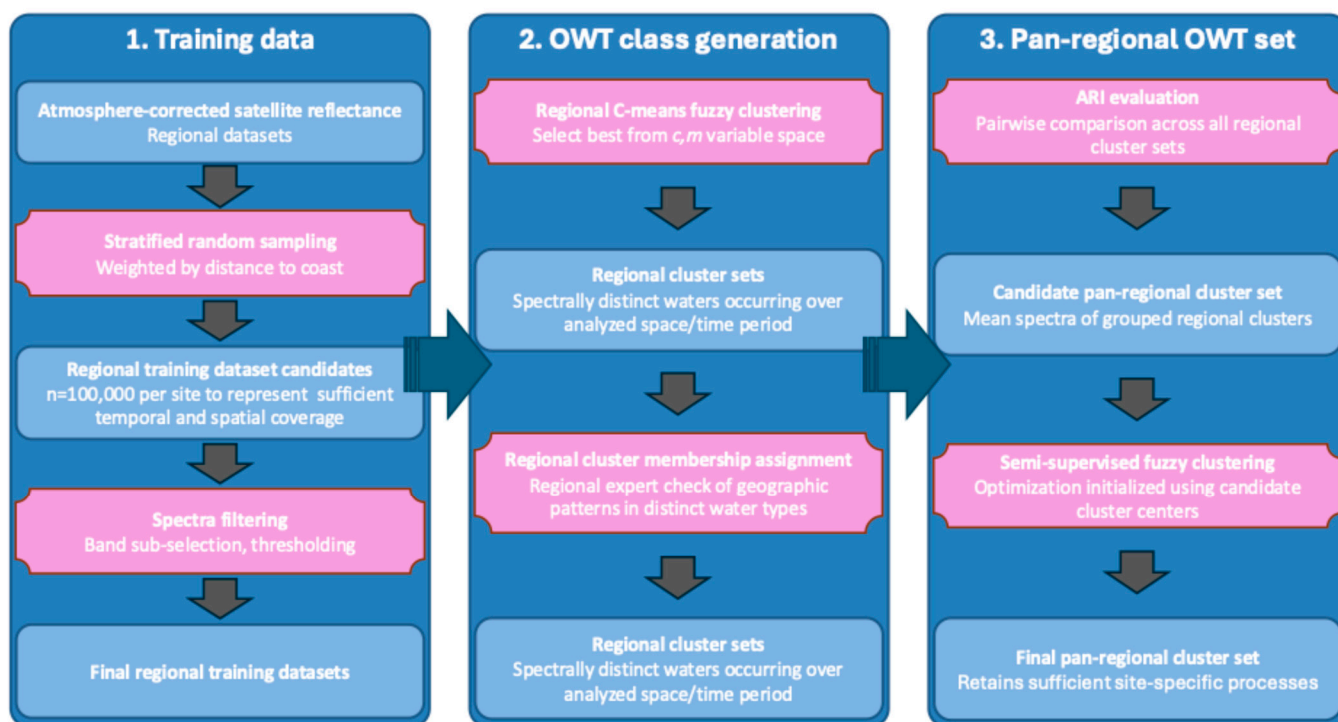
Figure 1 provides an overview of the methodological steps for regional and pan-regional OWT steps. In the first step, representative training data for each study site were compiled, taking into account sufficient temporal and spatial coverage to capture small-area and rare events in each region. In the second step, these training data were used to generate OWT classes, or clusters, with a fuzzy clustering method to represent spectrally distinct waters occurring over the analyzed space/time period for each site. Using the identified OWT classes, membership values were assigned to each cluster for all water pixels within a satellite image, highlighting through high membership values where geographically a particular spectrally distinct water type is dominant. Spatiotemporal patterns in OWT coverage were checked with regional site leads to determine if regional OWT clusters represented site-specific events or annual flux patterns. In a final step, regional clusters were compared pairwise to build a wider representative cluster set, which was then used as an input to a semi-supervised cluster optimization to produce a pan-regional cluster set able to sufficiently capture site-specific processes.

### 2.2. Study Areas

The six sites (Figure 2) consisted of three delta/lagoon systems and three estuary systems. The former group includes the Curonian Lagoon stretching between Lithuania and Russia bordering the Baltic Sea, the Danube Razelm–Sinoe Lagoon System in Romania flowing into the Black Sea, and the Venice Lagoon located in Italy connected to the north-western Adriatic Sea. The latter group comprises the Elbe Estuary in Germany flowing into the German Bight, the Tagus and Sado Estuaries located in Portugal, and the Tamar Estuary connected to Plymouth Sound in the UK.

For brevity, the Tagus and Sado Estuaries are used as the example system within the main text, while site overviews for the other areas are provided in the Supplemental Material. The Tagus Estuary covers 34,000 hectares, representing the largest estuary system in Western Europe, located within the city limits of Lisbon, Portugal. The Tagus River originates in Spain and eventually merges with the Atlantic Ocean after flowing through Lisbon. It is a semi-diurnal mesotidal system with an average tidal range at the seaward side of 2.4 m [25]. The estuary is characterized by a long deep inlet channel reaching depths of about 40 m and an inner bay with an average depth of 7 m. Tidal flats, occupying about 40%

of the estuary's total area [26], provide important wintering habitat for many waterfowl species. South of the Tagus, the Sado Estuary represents the second largest estuary in Portugal (circa 23,100 hectares). This system also features a shallow basin (average depth of 10 m) with a maximum depth of 50 m in the inlet channel used for navigation. The well-mixed estuary is subject to semi-diurnal mesotidal tides, with amplitude varying between 1.3 m during neaps and 3.5 m during springs [27]. Water circulation is mainly tidally driven given the low flow rate of the Sado river ( $0.7 \text{ m}^3/\text{s}$  in summer to  $60.0 \text{ m}^3/\text{s}$  in winter) [28].



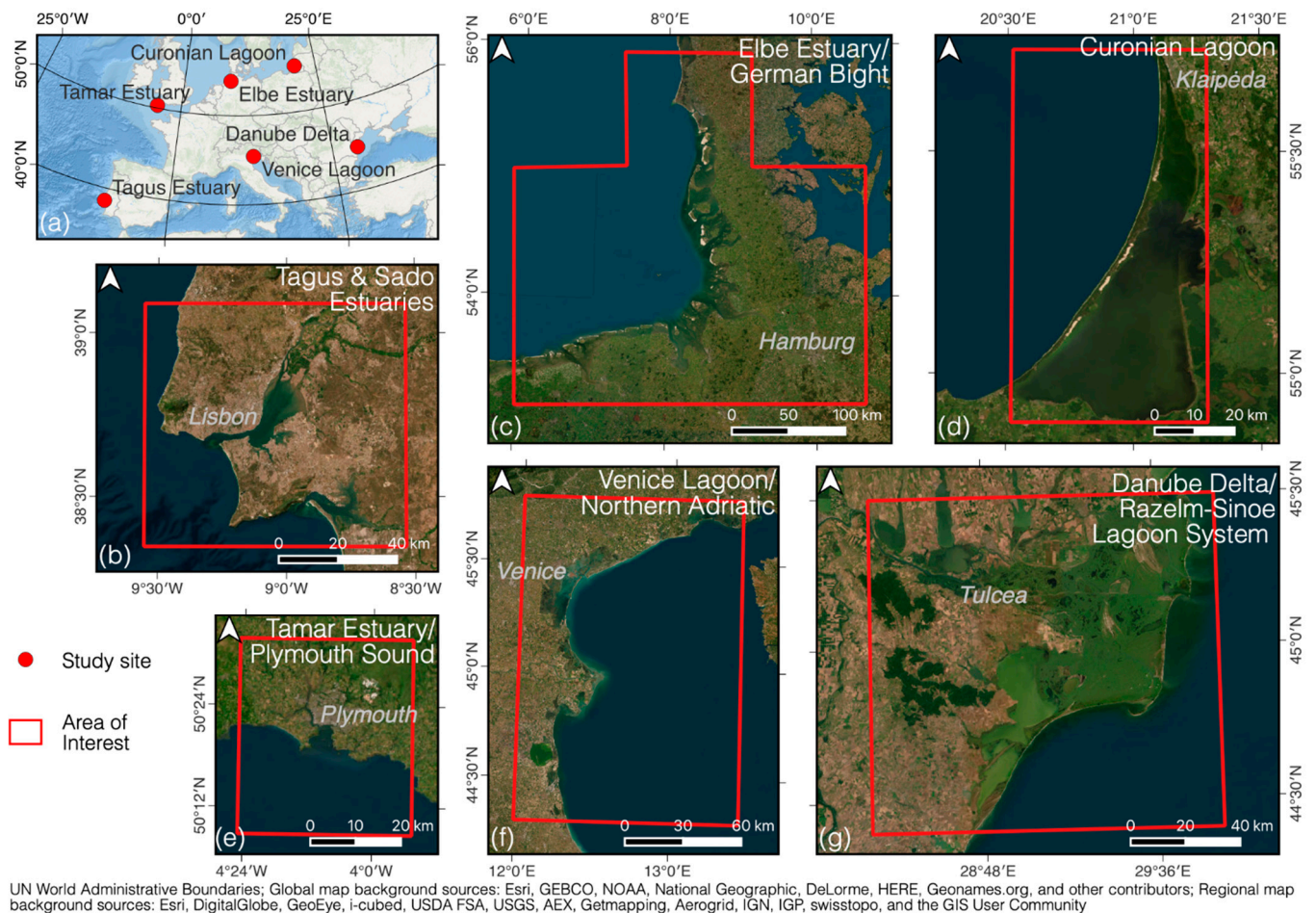
**Figure 1.** A flow chart summarizing the methodological approach to develop a framework for the regional to global extension of optical water type (OWT) classes.

### 2.3. Earth Observation Data

EO datasets included acquisitions by the Ocean and Land Colour Instrument (OLCI; Manufacturer: Thales Alenia Space, Cannes, France) onboard Sentinel-3 platforms (3A and 3B) and the MultiSpectral Instrument (MSI; Manufacturer: Airbus Defense and Space, Paris, France) onboard Sentinel-2 platforms (2A and 2B), both multi-satellite missions within the European Copernicus Programme. Sentinel-3A was launched in February 2016 and was joined by Sentinel-3B in April 2018. The OLCI sensor is an along-track (push broom) scanner providing 21 spectral bands from the optical to the near-infrared (400 nm to 1020 nm). Data have spatial resolution down to 300 m and are operationally managed by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). The Sentinel-3 constellation provides a revisit time of less than two days at the equator for OLCI data. Sentinel-2A was launched in June 2015, followed by Sentinel-2B in March 2017. The MSI sensor is also an along-track scanner, providing 13 spectral bands ranging from the optical to the short-wave infrared (490 nm to 1370 nm). Data are provided with spatial resolutions of 10, 20, and 60 m, dependent upon the band. High-resolution imagery from MSI is designed to complement SPOT-5 and Landsat-8/9 missions, with a core focus being land classification but, owing to key wavebands in the near and shortwave infrared, is also used for water remote sensing purposes, in particular smaller inland water bodies [29–31]. The Sentinel-2 constellation provides a revisit time of 2–3 days at mid-latitudes. Data from both sensors were atmospherically corrected using the *Calimnos* processing chain [32]. The



first half of the processing chain used Idepix (Identification of Pixel properties, in SNAP 8) masking to remove land, cloud, and spurious data points with high uncertainty [33]. In the second half of the processing chain, Polymer v4.15 is implemented for atmospheric correction [34,35], selected for its relatively high performance with Sentinel products [36,37], optimized for inland water remote sensing [33,38].



**Figure 2.** Locations of the six sites across Europe (a), together with true color images of each site showing study area bounds (red box) for the (b) Tagus and Sado Estuaries, (c) Elbe Estuary and German Bight, (d) Curonian Lagoon, (e) Tamar Estuary and Plymouth Sound, (f) Venice Lagoon and northwestern Adriatic Sea, and (g) the Danube Delta and Razelm–Sinoué Lagoon System. The largest population center closest to the transitional water system for each site is indicated (gray text).

#### 2.4. Building Training Data

The appropriate sourcing of training data for clustering is an area of active discussion, which one can separate into two positions: (i) in situ hyperspectral data, and (ii) satellite reflectance data. In this study, we use satellite reflectance data alone for model training, given the following considerations. In situ hyperspectral data (i) will be least impacted by corrections for atmospheric effects and provide a basis through convolution plus inverse atmospheric correction modeling to translate developed hyperspectral cluster sets to the multispectral band set of a particular satellite sensor. Fuzzy clustering efforts in coastal and inland waters to date, such as those in [5,20,39], used in situ hyperspectral data as training data. However, in situ data are often limited in their representation of the full variability over space and time of the entire study system [10], and thus run the risk of being less representative of system variability compared to a multi-year satellite acquisition database. Furthermore, the applicability of use on novel satellite sensor data is highly dependent on atmospheric correction model performance. Using satellite reflectance data

(ii) as the basis for cluster formation provides the second position. The quality of these reflectance data is tied to the sensor–atmospheric correction combination, but consistent errors will be accounted for in the cluster formation optimization step. Thus, the acquired cluster set should ideally be implemented on the same processed data as that on which the model has been trained. The capture of spatial and temporal variability in the training data, at least to the level possible with that particular sensor, is improved due to the comprehensive nature of satellite imagery coverage (e.g., a single image can cover the entire study system). The convolution of a satellite OWT set to another sensor may be possible, but results would be questionable. Where cluster sets have been generated and applied to multi-sensor records, such as within OC-CCI, the reflectance data are harmonized and homogenized across sensors before sampling and cluster creation, meaning that the data used for training remains consistently processed with the data to be classified. In this study, following (ii) as utilized in [10], we created clusters for a particular sensor and atmospheric correction combination that were only used with data from that same sensor plus atmospheric correction processing.

