Environmental Data Science (2025), 4: e32, 1–17
doi:10.1017/eds.2025.10005

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Classification-informed estimation: the role of water-type clustering to improve neural network generalization for salinity and temperature estimation in coastal waters

Solomon White[1], Encarni Medina-Lopez[1], Tiago Silva[2], Evangelos Spyrakos[3], Laurent Amoudry[4] and Adrien Martin[5]

[1]School of Infrastructure and Environment, University of Edinburgh, Edinburgh, UK
[2]Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Lowestoft, UK
[3]Biological and Environmental Sciences, University of Stirling, Stirling, UK
[4]Marine Physics and Ocean Climate, National Oceanography Centre (NOC), Southampton, UK.
[5]Noveltis, Labège, France
Corresponding author: Solomon White; Email: solomon.white@ed.ac.uk

Received: 14 June 2024; **Revised:** 09 March 2025; **Accepted:** 09 May 2025

Keywords: classification; machine learning; oceanography; remote sensing; segmentation

Abstract

Sea surface salinity and temperature are essential climate variables in monitoring and modeling ocean health. Multispectral ocean color satellites allow the estimation of these properties at a resolution of 10 to 300 m, which is required to correctly represent their spatial variability in coastal waters. This paper investigates the effect of pre-applying an unsupervised classification in the performance of both temperature and salinity inversion. Two methodologies were explored: clustering based solely on spectral radiances, and clustering applied directly to satellite images. The former improved model generalization by identifying similar water clusters across different locations, reducing location dependency. It also demonstrated results correlating cluster type with salinity and temperature distributions thereby enhancing regression model performance and improving a global ocean color sea surface temperature regression model RMSE error by 10%. The latter approach, applying clustering directly to satellite images, incorporated spatial information into the models and enabled the identification of front boundaries and gradient information, improving global sea surface temperature models RMSE by 20% and sea surface salinity models by 30%, compared to the initial ocean color model. Beyond improving algorithm performance, optical water classification can be used to monitor and interpret changes to water optics, including algal blooms, sediment disturbance or other climate change or antropogenic disturbances. For example, the clusters have been used to show the impact of a category 4 hurricane landfall on the Mississippi estuarine region.

Impact Statement

Understanding sea surface temperature (SST) and salinity (SSS) is critical for monitoring ocean health, particularly in coastal regions where marine heatwaves, eutrophication, and freshwater influx impact ecosystems and coastal communities. However, traditional regression models struggle to generalize across geographically diverse regions, often misinterpreting conditions in data-scarce areas like the Global South. Our study addresses this challenge by integrating a K-means clustering algorithm into ocean color models to classify water types based on spectral and spatial features, independent of location. This approach improves model accuracy by narrowing the range of SST and SSS predictions within each cluster, reducing location-specific biases. For

example, the model can correctly distinguish upwelling zones or low-salinity plumes in previously unseen regions, improving global applicability. The use of K-means clustering, rather than computationally intensive convolutional neural networks (CNNs), enables large-scale, long-term ocean monitoring with lower computational demands. This democratizes access to advanced data analysis for researchers and institutions with limited resources, particularly in developing countries, while also reducing the carbon footprint of processing large Earth observation datasets.

1. Introduction

Ocean temperature and salinity are the main properties that control water density, and with it, water column stability. These properties affect biological activity directly as physiological drivers and stressors, and indirectly by controlling the spatial distribution of nutrients, dissolved oxygen and prey (Lee and Gentemann, 2017; Thakur et al., 2018). Understanding and monitoring water temperature and salinity is crucial, especially in the coastal ocean, where climate change is increasing the frequency and intensity of marine extreme events including marine heat waves (Li et al., 2022; Dai et al., 2023), leading to increased thermal and freshwater stratification (Li et al., 2020), eutrophication (Breitburg et al., 2018), and hypoxic events (Altieri and Gedan, 2015).

Multispectral ocean color models, trained using in-situ data enable monitoring of these essential climate variables (ECVs) at the required high spatial and temporal resolution (Muller-Karger et al., 2018; Medina-Lopez, 2019; Bergsma and Almar, 2020; Hadjal et al., 2022), particularly in coastal waters, where there is a mixture of complex natural and anthropogenic influences at smaller spatial scales: river plumes provide influxes of organic and inorganic material as well as pollution, upwelling over the shelf break provides cooler nutrient-rich water, and coastal habitats such as reefs and mangroves host complex and highly productive ecosystems (Dickey and Bidigare, 2005; Salisbury et al., 2011; Fournier et al., 2016).

However, ocean color models predicting sea surface properties such as sea surface salinity (SSS) and temperature (SST) from satellite images are usually trained using point-wise ground-based in-situ temperature or salinity data (O'Reilly and Werdell, 2019; Wei et al., 2023). Learning relationships between spectral signatures and water column constituent concentrations (Cael et al., 2020; Casey et al., 2020; White et al., 2024). This works at pixel level and does not include any neighborhood information which can inform on spatial distribution of features. The result is a loss of spatial information of the final product.

Applying a clustering or segmentation approach to the input image can capture this spatial information by relating pixels to each other by spectral similarity or proximity. This paper determines how using unsupervised learning to create clusters for water classification improves the performance of ocean color regression models through appropriate algorithm selection as well as retaining spatial information.

Ocean color remote sensing can be generally split into case 1 and case 2 waters. Case 1 waters are waters where the Inherent Optical Properties (IOPs) are dominated by phytoplankton, found mostly in blue open ocean waters. Case 2 waters are all others, where the IOPs are influenced by Color Dissolved Organic Matter (CDOM) and inorganic particles (Matsushita et al., 2012). Effects of the accuracy of ocean color algorithms have been shown to be reduced in highly turbid, low salinity waters (found in sediment-high rivers), as well as those in colder waters with ice mixing (Giannini et al., 2013; White et al., 2025). Classification, therefore, can be a useful tool to select the appropriate ocean color algorithms and atmospheric correction model to be applied in a certain water type (Frouin et al., 2019). Several optical water type classifications have been proposed based on surface relectance spectrum (Moore et al., 2014; Spyrakos et al., 2017, 2018). Empirical ocean color algorithms depend on inter-relationships of in-water optical constituents which change as a function of space and time in optically complex waters, therefore different spectral band algorithms are needed in different water types (e.g. highly turbid regions) to accurately estimate IOPs. Classification has been shown to improve

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

accuracy of inversion of Chlorophyll concentration through thresholding and model selection (Sun et al., 2014; Brewin et al., 2015).