To capture interannual variability for each study site, per site multi-year datasets (2016–2021) were subsampled both temporally and spatially to balance processing efficiency during model training against sufficient coverage to capture rare or relatively small-area events (such as a cyanobacteria bloom or a wind-induced sediment resuspension event). Building a robust, representative training dataset benefits from extensive, long-term knowledge of the study site, and thus continual communication with the CERTO regional site leads during training dataset creation was of central importance. Based on these criteria, target training data size was set to 100,000 sample points per region. Spatial sampling was designed with a stratified random approach, with sampling density increasing closer to the coastline and decreasing further from shore. Stratified random sampling was selected to ensure that spectral diversity across the study area had balanced representation (in both time and space), with weighting the sample frequency by distance from land aimed to equalize input from smaller coastal and inland waters with that from larger ocean areas.

Winter months with low incident light levels were excluded (solar noon elevation  $< 30^\circ$  calculated using NOAA solcalc, <https://gml.noaa.gov/grad/solcalc/>, accessed on 11 June 2024, overview in Table 1). Bands from each sensor that were heavily affected by spurious atmospheric correction results were excluded, resulting in the following final band lists used in the cluster optimization: for OLCI, 400, 412, 443, 490, 510, 560, 620, 665, 674, 681, 709, 754, 779, 865, and 885 nm; for MSI, 443, 490, 560, 665, 705, 740, 783, and 865 nm. Furthermore, MSI data from the same relative orbit for each region were used to ensure that no part of the total study site was sampled more frequently than another.

**Table 1.** Subsampling and exclusion parameters used to build training dataset for each sensor/study site combination.

Study Site	Sensor	Temporal Coverage	Winter Months Excluded
Curonian Lagoon, Lithuania/Russia	OLCI	April 2016 to March 2021	October–February
	MSI	November 2016 to December 2020	
Razelm–Sinoe Lagoon System, Romania	OLCI	April 2016 to March 2021	November–January
	MSI	November 2016 to December 2020	
Elbe Estuary, Germany	OLCI	April 2016 to March 2021	October–February
	MSI	November 2016 to December 2020	
Tagus Estuary, Portugal	OLCI	April 2016 to March 2021	None
	MSI	November 2016 to December 2020	
Tamar Estuary and Plymouth Sound, UK	OLCI	April 2016 to March 2021	November–February
	MSI	November 2016 to December 2020	
Venice Lagoon, Italy	OLCI	April 2016 to March 2021	None
	MSI	November 2016 to December 2020	

### 2.5. Fuzzy Water Clustering: Scikit-Learn-Compatible Flexible Tool

To enable easy implementation of fuzzy c-means clustering and parameter optimization, we created the Fuzzy Water Clustering package in Python v3.9 ([https://github.com/CERTO-project/D4.3\\_Classification\\_toolbox](https://github.com/CERTO-project/D4.3_Classification_toolbox), accessed on 11 June 2024). It is extensible, integrates with the scikit-learn v1.0 machine learning framework and is not necessarily specific to water related data. At its heart is the c-means model, a fuzzy c-means clustering routine that can be combined with scikit-learn transformers, such as Principal Component Analysis (PCA), to form new models. Parameters for these models can be optimized using cross-validation routines from scikit-learn and their performance evaluated against several scoring metrics (Table 2). A wrapper is provided which changes the input and output of scikit-learn estimators from 2D arrays to xarray datasets, greatly improving processing efficiency. This is carried out under the assumption that each pixel is an independent measurement and each variable is a feature. In this manner, clustering models can be trained directly on opened netcdf datasets and the prediction of class membership can be performed at scale by using the `dask.array` package to address data on disk one chunk at a time.

**Table 2.** Commonly used scoring metrics for fuzzy c-means clustering.

Index	Definition *	Goal	Description
Xie–Beni [40]	$XB = J_m / \left( N \times \min_{i,j=1,\dots,C; i \neq j} \ v_i - v_j\ ^2 \right)$	min	Considers membership degree and dataset structure, measures overall average compactness and separateness of clusters.
Partition coefficient [41]	$PC = (1/N) \sum_{k=1}^C \sum_{i=1}^N u_{ki}^2$	max	Average squared membership degree to each cluster, summed across all clusters.
Modified partition coefficient [42]	$MPC = (C \times I_{PC} - 1) / (C - 1)$	max	Uses index value of PC to better capture cluster substructure.
Partition entropy [41]	$PE = -(1/N) \sum_{k=1}^C \sum_{i=1}^N u_{ki} \log u_{ki}$	min	Average measure of entropy in membership degrees, summed across all clusters.
Modified partition entropy [42]	$MPE = (N \times I_{PE}) / (N - C)$	min	Uses index value of PE and, similar to MPC, should better capture cluster substructure.
Silhouette coefficient [43]	$SC = \max_{k=1,\dots,C} \left( 1/N \sum_{i=1}^N s_{ik} \right)$ <p>where</p> $s_i = (b_i - a_i) / \max_{i=1,\dots,C} (b_i, a_i)$ <p>with</p> $a_i = \frac{1}{ C_k -1} \sum_{j \in C_k, i \neq j} d(i, j)$ <p>and</p> $b_i = \min_{k \neq i} \frac{1}{ C_k } \sum_{j \in C_k} d(i, j)$	max	Silhouette index ( $s_i$ ) ratio for each object/point built around difference of ( $a_i$ ) average dissimilarity between $i$ and all other objects in that cluster and ( $b_i$ ) lowest average dissimilarity to all points in any other cluster, of which $i$ is not member. SC is maximum average $s_i$ for particular cluster.
Fuzzy silhouette coefficient [44]	$FSI = \max_{k=1,\dots,C} \left( 1/N \sum_{i=1}^N f s_{ik} \right)$ <p>where</p> $f s_i = \frac{\sum_{i=1}^N (u_{ig} - u_{ig'}) s_i}{\sum_{i=1}^N (u_{ig} - u_{ig'})}$	max	Similar to SC, scores based on cluster compactness and separation distances. Silhouette index ( $s_i$ ) modified using difference between greatest and second greatest membership values for object $i$ .

Table 2. Cont.

Index	Definition *	Goal	Description
Davies–Bouldin index [45]	$DB = \frac{1}{C} \sum_{k=1}^C D_k$ <p>where</p> $D_k = \max_{k \neq l} \frac{S_k + S_l}{M_{k,l}}$ <p>with</p> $S_k = \left( \frac{1}{N} \sum_{i=1}^N  X_i - V_k ^p \right)^{1/p}$ <p>and</p> $M_{k,l} = \ V_k - V_l\ _p$	min	Average similarity measure of each cluster with its most similar cluster, where similarity is ratio of within-cluster distances to between-cluster distances. Clusters which are farther apart and less dispersed will result in better score ( $p$ is usually taken as 2). $M_{k,l}$ measures separation between centroids of two clusters $k$ and $l$ ; $S_k$ is within-cluster scatter for cluster $k$ .

\*  $J_m$  is the fuzzy c-means clustering objective function, as defined in [40];  $N$  is the number of sample points;  $C$  the number of clusters;  $v_i$  is the centroid of cluster  $i$ ;  $u_{ik}$  is the membership value of point  $i$  to cluster  $k$ ;  $I_{PC}$  and  $I_{PE}$  are the index values of PC and PE, respectively;  $d(i,j)$  is the distance between data points  $i$  and  $j$  in cluster  $C_k$ ;  $u_{ig}$  and  $u_{ig'}$  are the greatest and second greatest membership values for row  $i$  of membership matrix  $U$ ;  $X_i$  is the data point  $i$  in  $n$ -dimensional space (here,  $n$  is the number of spectral bands);  $V_k$  is the  $n$ -dimensional centroid of cluster  $k$ .

## 2.6. Regional Cluster Set Formation

Training data were transformed prior to clustering. Spectral curve integral normalization is well accepted [5,20,39], which allows cluster optimization to be focused on groups of spectral shape as opposed to amplitude. Using the integral of each sample introduces a problem of invertibility when implementing identified cluster centers to assign OWT memberships to new datasets in the original reflectance space. We thus elected a log transformation for each feature (i.e., satellite band) to retain invertibility and for consistency with earlier versions used in OC-CCI. Furthermore, a log transformation will retain normality of the log-normal distributed reflectance data while reducing amplitude differences. Some training data contained negative reflectance within particular bands, arising from the atmospheric correction step. In order to retain as much of this information as possible in the cluster formation step and allow the exploration of clustering as a tool to identify these problematic pixels, a small additive shift was implemented prior to log transformation, chosen to balance reducing data loss while keeping the shift as small as possible. A PCA was run on transformed data, with all components (equal to the number of input features from the training data) being used as input for the c-means clustering optimization. For c-means, two parameters must be set a priori to run the optimization: the number of clusters ( $c$ ) and degree of fuzziness ( $m$ ). We explored the expected parameter space for these two factors using a grid search routine as part of the Fuzzy Water Clustering package. C-means cluster optimization was carried out for all  $c/m$  parameter pair nodes using Euclidean distance, which in PCA transformed space is proportional to Mahalanobis distance [46,47]. The best performing  $c/m$  parameter configuration was chosen based on selected scoring metrics (Table 2), which provided the optimized cluster statistics (cluster centers, covariance matrix) for that configuration. To assess cluster membership performance across training data, non-constrained membership values are assigned using the squared Mahalanobis distance and an  $\chi^2$  distribution, following [10,48,49]. For visualization, cluster geospatial performance in novel imagery was assessed from each region over the entire study period (2016–2021) together with regional teams using constrained Euclidean distance memberships (with 1.0 indicating perfect cluster membership).

## 2.7. Pan-Regional Cluster Set Formation

The initial testing of pan-regional clustering with the full training dataset from all six sites ( $n = 600,000$ ) was unable to capture sufficient site-specific processes, thus regional cluster sets were used to build a better representative pan-regional cluster set. As stated above, some processes (such as cyanobacteria blooms or wind-induced sediment resuspension events) are very short-lived and can happen over relatively short timescales, posing chal-



lenges to building a balanced training dataset representative of the variability across all six study sites. Both parametric (Welch's *t*-test; given reflectance data within a cluster should be log-normal distributed but still indicate variance heterogeneity) and non-parametric methods were explored; only the latter are presented here for brevity.

Regional clusters were compared pairwise between sets using the Adjusted Rand Index (ARI), the corrected-for-chance version of the Rand Index (*R*), where:

$$R = \frac{(a + b)}{(a + b) + (c + d)}$$

with *a* being the number of element pairs that are in the same cluster in both regional sets being compared, *b* the number of element pairs that are in different clusters for both regional sets, and *c* + *d* the same subset in one but not in the other regional set. The index *R* can be understood as the ratio comparing the number of cluster pair assignment agreements (*a* + *b*) to all pairwise comparisons (*a* + *b* + *c* + *d*, or  $\binom{n}{2}$ , where *n* is the total number of elements). ARI builds on this basis while accounting for different models from random clustering (which can differ in number of clusters or cluster size distribution), where

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[ \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\left[ \sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] / 2 - \left[ \sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

with  $n_{ij}$  being the number of elements in common between cluster *i* and cluster *j* from each cluster set, respectively (i.e., the intersection of cluster *i* and cluster *j*),  $a_i$  the sum of elements in cluster set *i* and  $b_j$  the sum of elements in cluster set *j*.

To provide information on specific cluster pair similarity between regional sets, ARI was calculated based on one cluster within a set being successively retained and all other clusters within that same set being conglomerated to "other" (and the same being performed with the comparison cluster set). This results in all pairwise comparisons of clusters between the two sets being assigned an ARI score. ARI values can be negative up to one, with those closer to one indicating the cluster pair between regional sets being essentially the same. The grouping of similar regional clusters while retaining those regional clusters which prove unique across all regional cluster sets provided an estimated pan-regional cluster set, which was used for setting the parameter space and as the initialization configuration for a semi-supervised fuzzy clustering analysis combining training data across all six study sites.

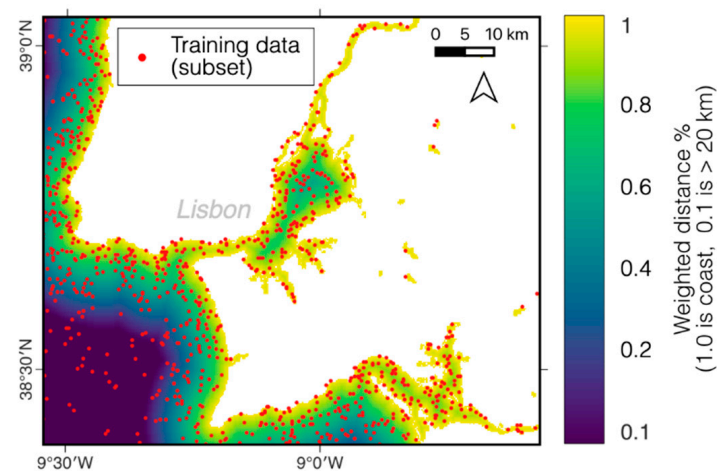
### 3. Results

As mentioned in Section 2.2, regional cluster analysis is presented for only one of the six study sites. Reports detailing results for the other sites are provided in the Supplemental Material.

#### 3.1. Regional Clustering

Training data from the satellite image timeseries were built using a random subsample of pixels weighted by distance from land to ensure the relatively equal representation of coastal areas to larger offshore areas in the dataset. Figure 3 shows an example of one day of stratified random sampling from the full timeseries for the Tagus and Sado Estuaries. A coastline buffer of 20 km was chosen for weighting to best represent a suitable balance between inland and coastal waters with offshore regions across the six study sites. This process was performed over both the MSI and OLCI timeseries data (2016 to 2021) to produce a training dataset for each sensor that had *n* = 100,000 for each region. To ensure that no more than 5% of training data were lost within any band across all study sites due to negative reflectance, the following small additive shifts were implemented prior to log transformation: 0.015 for OLCI, 0.003 for MSI. These values were based on a 5%

acceptance threshold of data losses across all bands due to there being negative values when transforming the data.



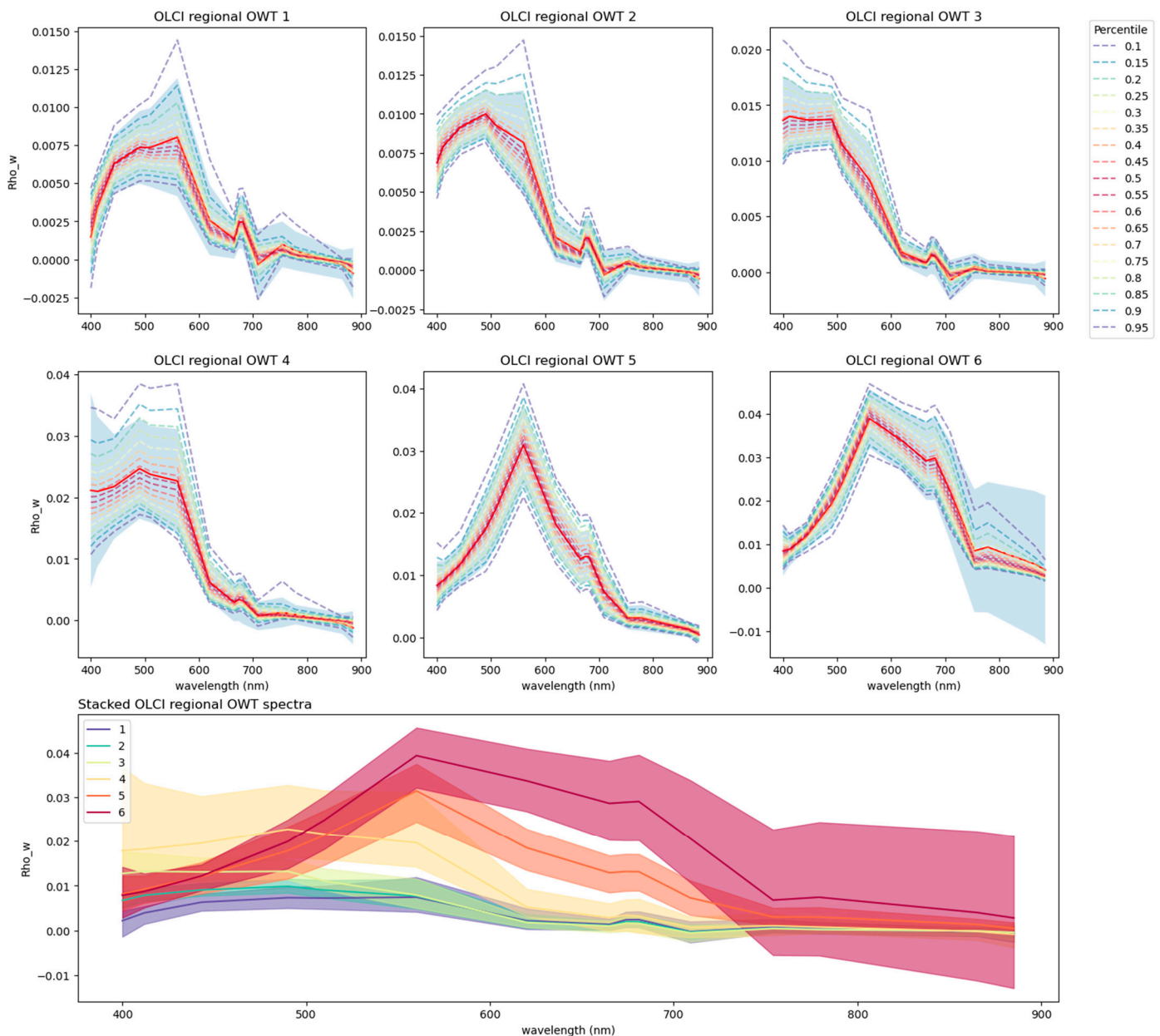
**Figure 3.** The training data spatial distribution (red dots) for a single day overlaid on the OLCI timeseries spatial grid, colored to represent the value weighting relative to the coastline used for stratified random sampling frequency (from dark blue, being 10%, to yellow, at 100% random sampling frequency).

### 3.1.1. Tagus OLCI Regional Cluster Set

Transformed OLCI training data PCA components were processed through a fuzzy c-means clustering scheme. As mentioned, fuzzy c-means clustering requires a number of clusters ( $c$ ) and a fuzziness factor ( $m$ , defining the degree of allowable cluster overlap) to be set a priori. We used a parameter space grid representative of the expected optical variability of the transitional water across all regions, specifically  $m = [1.2, 2.5] \in \mathbb{Q}$  with steps of 0.3 and  $c = [6, 12] \in \mathbb{N}$  with steps of 1. The optimization solution for each  $c/m$  combination was assigned a score (Xie–Beni, hard silhouette, fuzzy partition coefficient, and Davies–Bouldin were considered; see Table 2, of which Xie–Beni was found to be the most stable). The best performing score across all  $c/m$  combinations was selected, which for the Tagus region with OLCI data was  $c = 6$  and  $m = 2.1$ . The optimal cluster set is shown in Figure 4 in untransformed reflectance space (as water-leaving reflectance,  $\rho_w$ ), both as single clusters overlaid with the standard deviation and percentile distribution of the training data with the dominant membership of that cluster, and as the combined cluster set overlaid with the standard deviation. The single clusters display a variety of spectral shapes, with OWT 1,2,3 having pronounced peaks in blue/green bands and the 681 nm band. OWT 1 and 2 are close in shape with the exception of peak characteristics (the shift from 665 nm to 490 nm, respectively, and the markedly lower reflectance values at 400 and 412 nm for OWT 1). OWT 4 is also similar in shape to these three, albeit with a much higher overall reflectance and a blue/green peak that has shifted back to 490 nm. OWT 5 and 6 display a very different spectral shape, with the highest peak in the 560 nm band and an enhanced secondary peak in the 600 nm bands.

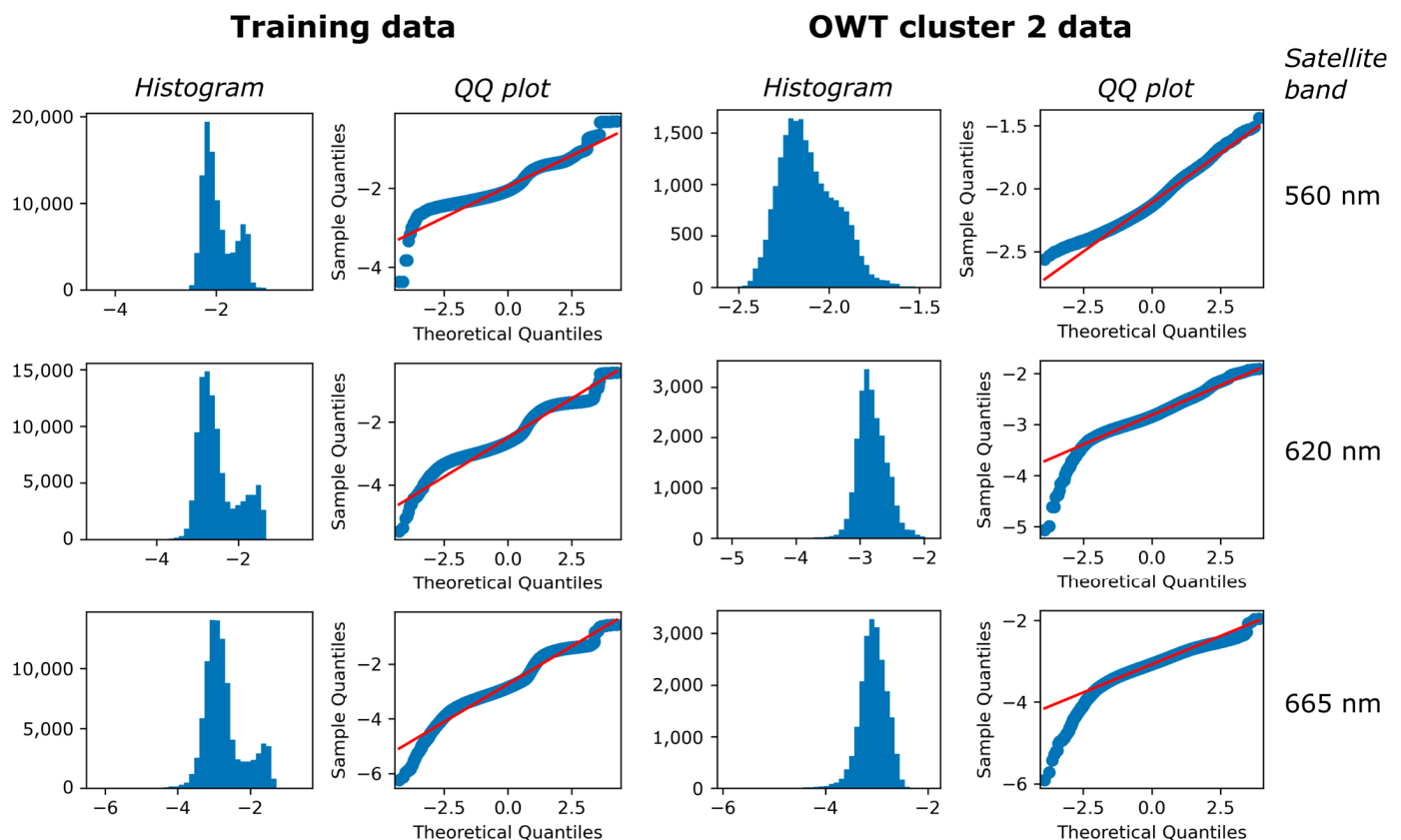
The log-normal distribution of reflectance spectra from a single target was used to additionally assess cluster performance. Training data displayed clear multimodal characteristics prior to clustering (left column Figure 5) in histograms of log-transformed reflectance for single bands. Quantile–quantile (QQ) plots compared sample quantiles to theoretical quantiles from a representative distribution (in this case, normal) and are a useful tool for qualitatively assessing sample distribution fit to expectation, as indicated by the red standardized one-to-one line. For the full training dataset, one sees marked steps in the QQ plots that indicate the multimodal structure present in the training data representative of various optical water types across space/time in the Tagus Estuary. After clustering, using OWT 2 as an example (right column Figure 5), one sees little indication of

a remnant multimodal structure in the histograms. Steps in the QQ plots have also been removed, supporting the successful removal of a multimodal structure within a single cluster, but indicating that reflectance data in some bands remain lightly skewed.



**Figure 4.** Regional optical water type (OWT) clusters created from Tagus OLCI training data, showing spectra for each cluster together with spectra distribution for those training data with dominant membership for that particular cluster (cluster center is solid red line,  $\pm 1$  standard deviation in gray shading, percentiles as broken lines with rainbow colors). Lower plot shows overlaid cluster center spectra (solid line) for all OWT classes with  $\pm 1$  standard deviation in shading of same color.

The OWT membership distribution in geographic space is shown for the Tagus Estuary (Figure 6) on a single date (6 September 2020). OWT classes 1 to 3 primarily represent Atlantic waters, but interestingly also some areas in the upper estuary. OWT 4 captures well the coastal areas, OWT 5 the lower estuary and water exiting the estuary, and OWT 6 the mid- to upper estuary.