In addition to improving algorithm performance, optical water classification has also been used to monitor the size and location of water types and hazardous events, such as harmful algal blooms (Medina-López et al., 2023). To that purpose these clusters can also be used purely for ocean type monitoring, by viewing cluster size and frequency change over time. This was applied in this paper in test regions to map inter-annual changes and was able to identify changes in estuarine plume size corresponding to damming upstream.

This study aims to improve the accuracy and robustness of predictive models applied to global ocean color data, as well as case studies in the Gulf of Mexico and the UK. The datasets exhibits significant variability due to diverse environmental conditions, ranging from freshwater regions to highly saline open ocean areas. To address this challenge, we employ K-means clustering to group water samples into distinct water types. Each cluster represents a specific environmental context, such as coastal, open ocean, or transitional zones.

This paper is structured as follows: i) introduce the ground-truth datasets (buoy data) and Sentinel-2 satellite imagery; ii) select the clustering algorithm and hyperparameter, iii) apply clustering algorithm to the spectral band point data, and then regions of the image that resulted from image segmentation iv) show results for the locations of the different cluster classes, as well as the dynamics between average spectral band for each cluster and the matching sea surface temperature and salinity cluster distribution.

2. Materials and methods

2.1. In situ and satellite datasets

This section introduces the matched in-situ and satellite data for two independent datasets from the UK and the Gulf of Mexico, as well as global SST and SSS from the Copernicus in-situ monitoring system (CMEMS TAC Data Team, 2021).

The global in-situ dataset comes from the Copernicus in situ Marine Environmental Monitoring Service (CMEMS) (with datasets coming from over 100 countries (European Commission, 2023). CMEMS provides pointwise data from various observing systems, including Argo float profiles, and observations from ships, moored buoys, drifting buoys, fixed platforms, gliders, ferry-boxes, and coastal observations (SeaDataNet, 2025). The Gulf of Mexico was chosen as a study region for its varied waters, including shallow coastal bays and deeper offshore regions, which provide an excellent opportunity to study the dynamics of coastal ecosystems. These areas are characterized by their proximity to land, complex bathymetry, and diverse marine habitats. The Gulf of Mexico Coastal Observing System (GCOOS) uses NOAA cruises and stations to monitor estuaries in the Gulf of Mexico (Jochens and Watson, 2013). UK smart buoy data is obtained from the Centre for Environment, Fisheries and Aquaculture Science of the UK (Cefas; Cefas, 2024).

The satellite used in this study is the European Space Agency (ESA) Sentinel-2 multispectral satellite system. Consisting of two satellites, Sentinel-2A and Sentinel-2B, it offers a revisit time of approximately 5 days at the equator, enabling frequent monitoring of Earth's surface in a polar orbit with a pixel spatial resolution of 10 m^2. The Sentinel-2 data comprise 13 spectral bands, each represented as 16-bit unsigned integers (UINT16) and scaled by a factor of 10,000 to obtain top-of-atmosphere (TOA) reflectance values. The spectral bands include: coastal aerosol (443 nm), red edge detection (705 nm) and near infrared band (842 nm), and a full list is available here: (European Space Agency, 2023).

2.2. Ocean color model

All of the in situ datasets underwent the same matching process with the multispectral satellite images. The Sentinel-2 data was processed on the Google Earth Engine Python API platform. This

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

allows geospatial analysis and processing of satellite images on Google Cloud computers, using scalable, high-performance computing resources. However, there is additional data transfer costs and the inherent risks from depending on a third party provider (Google Earth Engine, 2023). Latitude, longitude and time of measurement are taken for each in-situ data point. The Sentinel-2 image collection is filtered to the tiles that contain the point on the day and time when the measurement was taken, within one hour of Sentinel-2 overpass. One hour was chosen to improve accuracy, especially in the coastal zones where tidal effects and river flows can vary significantly over the course of hours.

The matched images for those points are clipped in 3x3 pixel windows about the in-situ data point, to retain the high resolution benefits from the Sentinel-2 satellite. The whole 3x3 pixel window is taken to avoid any random noise reflectance, wave effects or sun-glint errors occurring at the pixel level. The median value of the window for the spectral bands and metadata is then selected for the matched point. Time difference is also recorded between the in-situ measurement and satellite image, if there are multiple images within 1 hour of the in-situ point the smallest time difference is selected. The output is a table containing all satellite data (spectral bands and metadata properties) with the corresponding salinity and temperature data.

Various machine learning models were trained on the matched satellite and in-situ data with the inputs being the spectral bands and metadata and predicting SSS and SST independently as output. In this study, a feed-forward neural network architecture with multiple hidden layers (10) was designed, with a Tanh activation function to for propagating non-linearities, to capture the complex second order relationships between spectral signatures and the water physical properties. 18 nodes were used for the neural network initial layer to match the input data of bands and metadata. Dropout regularization helped prevent overfitting, while early stopping, based on validation set performance, further ensured that the model did not overtrain on the noise in the data. Additionally, the max-pooling layers, improved model ability to extract relevant information and improve generalization performance. Training the network with stochastic gradient descent (SGD) and optimizing with mean squared error (MSE) allowed for more efficient convergence over the course of 1000 epochs, resulting in a model that outperformed traditional methods like gradient boosting and XGBoost. The neural network's ability to capture complex, non-linear dependencies between features made it the most effective approach for this study.

All models were trained using a 70% training and 30% testing data split. The training data was further subjected to k-fold cross-validation with $k = 5$, where the model was trained on 4 folds and validated on the remaining fold, rotating through all folds. Early stopping and hyperparameter tuning were conducted using the validation folds within this k-fold process.

The models were optimized using RMSE as the primary error metric and evaluated on RMSE, RME, MAE, and MPSE. To ensure a robust assessment of model generalizability, certain buoy locations were completely withheld from both training and validation datasets, serving as an independent test set. The figures and metrics in the results section are derived from this unseen test data, providing an evaluation of the model's real-world performance.