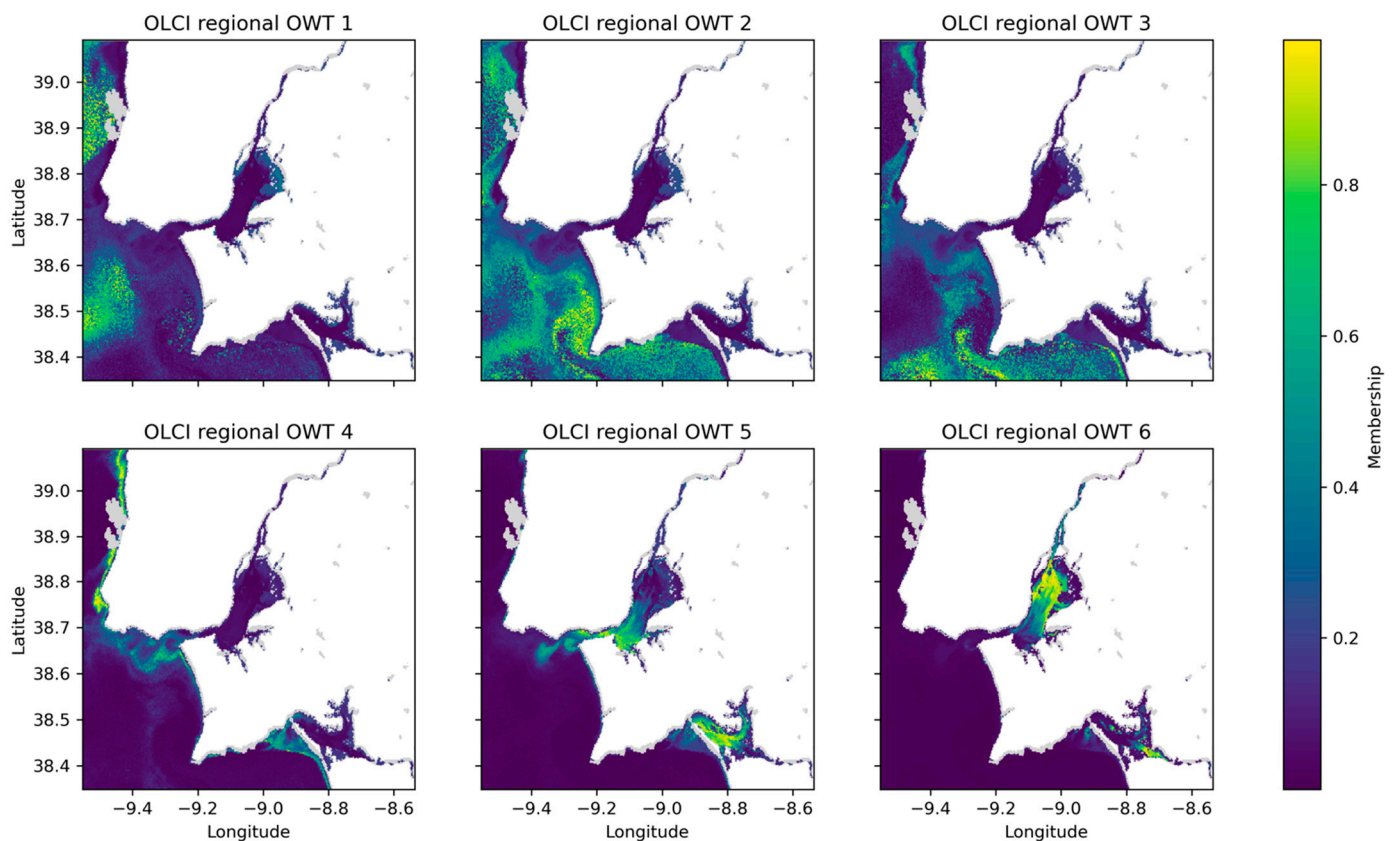


**Figure 5.** A comparison of log-transformed reflectance histogram and quantile–quantile (QQ) plots for single bands from the full training dataset (**left**) and for a particular cluster (here OWT 2, **right**). Through clustering, multimodality has been reduced and data better follow a normal distribution, as indicated by the disappearance of steps in the cluster QQ plots. The red line in the QQ plots is the standardized line, representing the expected order statistics scaled by the standard deviation of the given sample and then adding the mean.

### 3.1.2. Tagus MSI Regional Cluster Set

Similar to the process performed with OLCI, transformed MSI training data PCA components were fed into the fuzzy c-means clustering scheme. An analogous parameter space was utilized for the grid search,  $m = [1.2, 2.5]$  and  $c = [6, 12]$ , under the same assumption of expected optical variability across the sites. The best performing score across all  $c/m$  combinations for MSI data was  $c = 6$  clusters and  $m = 2.1$  as the fuzziness factor. Figure 7 presents the optimal MSI cluster set for the Tagus Estuary in untransformed reflectance space, again as single clusters overlaid with the standard deviation and percentile distribution of training data with dominant membership for that particular cluster, and as the full cluster set overlaid with their standard deviations. Single clusters again show a variety of spectral shapes, albeit to a lesser degree with the coarser spectral resolution MSI data as compared with OLCI (Figure 4). OWT classes 1, 2, and 3 have a pronounced peak in the 490 nm band and a much smaller peak in the 783 nm band. Of those three classes, OWT 1 is more distinct in the elevated 443 nm band relative to the 490 nm peak and the slight convex curve around the 560 nm band. OWT 4 displays a similar shape to these three but with the highest peak shifting to the 560 nm band. The last two OWT classes, 5 and 6, also have the highest peak in the 560 nm band but much reduced relative reflectance in the 443 and 490 nm bands. OWT 6 also shows increasing overall reflectance in the NIR (bands 665 and 705 nm). Checking log-normal distribution assumptions after clustering suggested that the multimodal structure from the full MSI training data was again reduced through clustering.





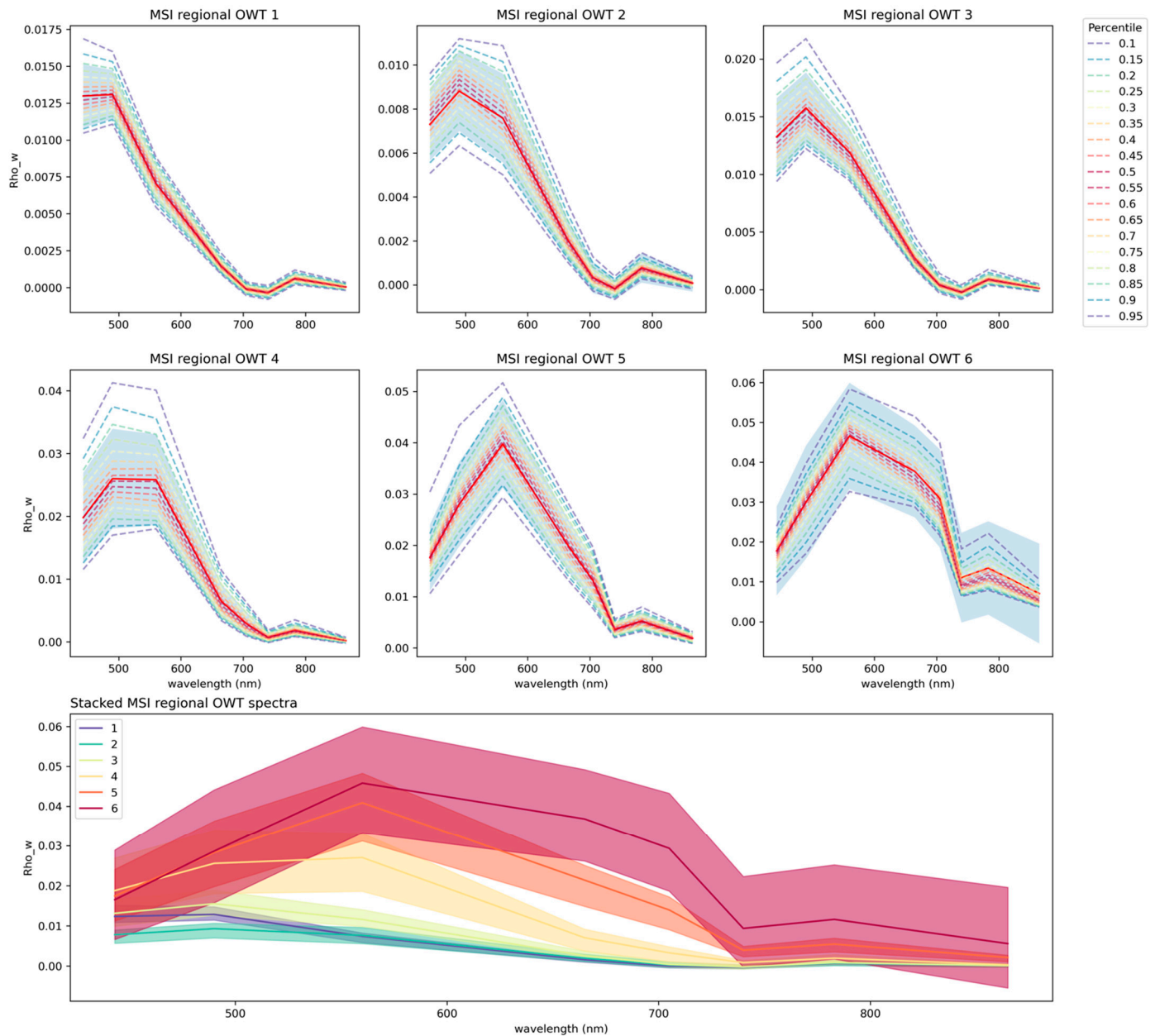
**Figure 6.** OLCI optical water type (OWT) cluster membership (%) distribution across the Tagus study site on 6 September 2020, with dark blue representing low membership to that cluster and yellow high cluster membership (masked water pixels are light gray).

The spatial distribution over time of regional OWT membership was also assessed based on the dominant OWT, representing the highest membership water class for a given pixel either over a month, year, or full timeseries. The MSI dominant OWT map for the Tagus Estuary over the full time series (2016–2021) is shown in Figure 8. OWT 1 and 2 primarily represent the Atlantic waters, while OWT 3 captures well the waters along the coastline. OWT 2 in particular captures a satellite along-track feature in the offshore waters, likely caused by sun glint at a particular but repeated overpass geometry. OWT 4 represents the mid-estuary waters and OWT 5 those of the upper estuary. Dominant OWT membership for particular months, across the full time series, is shown in Figure 9. The left panel is based on March data, representing when the Tagus river characteristically has high discharge rates. Coverage by dominant OWT 4 spreads much further into the Atlantic from the Tagus river mouth as compared with the full time series map (Figure 8) and this wider plume extends much further north along the coast. Some of the lowest Tagus river discharge rates are in August, shown in the right panel of Figure 8. In the summer, wind-driven resuspension and upwelling water intrusion into the Tagus Estuary can further play an important role in the overall water color. The spatial distribution of dominant OWT 4 is much reduced, spreading less far into the Atlantic and only present as a thin band along the coastline north of the river mouth.

### 3.2. Pan-Regional Clustering

To establish a best-guess pan-regional cluster initialization set, a comparison of regional cluster sets was performed to identify groups of clusters with similar spectral signatures and those clusters unique across all study sites. ARI scores indicated cluster membership similarity, with an  $ARI \leq 0$  indicating that two data cluster memberships do not agree on any pair of points, while an ARI value of 1 would indicate that compar-

ison cluster sets were exactly the same. A threshold of  $ARI \geq 0.35$  was used to group regional clusters.

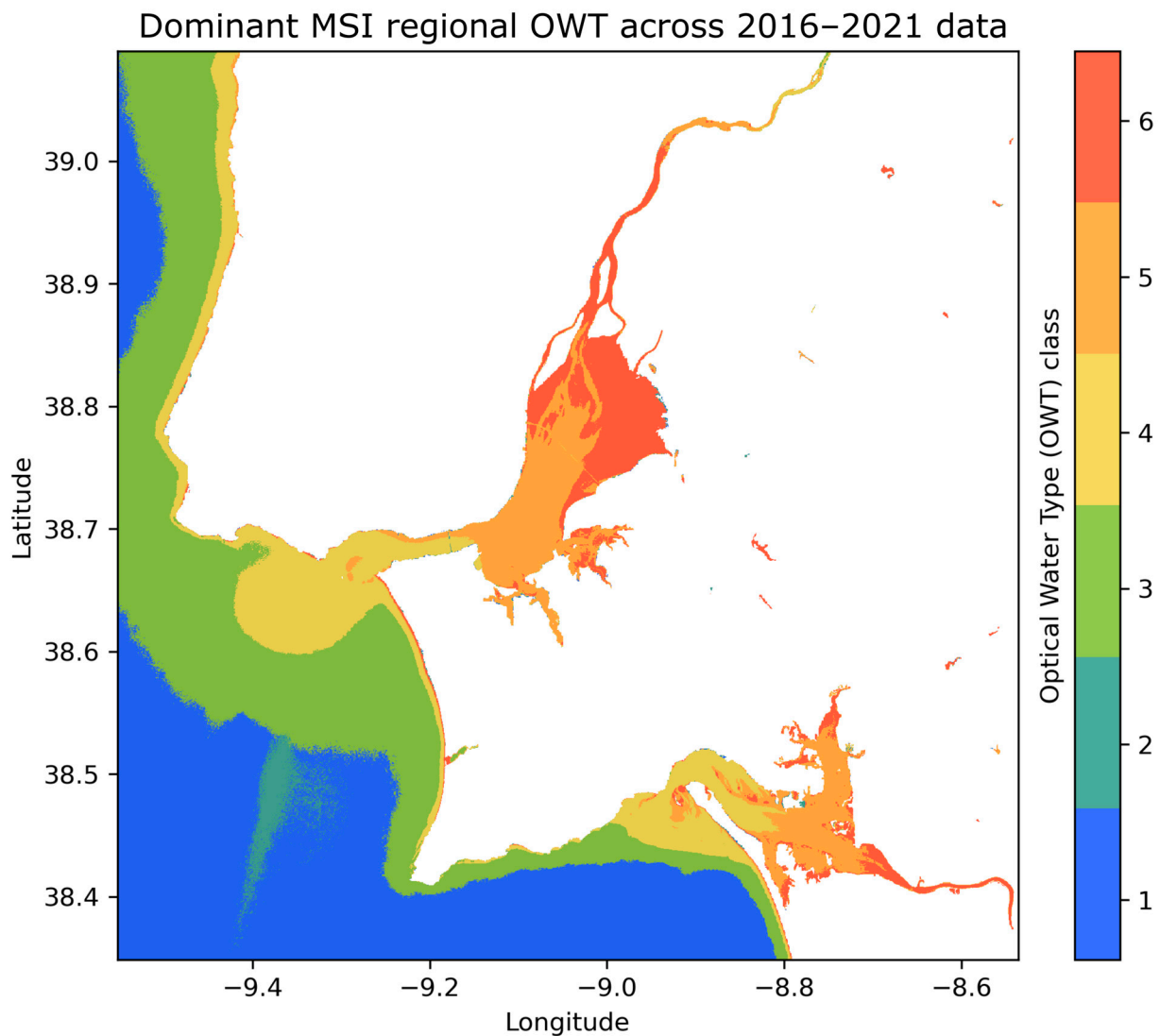


**Figure 7.** Regional optical water type (OWT) clusters created from the Tagus MSI training data, showing spectra for each cluster together with the spectra distribution for those training data with a dominant membership for that particular cluster (cluster center is solid red line,  $\pm 1$  standard deviation in gray shading, percentiles as broken lines with rainbow colors). The lower plot shows overlaid cluster center spectra (solid line) from all OWT classes with  $\pm 1$  standard deviation in the shading of the same color.

### 3.2.1. OLCI Cluster Set

Some sites were found to contain many regional cluster spectra similar to spectra from other study sites, such as the Danube Delta/Razelm–Sinoe Lake Complex, while those from the Curonian Lagoon and Tamar Estuary were found to be generally more unique. Further fine-tuned groupings through the visual comparison of regional cluster spectra resulted in a final set of 18 spectra, consisting of 16 grouped (example grouped spectra subset in Figure 10, left column) and 2 unique OWT spectra (full set provided in Supplemental

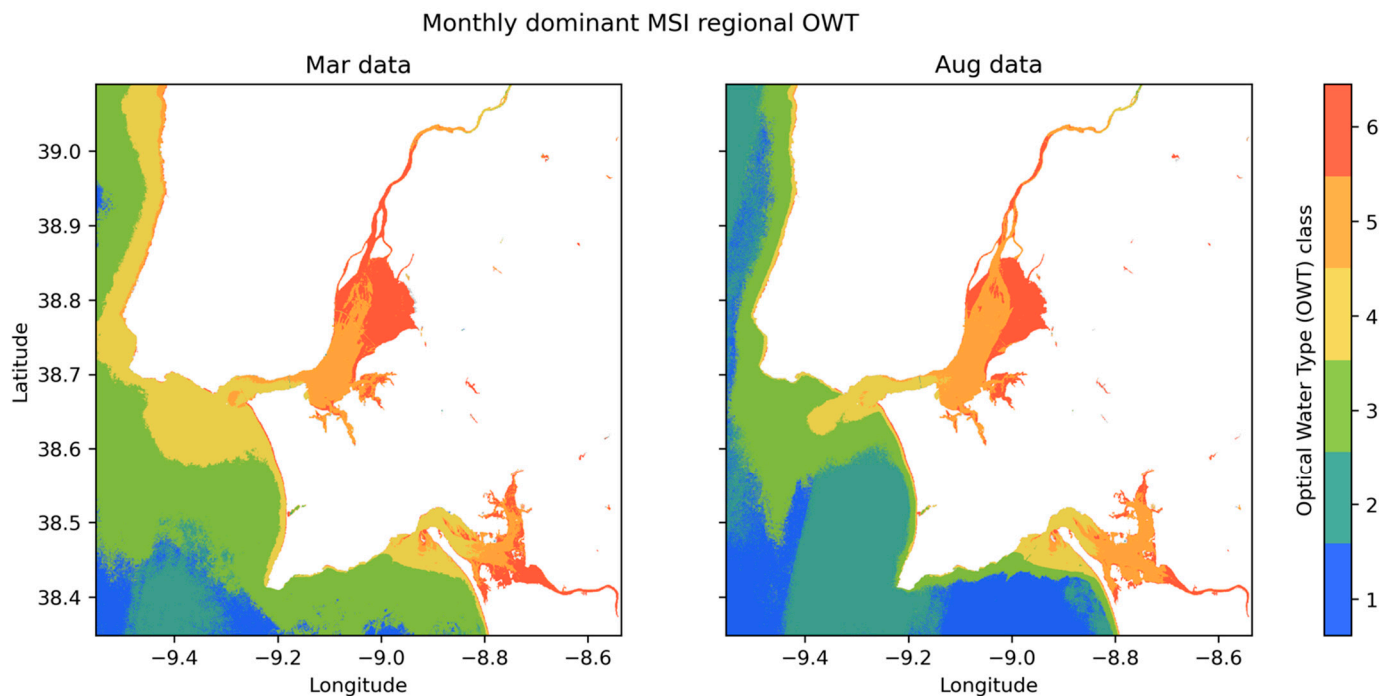
Material). Characteristic spectra for each group were estimated using an average, and the full set ( $c = 18$ ) was subsequently used as the initialization for a semi-supervised pan-regional clustering (associated pan-regional spectra subset in Figure 10, right column). The optimized pan-regional OWT spectra retain tight standard deviation estimates around cluster center spectra, which was the case for most OWT classes. Checking the geographic distribution of the 18 pan-regional OWT classes across the six separate CERTO study sites (for example Tagus and Sado Estuaries in Figure 11, other sites in Supplemental Material), a sensible structure was observed for each site. Clear coastal sea or ocean waters were well represented by the lower OWT classes, while the higher OWT classes better represented transitional waters.



**Figure 8.** The dominant MSI regional optical water type (OWT), based on summed membership for each pixel over the entire timeseries (2016 to 2021).

Examining the geographic distribution of the 18 OLCI pan-regional OWT classes across the Tagus Estuary based on the full timeseries dominant class (Figure 11, left panel) showed a generally similar spatial pattern, with pan-regional OWT classes 3 and 6 primarily representing offshore waters. The Tagus river plume is primarily represented by pan-regional OWT class 11, the mid-estuary area by OWT class 15, the northwestern portion of the estuary by class 16, and the eastern portion by class 17. One can compare these to the spatial distribution of OC-CCI v6.0 1 km OWT classes (noting that OC-CCI data have a spectral basis of MERIS-referenced merged sensor bands), based on the dominant class

from 2016 to 2021 data (Figure 12, right panel). The OC-CCI OWT set shows appropriate capture of the Tagus river plume but regional complexity is reduced to representation from only four classes. The OC-CCI classes 11 and 12 are capturing the offshore waters and spatially map fairly well to the coverage of pan-regional OWT classes 6 and 3, respectively. The differentiation of the largest estuary system in Western Europe, though, is reduced to a single OC-CCI class, which for the pan-regional OWT classes is primarily represented by four separate classes.



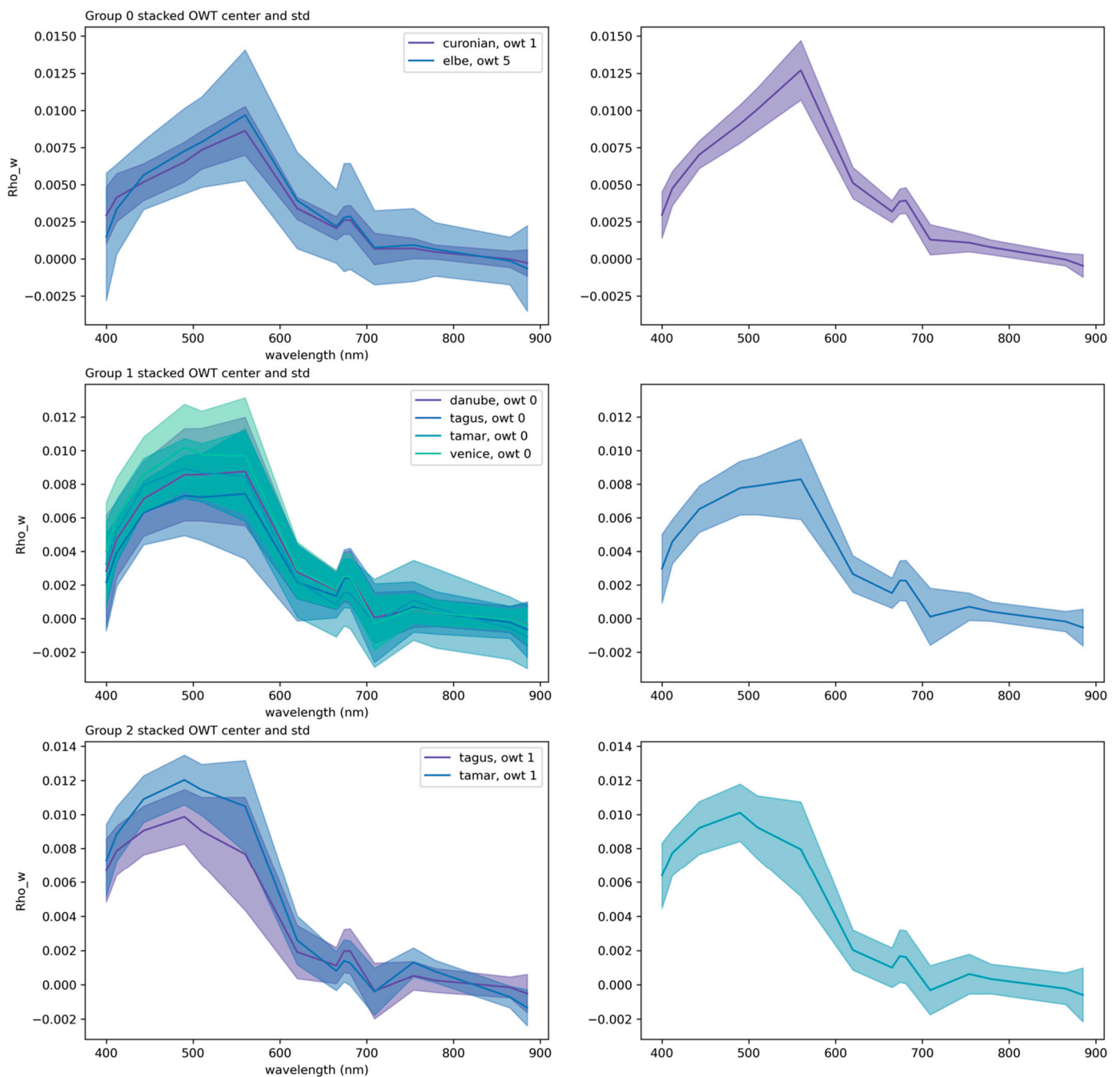
**Figure 9.** The dominant MSI regional optical water type (OWT), based on summed membership for a particular month, for each pixel over the entire timeseries (2016 to 2021). Data from March, representing peak Tagus River discharge, are on the left and on the right from August when river discharge is at its lowest.

### 3.2.2. MSI Cluster Set

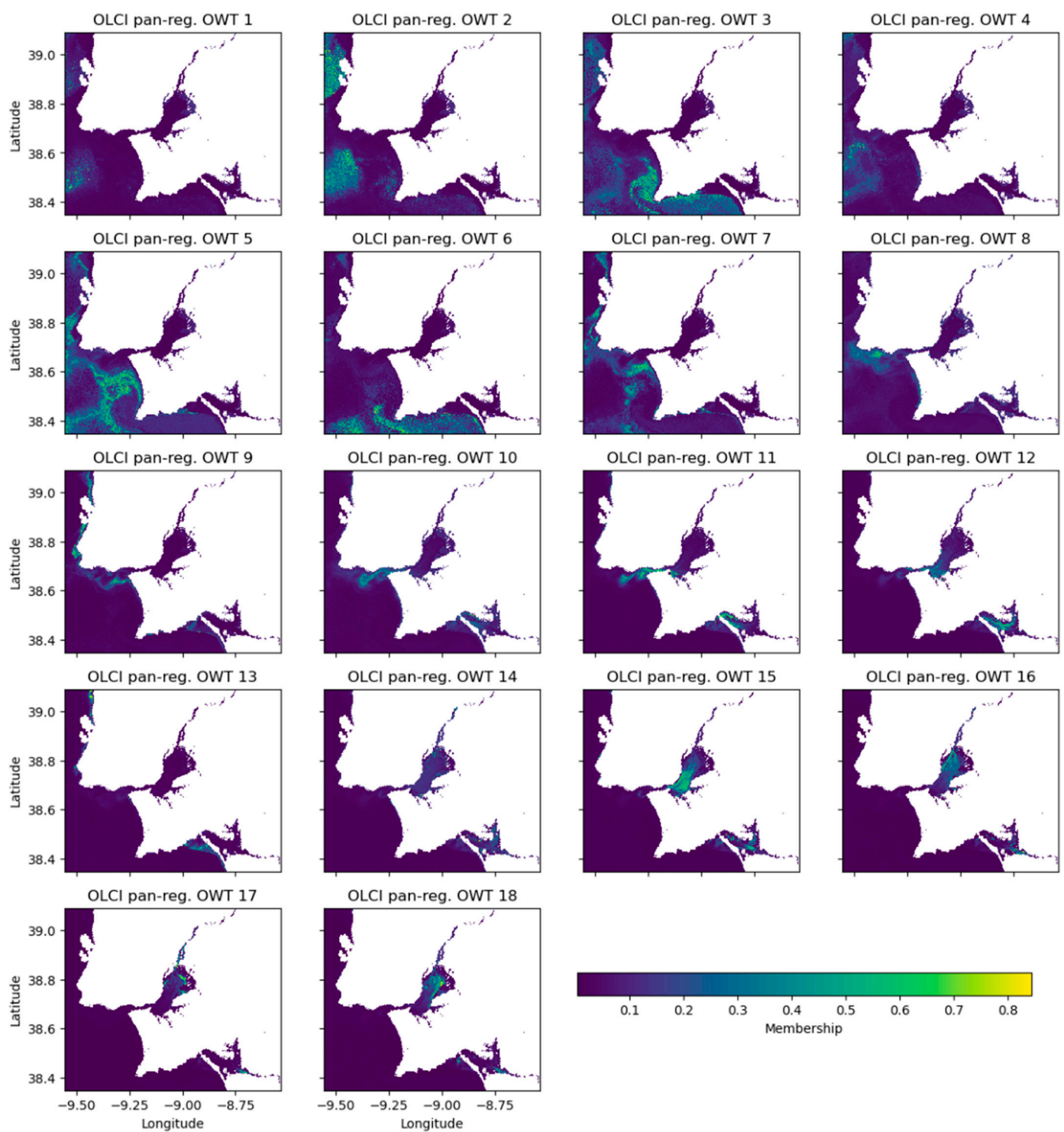
Regional clusters from MSI were more commonly found to overlap with similar spectra from other study sites, with those from the Danube Delta/Razelm–Sinoe Lake Complex all grouping with at least one cluster from another region. At the same time, there were also an increased number of spectra which did not group with any other regional clusters. ARI thresholding together with the visual comparison of regional cluster spectra resulted in a final set of 17 spectra, representing 9 groups of clusters and 7 unique OWT spectra (grouped spectra subset in Figure 13, left column, full set in Supplemental Material). As for OLCI, the characteristic spectra of each grouping were estimated using the mean and the full set ( $c = 17$ ) used as initialization for the semi-supervised pan-regional clustering (associated pan-regional spectra subset in Figure 13, right column).