2.3. Unsupervised machine learning (clustering)

Unsupervised learning, particularly clustering, is a crucial method in data analysis for uncovering intrinsic patterns within datasets without predefined labels (Caron et al., 2018). In clustering, the goal is to group similar data points based on certain similarity metrics, by optimizing an objective function, bringing coherence to the data and enabling exploratory analysis. In remote sensing, unsupervised clustering is particularly useful for analyzing multispectral or hyperspectral imagery. Algorithms like K-means help reveal patterns in spectral signatures without the need for labeled training data (Melgani and Pasolli, 2013; Naeini et al., 2014). The algorithm classifies pixels into "K" clusters based on spectral similarities, revealing inherent patterns in the data. Other clustering methods include hierarchical clustering, which builds a tree-like structure of nested clusters, iteratively merging or splitting clusters based on their

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

pairwise dissimilarities; or DBSCAN, which identifies dense regions of data points separated by sparser areas. These approaches are particularly valuable in scenarios where labeled training data for supervised methods is scarce.

In this paper, we test four clustering algorithms: K-means, Fuzzy C-Means (FCM), Spectral Clustering, and Hierarchical Clustering.

2.3.1. K-means

K-means is a popular unsupervised machine learning algorithm used for clustering data into K distinct groups or clusters (Jin and Han, 2008). The algorithm aims to partition n data points into K clusters in such a way that the within-cluster variation (or inertia) is minimized. Given a set of data points $X = \{x_1, x_2, \ldots, x_n\}$ and K cluster centroids $C = \{c_1, c_2, \ldots, c_k\}$, the objective is to assign each data point to the cluster with the nearest centroid, minimizing, for example, the objective function presented in Eq. 2.1. The process then iteratively assigns data points to the cluster with the nearest centroid and updates the centroid based on the mean of the assigned points. This process continues until convergence occurs, resulting in K clusters characterized by their centroids.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{1}$$

k is the number of clusters, C_i is the ith cluster, μ_i is the centroid (mean) of cluster C_i.

K-means is simple and easy to implement and can result in fast convergence. However, the effectiveness hinges on the appropriate choice of K, which can be determined by silhouette analysis, or the elbow method to find the optimal number of clusters (Yuan and Yang, 2019; Umargono et al., 2020). Silhouette analysis gives an idea of the separation distance between resulting clusters (Wang et al., 2017). The silhouette coefficients have a range from (SeaDataNet, 2025) with +1 indicating the sample is far from the neighboring clusters and − 1 indicating a datapoint may be assigned to the wrong cluster. It can as well be sensitive to the initial centroid locations, therefore validation methods such as K cross fold validation are useful to avoid fitting local minimums (Mayo, 2022).

2.4. Clustering for classification of water types

Two main methodologies are tested in this paper. The first approach undertakes clustering purely on spectral radiances. Although this does not contain spatial information, this helps improve the generalization of the models by learning similar clusters in unfamiliar test locations, reducing location dependencies: e.g. if the water classification in a region in Patagonia with no in-situ training data is the same as a UK region with ground-truthed data, the model will infer similar distributions of sea surface properties improving regression estimations.

The clustering of the matched satellite data was done solely using the spectral bands as inputs to the clustering algorithms. Different clustering methods were trialed, K-means, Fuzzy C-means, Spectral clustering, and hierarchical clustering. K-means was selected due to the computational speed and resistance to outliers, good scalability and only needing one hyperparameter: K, the number of clusters. Principal component analysis (PCA) was applied to the 13 spectral bands, for dimensionality reduction, improving estimations with noise and effectiveness of the K-means clustering (Ding and He, 2004), with 95% and 99% variance resulting in 1 and 4 resultant bands respectively.

The number of optimal clusters (K) was determined, using a mixture of silhouette analysis and the elbow method. The 99% variance PCA with number of cluster of 9 had the highest silhouette score with 0.99. Figure 1 shows the silhouette plot for each cluster in the 9 cluster selection, as well as the visualization of the clusters for the first two features (B1 and B2). This also corresponded with distinct corners on the Elbow method graphs for both inertia and distortion. Other studies on water type classification, set 21 clusters for water types of which 9 were selected for coastal waters so our value of 9 agrees with this decision (Spyrakos et al., 2018).

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

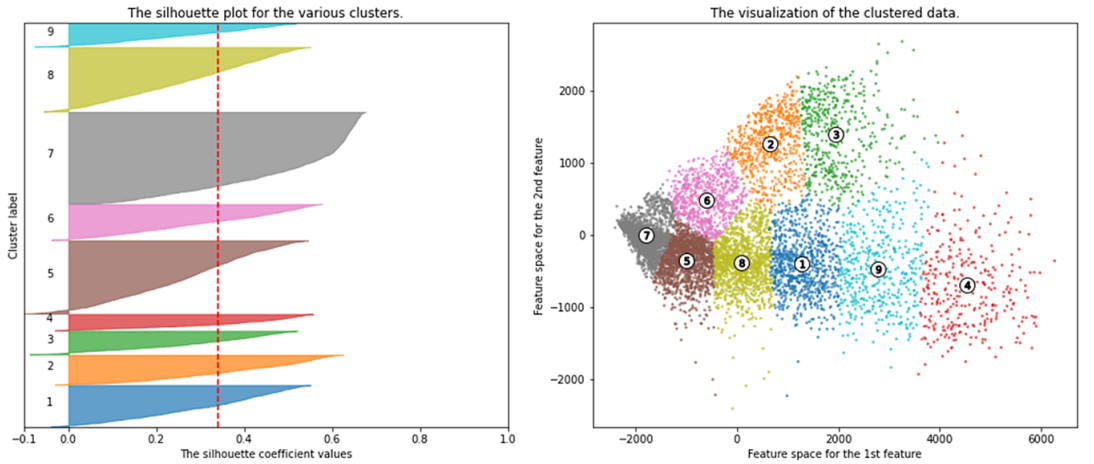


Figure 1. *Silhouette score and cluster visualization for number of clusters = 9. Feature space visualization for first and second principle components with PCA variance 99% (number of components = 5).*

K means clustering can be done in a number of ways—here the models are trained on pointwise matched spectral data from Sentinel-2. Latitude and longitude data were not used in the K means model to avoid overfitting problems and the model not being able to extrapolate to unseen locations. Furthermore, the link between location and SSS and SST is not often first order—e.g. a fresh water river further south maybe colder and fresher than "typical" sea water, so this location data could hinder the later model accuracy.