Study site comparisons of the regional and pan-regional dominant OWT are shown in Figure 14, for the respective low river discharge month of that region. Here, we show the Tagus and Sado estuaries as well as two further sites to expand the comparison (other sites in Supplemental Material) and understanding of the utility of the pan-regional set across different areas. The geographic distributions of the 17 MSI pan-regional OWT classes across the CERTO study sites show similar general spatial pattern to their regional cluster sets but often with finer definition of local features. Offshore waters are best captured with the pan-regional OWT classes 1, 2, and 4, while the outflowing river plumes that were primarily represented with one regional class were better covered by two pan-regional classes (Danube classes 12, 14, and 16; Tagus 7, 8, and 10; Tamar 7 and 9).

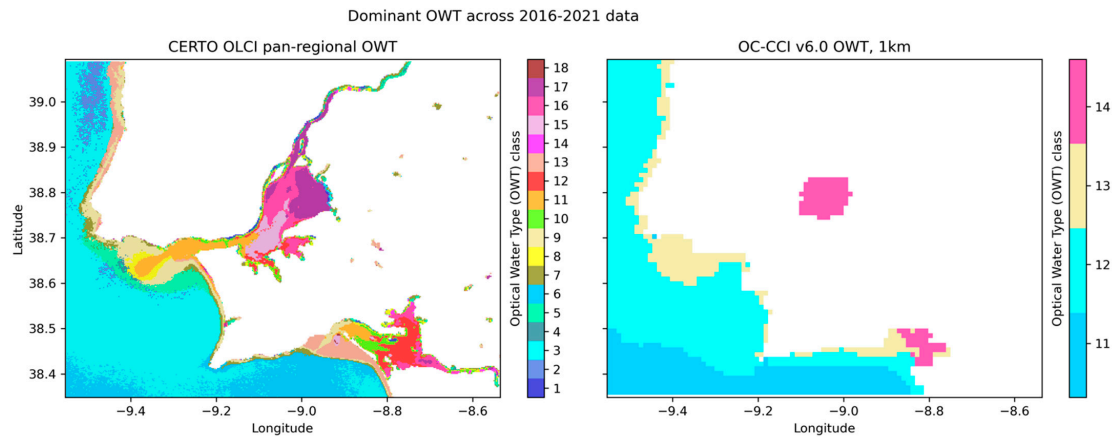




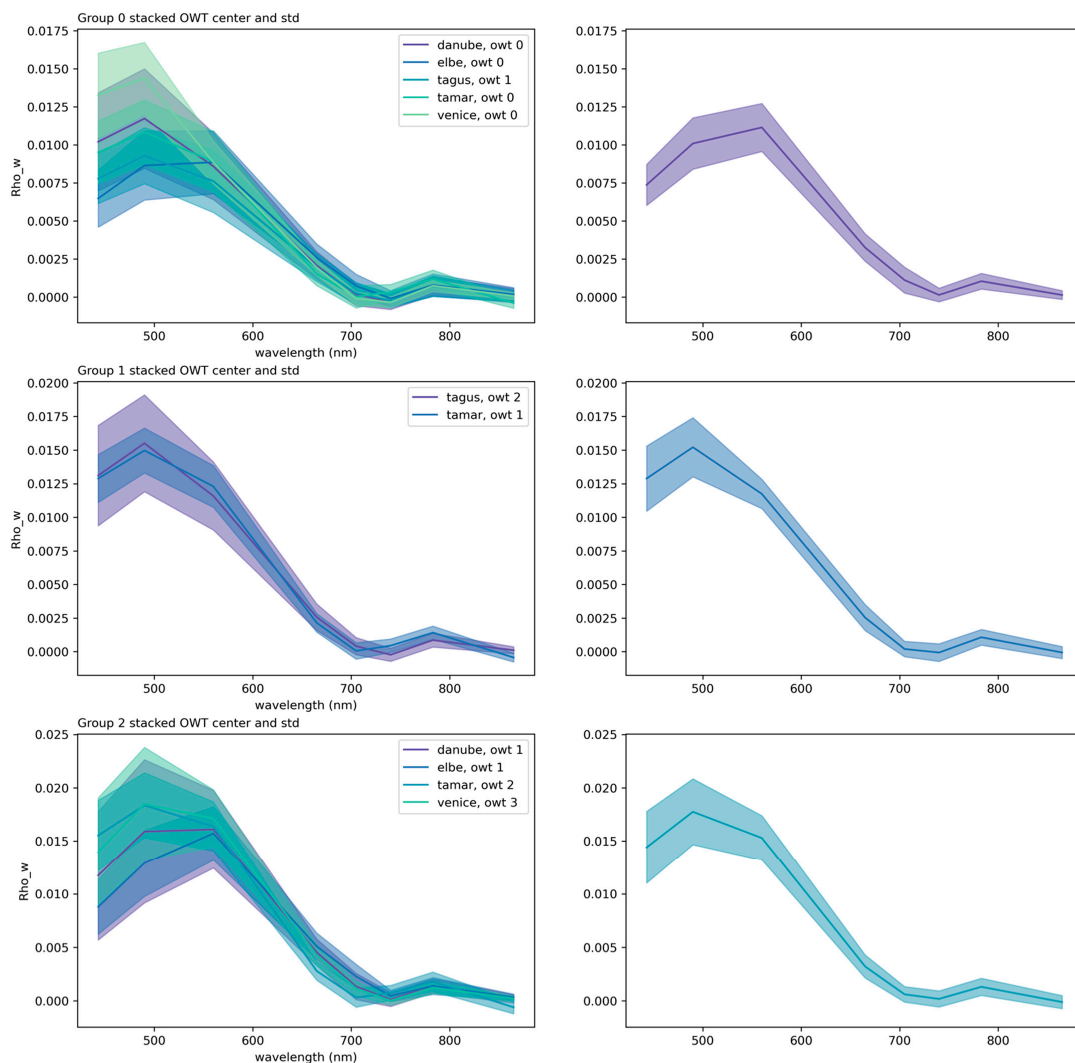
**Figure 10.** Left column contains example subset of grouped OLCI regional optical water type (OWT) cluster spectra (solid line, standard deviation as shaded region) based on Adjusted Rand Index  $\geq 0.35$  threshold groups. Full grouping set is presented in Supplemental Material. Grouped regional OWT spectra were used to estimate initialization cluster center for semi-supervised global c-means analysis; right column contains associated OLCI pan-regional cluster spectra.



**Figure 11.** Constrained Euclidean distance memberships (%) to the 18 OLCI pan-regional optical water type (OWT) classes for a single date (6 September 2020) from the Tagus Estuary.



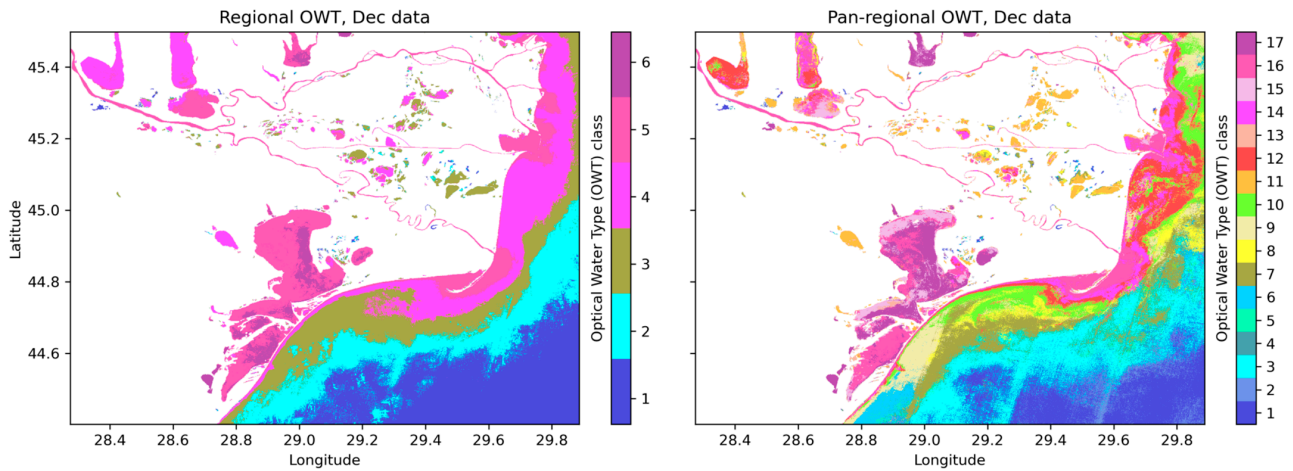
**Figure 12.** Dominant optical water type (OWT), based on summed membership for each pixel, over timeseries (2016 to 2021) for OLCI pan-regional (**left panel**) as compared with OC-CCI v6.0 1 km product (**right panel**).



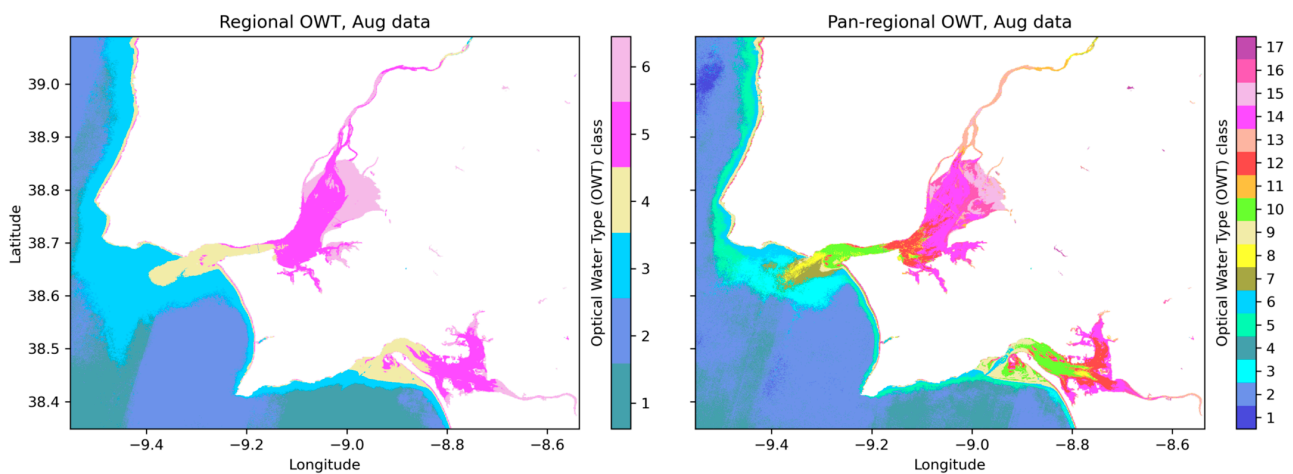
**Figure 13.** Left column contains example subset of grouped MSI regional optical water type (OWT) cluster spectra (solid line, standard deviation as shaded region) based on Adjusted Rand Index  $\geq 0.35$  threshold groups. Full grouping set is presented in Supplemental Material. Grouped regional OWT spectra were used to estimate initialization cluster center for semi-supervised global c-means analysis; right column contains associated MSI pan-regional cluster spectra.

## MSI regional vs pan-regional OWT coverage, monthly composites

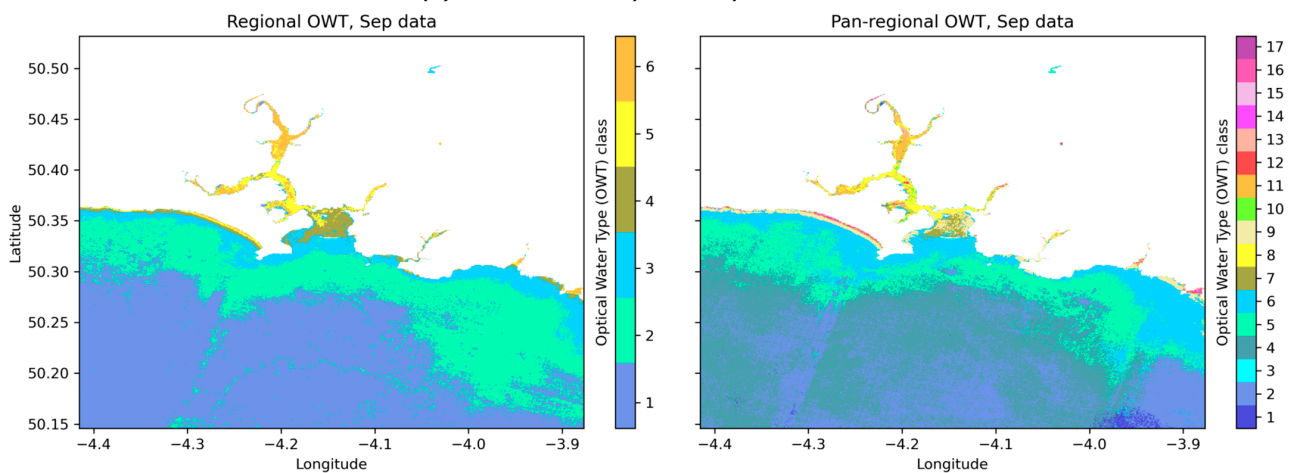
### (a) Danube Delta and Razelm–Sinoué Lagoon System



### (b) Tagus and Sado Estuaries



### (c) Tamar Estuary and Plymouth Sound



**Figure 14.** A comparison of MSI regional (left column) and pan-regional (right column) cluster set geographic coverage by dominant optical water type (OWT) for three study sites, based on dominant summed membership for the month (from full timeseries 2016 to 2021) with low river discharge for that site. Sites are (a) the Danube Delta and Razelm–Sinoué Lagoon System for the low river discharge month December, (b) the Tagus and Sado Estuaries for the month of August, and (c) the Tamar Estuary and Plymouth Sound for September.



#### 4. Discussion

Regional fuzzy clustering was found to provide useful site-specific information for the transitional water systems considered. This was demonstrated through the ability to represent transition zones of freshwater mixing into saline coastal water with per-pixel maximum membership values in excess of 0.60 across all study sites. In particular, this held true for transitional water systems which have proven challenging for monitoring with satellite imagery due to artificial boundaries and artifacts introduced from varying optically active constituent concentrations (phytoplankton- vs SPM- vs CDOM-dominated waters) or differing atmospheric correction algorithm performances between clearer offshore waters and more turbid transition and inshore waters. Optimal cluster parameters for  $c$  and  $m$  were determined via a comprehensive grid search. A final regional cluster set was found to reduce multimodality contained within the regional training dataset, better fulfilling the expected log-normal distribution of reflectance from a cohesive single target within an OWT class. The spatial distribution of dominant OWT coverage across the study sites conformed with regional teams' expert understandings of dynamics for their particular region. Seasonal patterns, due to forcing such as a variation in the river discharge rate, were evident in monthly composite images. It should be noted that despite best efforts, the training data are not free of atmospheric correction impacts from land-water mixed pixels, adjacency effects, and optically shallow waters, as can in part be evidenced by the negative reflectance values seen in Figures 3, 6 and 9. A benefit of using an OWT set on novel data which was trained on data processed the same is that these impacts should inherently be part of the optimization procedure and thus have representation in the OWT classes obtained.

Comparison between regional cluster sets provided the basis for the creation of a pan-regional OWT set that retained site-specific cluster features while combining common OWT classes. Using the non-parametric ARI score, patterns in membership occurrence over both space and time are the basis for grouping similar regional OWT classes. Regional cluster groupings are further confirmed through visual inspection of the spectral curves. The grouped set is used in a semi-supervised cluster analysis to produce a pan-regional OWT set. The pan-regional cluster set has tight standard deviations around cluster center spectra for most classes. Geographic coverage by dominant membership from the pan-regional OWT set suggests that site-specific features highlighted in the regional analyses were retained, but often represented with finer definition (more classes covering spatial variability, as observed in Figure 13). Viewing the spatial distribution with the dominant OWT from membership values is a useful tool for a simplified understanding of which areas are primarily represented by which OWT classes, but one should remember that this hides the "fuzzy" aspect of the c-mean cluster optimization algorithm. Two or more OWT classes from the pan-regional set could be sufficiently similar in spectral space such that a pixel has high memberships to those classes, with the dominant OWT only determined by minimal differences in membership. The comparison method can be used when analyzing new study sites in order to determine if unique OWT classes occur, and the pan-regional cluster set expanded to better encompass spectral variability from a wider assemblage of transitional water systems. ARI scoring is independent of the cluster method or data transformation implemented. So long as membership values can be calculated based on a common dataset, ARI scoring presents an ideal method for the comparison of OWT cluster sets from different studies.

The impact of increasing spectral resolution, using the Tagus regional clusters from MSI (Figure 6) to OLCI (Figure 3) as an example, suggests that while some optically active spectral features are lost with the lower spectral resolution of MSI, regional cluster sets from the two sensors are comparable. The more turbid OWT classes 5 and 6 from both sensors have their max peak in the 560 nm band, and OWT class 6 displays a convex curve between this max peak and the mid-700 nm bands for both OLCI and MSI. OWT classes 3 and 4 from both sensors have peaks at 490 nm, albeit for MSI OWT 4 this peak is secondary to one at 560 nm. Comparing the clearer offshore OWT classes 1 and 2, cluster numbering

appears switched between the sensors, with OLCI OWT 2 having a small but relatively pronounced peak at 490 nm, matching the same feature in MSI OWT 1. The OLCI OWT classes display the Chl-a absorption feature between 650 and 700 nm, which is missing from the MSI OWT classes due to the lower spectral resolution of this sensor. A comparison of the geographic distribution between the regional OLCI (Figure 5) and MSI (Figure 7) OWT classes shows a similar partition of the Tagus Estuary, although it should be noted that the tidal condition between the Sentinel-2 overpass time and that of Sentinel-3 will be different. OWT classes 1 and 2 from both sensors represent the offshore waters, with a switch in predominant coverage by OLCI OWT 2 to MSI OWT 1 (matching the spectral comparison between the sensor regional OWT sets).

Coherent OWT class sets created from different sensors offer a method for inter-sensor harmonization of EO products. OLCI sensors provide medium-resolution (300 m) daily imagery with a high Signal-To-Noise Ratio (SNR), while MSI sensors provide high-resolution imagery (10–60 m) every 5–10 days with a lower SNR. Across the transitional water sites focused upon within CERTO, large systems such as the Tagus and Sado Estuaries can be well characterized spatially with medium-resolution OLCI data. But smaller systems, such as the Tamar Estuary, are too small in area for sufficient valid pixel coverage by OLCI for insightful characterization. Differences in the estimates of WQ parameters from disparate sensors are an issue, which inhibit the cohesive use of the full satellite imagery portfolio. Coherent sensor-specific OWT classes between MSI and OLCI could be used in a variety of ways, such as filling the temporal gaps between MSI acquisitions with coarser resolution OLCI imagery, or provide high-quality WQ estimates for MSI images based on the same OWT class estimates from OLCI imagery with a higher SNR, allowing for high spatial resolution estimates of WQ in smaller transitional water systems.

We used our determination of common OWT classes between different regional cluster sets to support the coordination of CERTO field campaigns between the respective sites. This use case allowed us to create a consistent dataset across all study sites, representing the full range of WQ conditions present within these regions. The use of harmonized OWT classes helps determine which water masses are well sampled and which should receive more focus in future sampling efforts. In general, OWT coverage maps provide useful information for sampling location planning within field campaigns in order to increase sampling within rarer OWT classes. Furthermore, it is possible to infer the WQ characterization of a specific OWT class at a site where no in situ sampling has occurred, if the equivalent OWT has been sampled and analyzed in other sites.

Future work should focus on making comparisons between pan-regional OWT sets and other widely implemented OWT classifications, such as that of OC-CCI and Lakes-CCI. Each of these cluster sets are based on different types of training data, with OC-CCI clustering based on multi-sensor global satellite reflectance and Lakes-CCI based on in situ hyperspectral data from LIMNADES. As discussed in Section 2.4, each of these training data approaches provide various benefits and pitfalls, which need to be considered in the context of the intended application of the OWT classes. Wei et al. [14] performed a pairwise comparison of clusters across seven cluster sets, all trained on different data types and normalization schemes, using minimum cosine distance that focuses on comparing cluster center spectral shape. Nearness in spectral space is a clear indicator of cluster center similarity but, given that one can use OWT class partitioning to decipher the optical complexity and unravel the optical diversity of natural waters globally [50], a cluster comparison method that takes account of occurrence patterns in geographic and temporal space is just as important. A method such as the ARI comparison technique is well suited for the task of cluster set comparison given that the direct comparison of cluster center spectral curves may be difficult to impossible, depending on data formats. A cross-comparison would allow for determination if a particular OWT classification scheme had identified a unique water type that may have been missing.

## 5. Conclusions

Fuzzy c-means clustering is a classification tool well suited for transitional water systems through membership representation of mixing processes occurring within river mouths, estuaries, lagoons, and deltas that affect water-leaving reflectance. Cluster analysis at the regional level proved to be valuable to identify site-specific information from transitional water systems, and effectively captured transition zones where freshwater and coastal waters mix. The spatial distribution of dominant OWT coverage across the study sites accurately reflected the expected dynamics for each individual site. We presented a novel cluster set comparison method using ARI scoring over memberships to build a representative pan-regional cluster set that retained site-specific features. The pan-regional OWT set demonstrated here can be used as a basis for per-OWT class calibration of WQ algorithms, from which the optimum performing algorithm can be selected and membership values for those OWT classes used as a basis for a weighted blending of algorithms to produce a final WQ product. This method provides a first attempt to harmonize WQ data products across oceans (OC-CCI, C3S, NASA), regional seas (CMEMS), and inland waters (Lakes-CCI, CLMS). A harmonized EO monitoring system that represents well the continuum from inland aquatic systems to the open ocean would improve coastal water quality monitoring capabilities needed to address international water quality directives.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/rs16173267/s1>, File S1: Tagus Estuary, File S2: Elbe Estuary and German Bight, File S3: Curonian Lagoon, File S4: Tamar Estuary and Plymouth Sound, File S5: Venice Lagoon and northwest Adriatic Sea, File S6: Danube Delta and Razelm–Sinouie Lagoon System, File S7: OLCI pan-regional grouped cluster set, File S8: MSI pan-regional grouped cluster set.

**Author Contributions:** Conceptualization, T.J., S.S. and E.C.A.; methodology, E.C.A., T.J., A.L., B.F.J. and G.S.; software, A.L., E.C.A., T.J., O.D. and B.F.J.; validation, E.C.A., T.J. and A.L.; formal analysis, E.C.A.; investigation, E.C.A.; resources, E.S., D.J., E.C.A., G.S., N.S. and S.S.; data curation, E.C.A., T.J. and A.L.; writing—original draft preparation, E.C.A.; writing—review and editing, E.C.A., T.J., A.L., B.F.J., E.S., G.S., D.J., N.S., S.S., A.T. and S.G.; visualization, E.C.A., B.F.J. and T.J.; supervision, T.J. and B.F.J.; project administration, T.J., E.C.A., S.G. and A.T.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the European Commission through the EU Horizon-2020 project CERTO (Copernicus Evolution—Research for harmonised and Transitional water Observation), grant number 870349, and by the European Commission through the EU Horizon-2020 project DOORS (Developing Optimal and Open Research Support for the Black Sea), grant number 101000518. The APC was funded by DOORS. G.S. was funded by a PhD grant awarded by Fundação para a Ciência e a Tecnologia (FCT) within the scope of the MIT Portugal Program.

**Data Availability Statement:** The data presented in this study are available online at <https://engage.certo-project.org/data/>, accessed on 11 June 2024. These data were derived from the following resources available in the public domain: <https://dataspace.copernicus.eu/>, accessed on 11 June 2024.

**Acknowledgments:** The authors very sincerely acknowledge the extensive support of the wider CERTO consortium, who were essential to the completion of this study. The manuscript was greatly improved by the comments and suggestions provided by three reviewers, for which the authors are grateful.

**Conflicts of Interest:** The authors declare no conflicts of interest. Although the author O.D. is affiliated with a commercial entity (Brockmann Consult GmbH), this does not alter their adherence to journal policies on sharing data and materials.