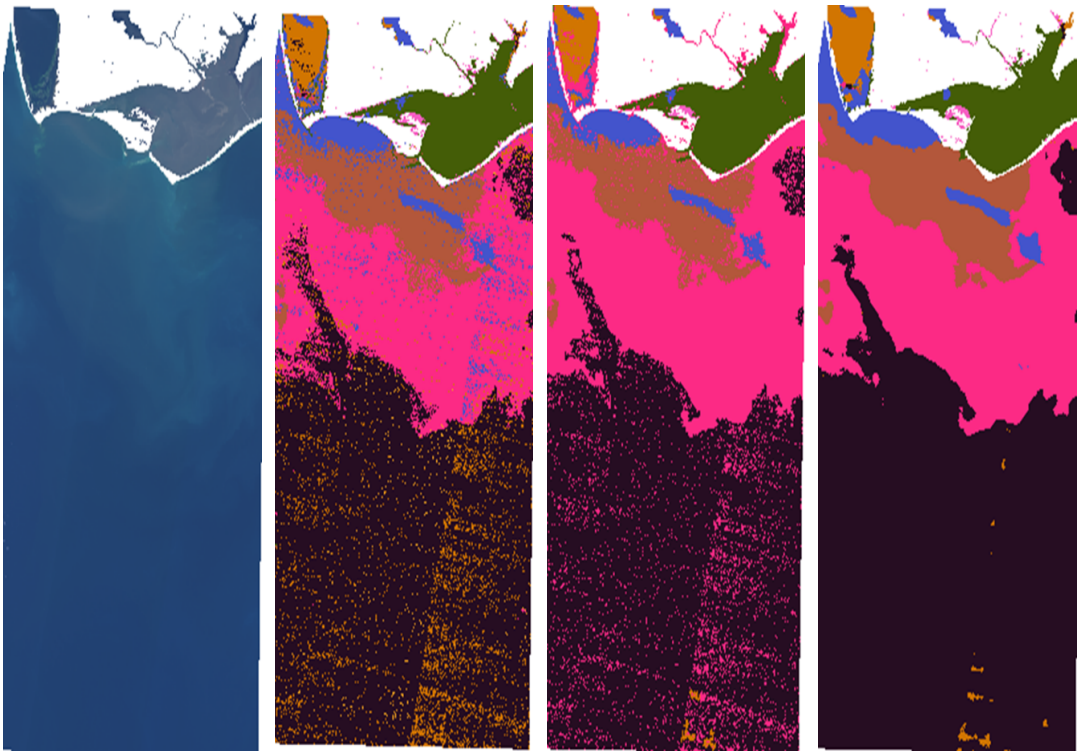
2.5. Image-based clustering approach

The first approach applies K-means clustering to the matched pointwise spectral data, which while separating the data into spectra based types such as seen from water color e.g. brown estuarine sediment, does not contain any spatial information from the satellite image. The second approach applies the clustering directly to the satellite image to incorporate the spatial information throughout the models. The image was segmented into distinct clusters based on proximity as well as spectral coherence. As before these clusters reduced variability of the SST and SSS distributions for each cluster, but also allowed for identification of front boundaries which can be fed into the model training alongside gradient information.

K means uses image neighborhood data based on the similarity between pixels—this means that although location data is not included—two pixels which are close together can be grouped in the same class. This is useful for the model to pick out water features such as a river plume or an upwelling—which have distinct SST and SSS properties. While we explored the Segment Anything Model (SAM) (Kirillov et al., 2023), we encountered challenges when applying it to multispectral Sentinel-2 image data. SAM, originally trained on RGB images, struggled to adapt effectively, resulting in numerous overlapping segments—particularly in ocean regions. Additionally, we experimented with convolutional neural networks (CNNs), but due to the absence of ground truth for the clusters, the computational costs outweighed the potential benefits.

First the satellite image, if coastal, used Sentinel-1 radar to mask out any land. K-means with 9 clusters were trained on >100 satellite images, covering a full range of seasonality, coastal environments, and latitudes. The matched satellite image, with an overpass time within one hour of an in-situ ground truth measurement, was clipped to 1 km about the measurement point to ensure the capture of any submesoscale processes and the trained K-means algorithm applied to the clipped image. The original K-means algorithm was tuned to allow more homogenous clusters to be formed. Figure 2 shows an example

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press



(a) Original Satellite image (b) Clustered image (c) Merged clustered image (d) Dilation clustered

Figure 2. *Clustered images, showing the coast adjacent to the Apalachicola River Wildlife and Environmental Area. With land masked in white. Clusters show agreement between different processes with less noise in the dilation clustering. a) Original Image, b) Simple clustered image, c) Merged clustered image (with minimum cluster size 25 pixels), d) Dilation smoothed clustered image.*

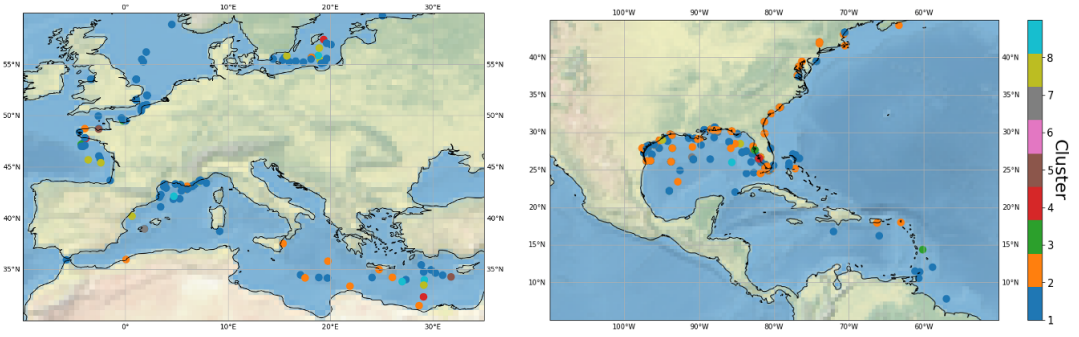
satellite image, with the original clustering applied (2b), this shows large amounts of speckled noise with heterogeneous clustering. Merged clustering was developed, (2c) to refine the clustering, defining a minimum cluster size (25 pixels) where clusters that are smaller than the defined minimum are identified and merged into their neighboring larger clusters. This removed some smaller noisy clusters but struggled with image boundaries and linear clusters. Dilation and erosion operations are used to merge noisy clusters, acting as a smoothing post-processing step. The "focal mode" image filtering function is used with a radius of 3 and 1 iteration for smoothing. Figure 2d shows the dilation smoothed clustering with uniform clusters agreeing with both image inspection and underlying physical processes. These all are post-processing steps to the K means clustered image. We also considered the smoothing of the input satellite image, but this increases coastal errors and reduces the benefits of the initial high-resolution satellite data.

3. Results

Figure 3 shows the locations and distributions of the image-based segmented clustering. Figure 3a shows the clustering in the Mediterranean Sea with cluster 1 as the dominant cluster. Figure 3b shows the North Atlantic coastline and the Gulf of Mexico cluster locations with a split between cluster 2 in the very coastal regions moving to cluster 1 in far shore conditions.

Figure 4 shows the 3D visualization for the 9 clusters against the input features, (e.g., B3, B2, B1, and other combinations). Figure 4a shows the visualization for the K-means algorithm purely done on spectral

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press



(a) Cluster locations in Europe. (b) Cluster locations around Gulf of Mexico.

Figure 3. *Cluster spread from the Global dataset focused on the Mediterranean (a) and the North-West Atlantic coast and Gulf of Mexico (b).*

bands. There is clear cluster separation in B3 to B7, showing the link between ocean color and cluster class, compared to the noisier variables in the SWIR range. Figure 4b shows the visualization of the bands for the image-based K-means, while there is a similar separation of the features by cluster it is less defined showing the reliance of the image-based algorithm on other input features such as neighborhood information.

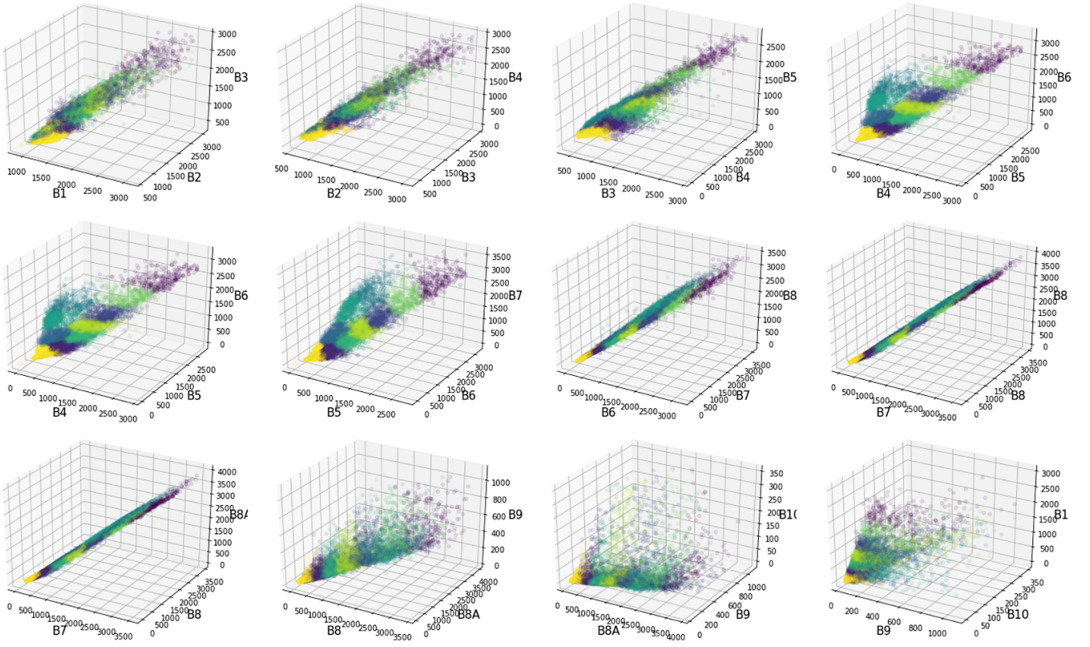
Figure 5 shows the mean cluster spectra and corresponding SST and SSS distributions for the global dataset and the two case studies in the Gulf of Mexico and the UK. The case studies in Mexico and the UK waters show very correlated spectra compared to the variety of spectra seen in the global dataset. This can also be seen reflected in the kernel density functions representing the distributions for temperature in each cluster. SSS distributions vary more depending on cluster reflecting the closer link between spectra and salinity variability measurements across different oceanic regions. This follows from the close relationship salinity has with ocean color from colored dissolved organic matter and chlorophyll. The SSS distributions for clusters 1 and 2 for Mexico (Figure 5f correlate well with the locations seen in Figure 3b, with the nearshore cluster 2 having a freshwater peak compared to the more saline open ocean cluster 1. The UK region stands out, it only detected 2 classes, clusters 1 and 2. It also has extremely similar SSS kernel density functions (KDFs), due to its low variability in salinity data. Testing showed good results with Pearson correlation coefficient, cosine similarity, and spectral angle mapping all showing separations of SST and SSS distributions purely based on spectral cluster type.

3.1. Cluster impact on neural network SST and SSS prediction

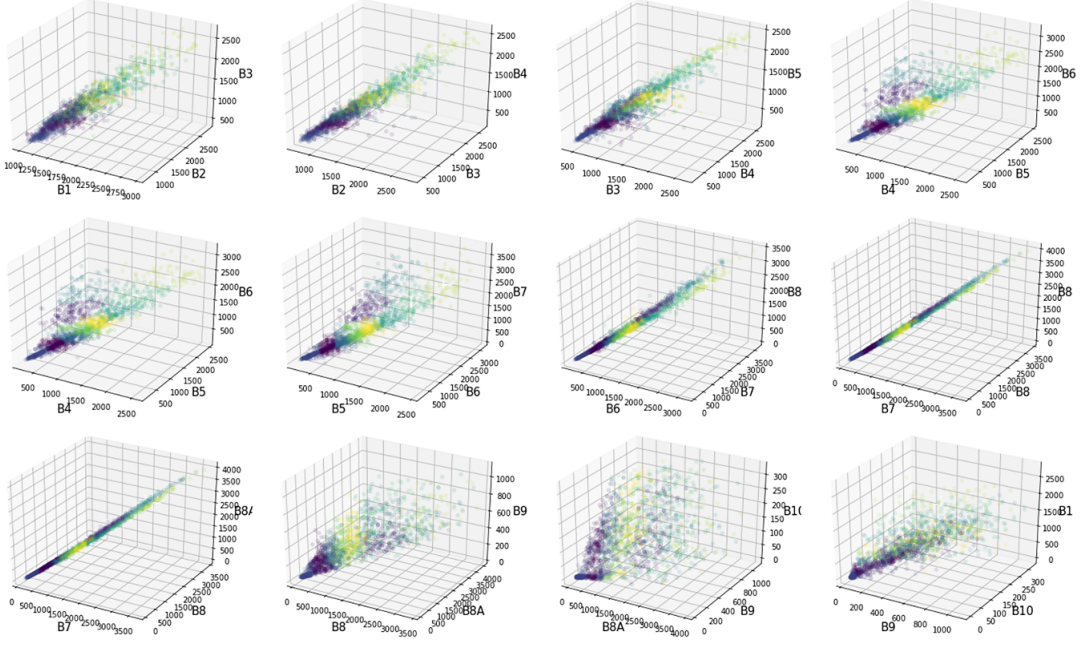
The plots in Figure 6 show the spread for the neural network predicted SST against actual values for each cluster for the global dataset. Where the model is trained only on each individual cluster. By training separate models on these individual water types, we reduce the variable range within each subset. Consequently, models can achieve better performance, as they focus on more homogeneous data segments. There is an improvement overall compared to the model with no cluster data impact as seen in Table 1 which shows the model results for RMSE, R^2 , and the variable range for the predicted SST and SSS for original models, spectrally clustered models, and the image segmented clusters model.