## References

1. Gattuso, J.-P.; Frankignoulle, M.; Wollast, R. Carbon and Carbonate Metabolism in Coastal Aquatic Ecosystems. *Annu. Rev. Ecol. Syst.* **1998**, *29*, 405–434. [\[CrossRef\]](#)
2. Giraud, X.; Quéré, C.L.; da Cunha, L.C. Importance of Coastal Nutrient Supply for Global Ocean Biogeochemistry. *Glob. Biogeochem. Cycles* **2008**, *22*, GB2025. [\[CrossRef\]](#)

3. Halpern, B.S.; Walbridge, S.; Selkoe, K.A.; Kappel, C.V.; Micheli, F.; D'Agrosa, C.; Bruno, J.F.; Casey, K.S.; Ebert, C.; Fox, H.E.; et al. A Global Map of Human Impact on Marine Ecosystems. *Science* **2008**, *319*, 948–952. [[CrossRef](#)] [[PubMed](#)]
4. Feist, B.E.; Levin, P.S. Novel Indicators of Anthropogenic Influence on Marine and Coastal Ecosystems. *Front. Mar. Sci.* **2016**, *3*, 113. [[CrossRef](#)]
5. Spyarakos, E.; O'Donnell, R.; Hunter, P.D.; Miller, C.; Scott, M.; Simis, S.G.H.; Neil, C.; Barbosa, C.C.F.; Binding, C.E.; Bradt, S.; et al. Optical Types of Inland and Coastal Waters. *Limnol. Oceanogr.* **2018**, *63*, 846–870. [[CrossRef](#)]
6. Mélin, F.; Vantrepotte, V. How Optically Diverse Is the Coastal Ocean? *Remote Sens. Environ.* **2015**, *160*, 235–251. [[CrossRef](#)]
7. Moore, T.S.; Dowell, M.D.; Bradt, S.; Verdu, A.R. An Optical Water Type Framework for Selecting and Blending Retrievals from Bio-Optical Algorithms in Lakes and Coastal Waters. *Remote Sens. Environ.* **2014**, *143*, 97–111. [[CrossRef](#)]
8. Xue, K.; Ma, R.; Wang, D.; Shen, M. Optical Classification of the Remote Sensing Reflectance and Its Application in Deriving the Specific Phytoplankton Absorption in Optically Complex Lakes. *Remote Sens.* **2019**, *11*, 184. [[CrossRef](#)]
9. Jerlov, N.G. Classification of Sea Water in Terms of Quanta Irradiance. *ICES J. Mar. Sci.* **1977**, *37*, 281–287. [[CrossRef](#)]
10. Jackson, T.; Sathyendranath, S.; Mélin, F. An Improved Optical Classification Scheme for the Ocean Colour Essential Climate Variable and Its Applications. *Remote Sens. Environ.* **2017**, *203*, 152–161. [[CrossRef](#)]
11. Eleveld, M.A.; Ruescas, A.B.; Hommersom, A.; Moore, T.S.; Peters, S.W.M.; Brockmann, C. An Optical Classification Tool for Global Lake Waters. *Remote Sens.* **2017**, *9*, 420. [[CrossRef](#)]
12. Feng, H.; Campbell, J.W.; Dowell, M.D.; Moore, T.S. Modeling Spectral Reflectance of Optically Complex Waters Using Bio-Optical Measurements from Tokyo Bay. *Remote Sens. Environ.* **2005**, *99*, 232–243. [[CrossRef](#)]
13. Jia, T.; Zhang, Y.; Dong, R. A Universal Fuzzy Logic Optical Water Type Scheme for the Global Oceans. *Remote Sens.* **2021**, *13*, 4018. [[CrossRef](#)]
14. Wei, J.; Wang, M.; Mikelsons, K.; Jiang, L.; Kratzer, S.; Lee, Z.; Moore, T.; Sosik, H.M.; Van der Zande, D. Global Satellite Water Classification Data Products over Oceanic, Coastal, and Inland Waters. *Remote Sens. Environ.* **2022**, *282*, 113233. [[CrossRef](#)]
15. Zhang, F.; Li, J.; Shen, Q.; Zhang, B.; Wu, C.; Wu, Y.; Wang, G.; Wang, S.; Lu, Z. Algorithms and Schemes for Chlorophyll a Estimation by Remote Sensing and Optical Classification for Turbid Lake Taihu, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 350–364. [[CrossRef](#)]
16. Bi, S.; Li, Y.; Liu, G.; Song, K.; Xu, J.; Dong, X.; Cai, X.; Mu, M.; Miao, S.; Lyu, H. Assessment of Algorithms for Estimating Chlorophyll-a Concentration in Inland Waters: A Round-Robin Scoring Method Based on the Optically Fuzzy Clustering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [[CrossRef](#)]
17. Hieronymi, M.; Müller, D.; Doerffer, R. The OLCI Neural Network Swarm (ONNS): A Bio-Geo-Optical Algorithm for Open Ocean and Coastal Waters. *Front. Mar. Sci.* **2017**, *4*, 140. [[CrossRef](#)]
18. Lubac, B.; Loisel, H. Variability and Classification of Remote Sensing Reflectance Spectra in the Eastern English Channel and Southern North Sea. *Remote Sens. Environ.* **2007**, *110*, 45–58. [[CrossRef](#)]
19. Shi, K.; Li, Y.; Zhang, Y.; Li, L.; Lv, H.; Song, K. Classification of Inland Waters Based on Bio-Optical Properties. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 543–561. [[CrossRef](#)]
20. Vantrepotte, V.; Loisel, H.; Dessailly, D.; Mériaux, X. Optical Classification of Contrasted Coastal Waters. *Remote Sens. Environ.* **2012**, *123*, 306–323. [[CrossRef](#)]
21. Hommersom, A.; Wernand, M.R.; Peters, S.; Eleveld, M.A.; van der Woerd, H.J.; de Boer, J. Spectra of a Shallow Sea—Unmixing for Class Identification and Monitoring of Coastal Waters. *Ocean Dyn.* **2011**, *61*, 463–480. [[CrossRef](#)]
22. Ye, H.; Li, J.; Li, T.; Shen, Q.; Zhu, J.; Wang, X.; Zhang, F.; Zhang, J.; Zhang, B. Spectral Classification of the Yellow Sea and Implications for Coastal Ocean Color Remote Sensing. *Remote Sens.* **2016**, *8*, 321. [[CrossRef](#)]
23. Mustapha, Z.B.; Alvain, S.; Jamet, C.; Loisel, H.; Dessailly, D. Automatic Classification of Water-Leaving Radiance Anomalies from Global SeaWiFS Imagery: Application to the Detection of Phytoplankton Groups in Open Ocean Waters. *Remote Sens. Environ.* **2014**, *146*, 97–112. [[CrossRef](#)]
24. Neil, C.; Spyarakos, E.; Hunter, P.D.; Tyler, A.N. A Global Approach for Chlorophyll-a Retrieval across Optically Complex Inland Waters Based on Optical Water Types. *Remote Sens. Environ.* **2019**, *229*, 159–178. [[CrossRef](#)]
25. Fortunato, A.B.; Baptista, A.M.; Luettich, R.A. A Three-Dimensional Model of Tidal Currents in the Mouth of the Tagus Estuary. *Cont. Shelf Res.* **1997**, *17*, 1689–1714. [[CrossRef](#)]
26. Neves, F.D.S. Dynamics and Hydrology of the Tagus Estuary: Results from In Situ Observations. Ph.D. Thesis, Ciências Geofísicas e da Geoinformação (Oceanografia), Universidade de Lisboa, Faculdade de Ciências, Lisbon, Portugal, 2010.
27. Freitas, M.C.; Andrade, C.; Cruces, A.; Munhá, J.; Sousa, M.J.; Moreira, S.; Jouanneau, J.M.; Martins, L. Anthropogenic Influence in the Sado Estuary (Portugal): A Geochemical Approach. *J. Iber. Geol.* **2008**, *34*, 271–286.
28. Neto, J.; Caçador, I.; Caetano, M.; Chaínho, P.; Costa, L.; Gonçalves, A.; Pereira, L.; Pinto, L.; Ramos, J.; Seixas, S. Capítulo 16: Estuários. In *Rios de Portugal—Comunidades, Processos e Alterações*; Coimbra University Press: Coimbra, Portugal, 2019; pp. 381–421.
29. Toming, K.; Kutser, T.; Laas, A.; Sepp, M.; Paavel, B.; Nöges, T. First Experiences in Mapping Lake Water Quality Parameters with Sentinel-2 MSI Imagery. *Remote Sens.* **2016**, *8*, 640. [[CrossRef](#)]
30. Sent, G.; Biguino, B.; Favareto, L.; Cruz, J.; Sá, C.; Dogliotti, A.I.; Palma, C.; Brotas, V.; Brito, A.C. Deriving Water Quality Parameters Using Sentinel-2 Imagery: A Case Study in the Sado Estuary, Portugal. *Remote Sens.* **2021**, *13*, 1043. [[CrossRef](#)]



31. Salama, M.S.; Spaias, L.; Poser, K.; Peters, S.; Laanen, M. Validation of Sentinel-2 (MSI) and Sentinel-3 (OLCI) Water Quality Products in Turbid Estuaries Using Fixed Monitoring Stations. *Front. Remote Sens.* **2022**, *2*, 808287. [\[CrossRef\]](#)
32. Simis, S.; Stelzer, K.; Müller, D.; Selmes, N.; Warren, M. Copernicus Global Land Operations 'Cryosphere and Water'. Copernicus Global Land Operations—Lot 2. 2020, p. 46. Available online: <https://land.copernicus.eu/en/technical-library/algorithm-theoretical-basis-document-lake-water-quality-v1.0/@download/file> (accessed on 29 August 2024).
33. Warren, M.A.; Simis, S.G.H.; Martinez-Vicente, V.; Poser, K.; Bresciani, M.; Alikas, K.; Spyarakos, E.; Giardino, C.; Ansper, A. Assessment of Atmospheric Correction Algorithms for the Sentinel-2A MultiSpectral Imager over Coastal and Inland Waters. *Remote Sens. Environ.* **2019**, *225*, 267–289. [\[CrossRef\]](#)
34. Steinmetz, F.; Deschamps, P.-Y.; Ramon, D. Atmospheric Correction in Presence of Sun Glint: Application to MERIS. *Opt. Express* **2011**, *19*, 9783. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Steinmetz, F.; Ramon, D. Sentinel-2 MSI and Sentinel-3 OLCI Consistent Ocean Colour Products Using Polymer. *Remote Sens. Open Coast. Ocean Inland Waters* **2018**, 10778, 107780E. [\[CrossRef\]](#)
36. Mograne, M.A.; Jamet, C.; Loisel, H.; Vantrepotte, V.; Mériaux, X.; Cauvin, A. Evaluation of Five Atmospheric Correction Algorithms over French Optically-Complex Waters for the Sentinel-3A OLCI Ocean Color Sensor. *Remote Sens.* **2019**, *11*, 668. [\[CrossRef\]](#)
37. Giannini, F.; Hunt, B.P.V.; Jacoby, D.; Costa, M. Performance of OLCI Sentinel-3A Satellite in the Northeast Pacific Coastal Waters. *Remote Sens. Environ.* **2021**, *256*, 112317. [\[CrossRef\]](#)
38. Liu, X.; Steele, C.; Simis, S.; Warren, M.; Tyler, A.; Spyarakos, E.; Selmes, N.; Hunter, P. Retrieval of Chlorophyll-a Concentration and Associated Product Uncertainty in Optically Diverse Lakes and Reservoirs. *Remote Sens. Environ.* **2021**, *267*, 112710. [\[CrossRef\]](#)
39. Bi, S.; Li, Y.; Xu, J.; Liu, G.; Song, K.; Mu, M.; Lyu, H.; Miao, S.; Xu, J. Optical Classification of Inland Waters Based on an Improved Fuzzy C-Means Method. *Opt. Express* **2019**, *27*, 34838. [\[CrossRef\]](#)
40. Xie, X.L.; Beni, G. A Validity Measure for Fuzzy Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 841–847. [\[CrossRef\]](#)
41. Bezdek, J.C. Cluster Validity with Fuzzy Sets. *J. Cybern.* **1973**, *3*, 58–73. [\[CrossRef\]](#)
42. Dave, R.N. Validating Fuzzy Partitions Obtained through C-Shells Clustering. *Pattern Recognit. Lett.* **1996**, *17*, 613–623. [\[CrossRef\]](#)
43. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley: New York, NY, USA, 2005; ISBN 978-0-471-73578-6.
44. Campello, R.; Hruschka, E. A Fuzzy Extension of the Silhouette Width Criterion for Cluster Analysis. *Fuzzy Sets Syst.* **2006**, *157*, 2858–2875. [\[CrossRef\]](#)
45. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, PAMI-1, 224–227. [\[CrossRef\]](#)
46. Deza, E.; Deza, M.M. *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009; ISBN 9783642002335.
47. Brereton, R.G. The Mahalanobis Distance and Its Relationship to Principal Component Scores. *J. Chemom.* **2015**, *29*, 143–145. [\[CrossRef\]](#)
48. Moore, T.S.; Campbell, J.W.; Dowell, M.D. A Class-Based Approach to Characterizing and Mapping the Uncertainty of the MODIS Ocean Chlorophyll Product. *Remote Sens. Environ.* **2009**, *113*, 2424–2430. [\[CrossRef\]](#)
49. Moore, T.S.; Campbell, J.W.; Feng, H. A Fuzzy Logic Classification Scheme for Selecting and Blending Satellite Ocean Color Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 1764. [\[CrossRef\]](#)
50. IOCCG. *Partition of the Ocean into Ecological Provinces: Role of Ocean-Color*; Dowell, M., Platt, T., Eds.; International Ocean Color Coordinating Group (IOCCG): Dartmouth, NS, Canada, 2009; p. 99.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.