Incorporating cluster data into the training phase of regression models enhances their performance by increasing the models' understanding of SST or SSS distributions within each cluster. As indicated in Table 1, the Root Mean Square Error (RMSE) and other accuracy metrics for the comprehensive model are presented for each location and case study, juxtaposed with the clustered data outcomes.

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press



(a) K-means clustering on pointwise spectral data.



(b) Satellite image segmentation resulting in K-means clustering

Figure 4. *K-means clustering shows the different cluster classes against different input feature groups, enabling visualization of the feature weight and spectral variability to the classified cluster. a) K-means on pointwise spectral data shows separation of clusters by band b) K-means on segmented image has less clear band importance.*

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

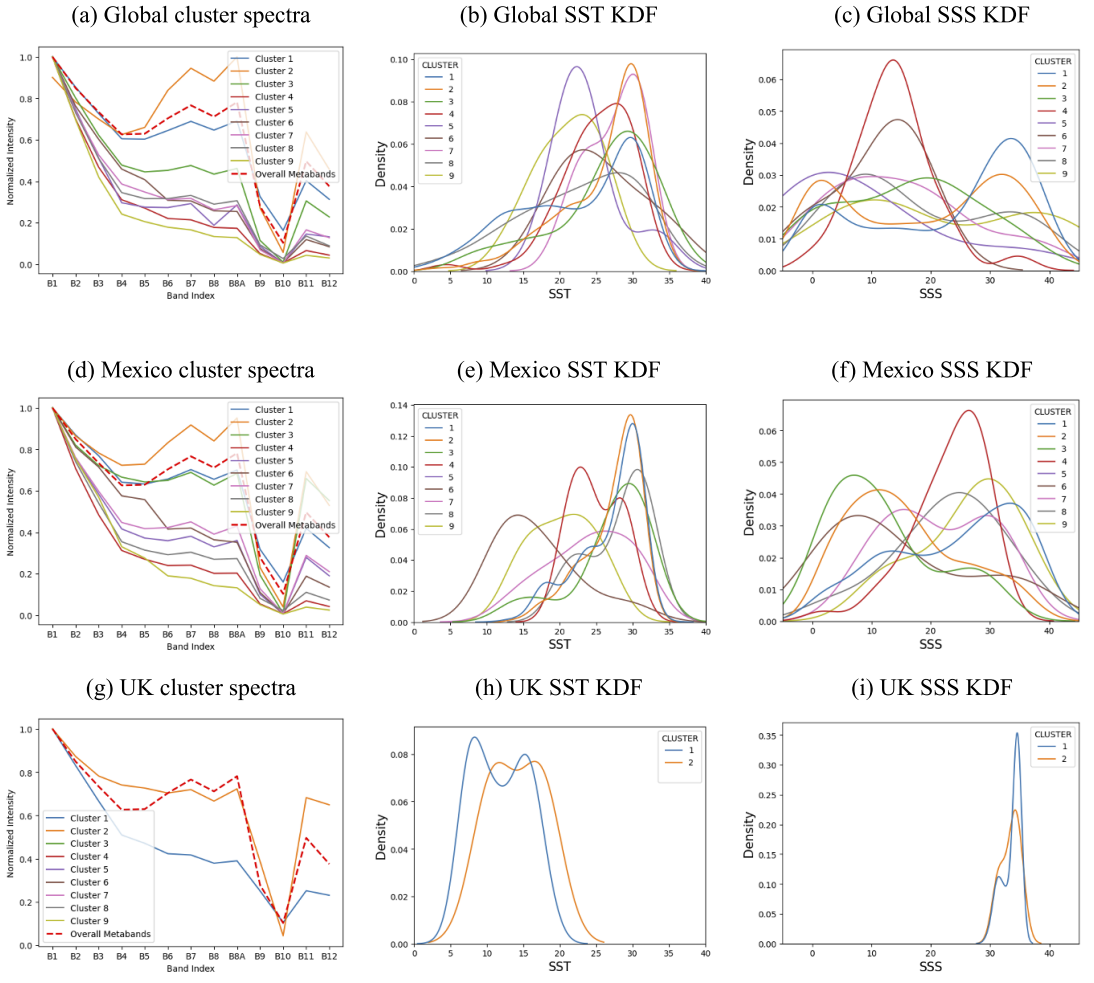


Figure 5. *Spectra for the different clusters for the global dataset (a), Gulf of Mexico (d), and the UK (g), and the corresponding kernel density functions for the sea surface temperature (°C) and salinity (PSU) distributions.*

The spectral clustered models for SSS and SST show a reduction in Train and test RMSE for the Global and Mexico dataset. Yet the Test RMSE values are higher than the initial models for the UK locations, suggesting that while the models fit the cluster data well, they may not be capturing the variability necessary for accurate predictions on the test set. The global model benefits from a reduction in prediction set variability as does the Mexico data, but this advantage is offset by the diminished training data, which adversely affects the models for the smaller UK regions. The UK SST model shows a higher Normalized Test RMSE for the cluster average, which could imply overfitting to the cluster data or that the clustered data does not represent the broader dataset well.

The clustering based on segmentation of the satellite image, it further improves the global SST and SSS RMSE as well as aiding the Mexico result. Improving global sea surface temperature models RMSE by 20% and sea surface salinity models by 30%, compared to initial ocean color model.

3.2. Image application and visualization

Figure 7 shows a true color Sentinel 2 image compared to the clustered image. The clustering was applied to the Mississippi River outflow in the Gulf of Mexico, with the black outline showing the

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

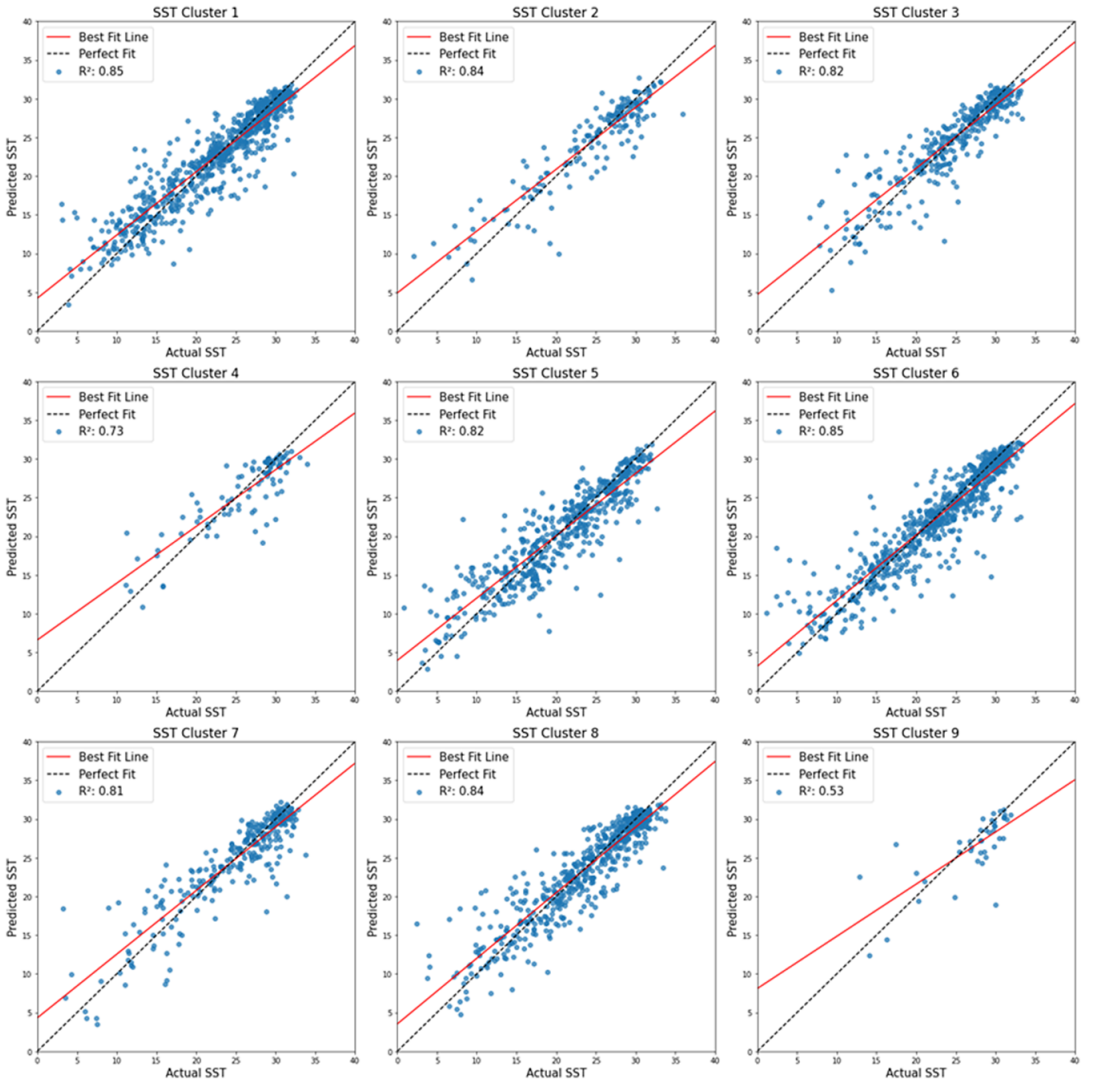


Figure 6. *Global Scatter plots of predicted against actual sea surface temperature for each segmented image model trained only on the selected cluster class.*

land and cloud masking applied to the pre-classified satellite image. The images were both taken in September, the first in 2022 (for reference of unaffected conditions) and the second in 2021, when there was significant flooding both upstream in the Mississippi delta and increased mixing due to storm winds from Hurricane Ida. As can be seen from the RGB images this flooding and storm activity has increased the size and distribution of the sediment plume. This is captured in the clustered images with the pink cluster (cluster 5) capturing the full extent of the sediment (and even some which are not visible in the true color image).

Figure 8 shows the average cluster area (in thousands of km^2) for the seasons in the years 2018–2023, in the region shown above in Figure 7. Cluster 5, as seen as the outflow class, is the most dominant in spring, but in summer, it is the 5th biggest cluster (showing the extremity of the flooding in Summer 2021 7d). In winter, with less outflow, cluster 8 (corresponding more to Case 1 open ocean waters) is the biggest class. Interestingly class 1 and class 2, appear the most in summer. These seem to be correlated with warmer, highly productive waters appearing around the delta region of the Mississippi.

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

Table 1. *Model performance metrics for overall sea surface temperature and salinity models, with the best-performing model per region highlighted in gray*

Model	Train RMSE	Test RMSE	Train R^2	Test R^2	Range Y
		SST			
Global initial	1.47	2.41	0.95	0.87	35.94
Global cluster Avg.	0.39	2.21	1.00	0.80	32.9
Global segmented image	0.52	**1.96**	1.00	0.82	15.34
Mexico initial	0.77	1.43	0.96	0.88	29.06
Mexico cluster avg.	0.20	1.78	1.00	0.79	23.28
Mexico segmented image	0.46	**1.29**	1.00	0.88	17.73
UK initial	0.27	**1.57**	1.00	0.84	16.87
UK cluster avg.	0.08	2.45	1.00	0.65	14.03
UK segmented image	0.07	1.89	1.00	0.72	12.46
		SSS			
Global SSS initial	3.93	4.88	0.86	0.85	51.98
Global cluster avg.	1.31	4.27	1.00	0.86	51.44
Global segmented image	2.78	**3.51**	1.00	0.86	43.95
Mexico SSS initial	3.45	3.71	0.90	0.89	38.97
Mexico cluster Avg.	0.03	3.42	1.00	0.61	38.45
Mexico segmented image	1.76	**2.60**	0.99	0.90	36.50
UK SSS initial	0.00	**0.43**	1.00	0.89	4.97
UK cluster avg.	0.22	0.92	1.00	0.28	4.7
UK segmented image	0.12	0.86	1.00	0.54	4.8

Note. Model performance metrics for overall sea surface temperature and salinity models, with the best-performing model per region highlighted in bold.

Figure 9 shows the correlation between the cluster classes and the SST distribution for a region in the Gulf of Mexico, there is a clear overlap between the warm cold front seen in the temperature plot compared to the plot of the cluster, which shows why there is such an improvement in the RMSE errors of the regression prediction model when the cluster spatial information is included.

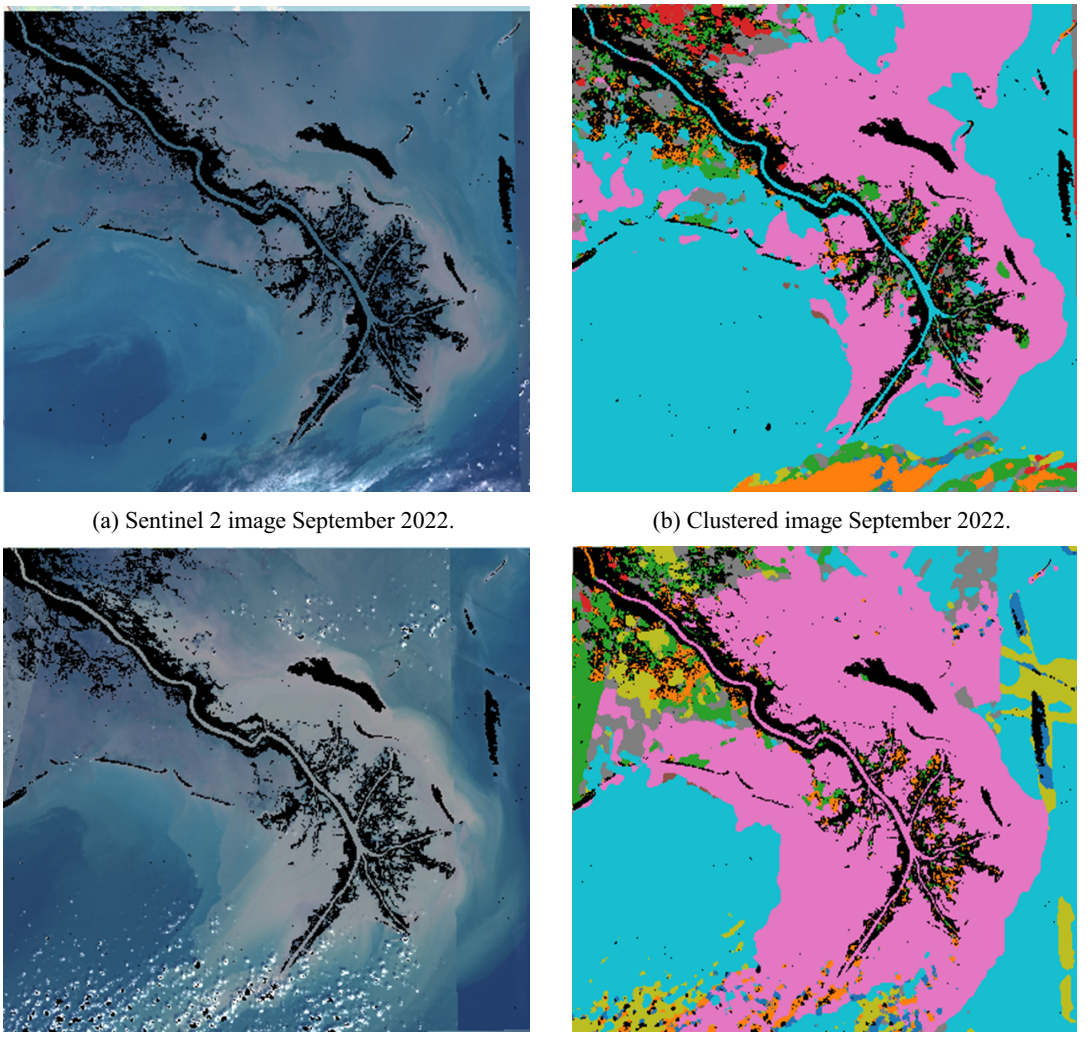
4. Conclusion

Implementing these classification algorithms has demonstrated an improvement in the regression model's Root Mean Square Error (RMSE), indicating enhanced predictive accuracy. Nonetheless, it is important to ensure that the training data encompasses a full distribution of all sea surface types. This comprehensive coverage is crucial for the accurate classification of specific water types. With purely spectral-based clustering, no additional information is added to the model, which is still trained on pointwise data, however, the multi-model approach can reduce the variance of each predicted variable. Applying the clustering directly to the satellite image in the segmentation approach, allows spatial information to be included in the model capturing physical processes.

The image-segmented clusters can also be used to track the impact of changes by anthropogenic or climate changes, both seasonally and interannually, to help guide policy and understand the challenges faced by ecosystems, and calculate the probability and extent of changes due to predicted climate warming.

A recurring issue with traditional ocean color temperature and salinity regression models is their inability to accurately handle regions that differ from the ones they are trained on. Specifically, when the model is used on data from a new location, it tends to incorrectly assume that the data distribution is the

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press



(a) Sentinel 2 image September 2022.

(b) Clustered image September 2022.

(c) Sentinel 2 image September 2021.

(d) Clustered image September 2021.

Figure 7. *Cluster visualization for Mississippi outflow in the Gulf of Mexico against a true color Sentinel 2 image. The images are from September 2022 and September 2021, when there was significant flooding due to the landfall of Hurricane Ida, a Category 4 storm.*

same as that of the training data. For instance, if the model was trained on data from the UK, it might predict an average temperature of 14 ° C, which is typical for the UK. However, when applied to data from Mexico, where the actual average temperature is 25 ° C, the model's prediction could be significantly off. This can lead to skewed results due to the uneven availability of training data, particularly from regions like the Global South. Classification can help to overcome these challenges, as the clustering was not based on location, it can pick out processes such as upwelling, and identify regions such as low saline or warmer water. The overarching goal was to achieve similar types of classification regardless of geographical location. However, the pre-classification of ocean color data can introduce a significant limitation on later regression models by constraining the model's output range. This could result in an over-reliance on simpler algorithms, such as K-means, to perform the heavy lifting of defining these constraints. Care must be taken to monitor the importance that deep learning models place on the clustering inputs (e.g., by weight tracking).

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

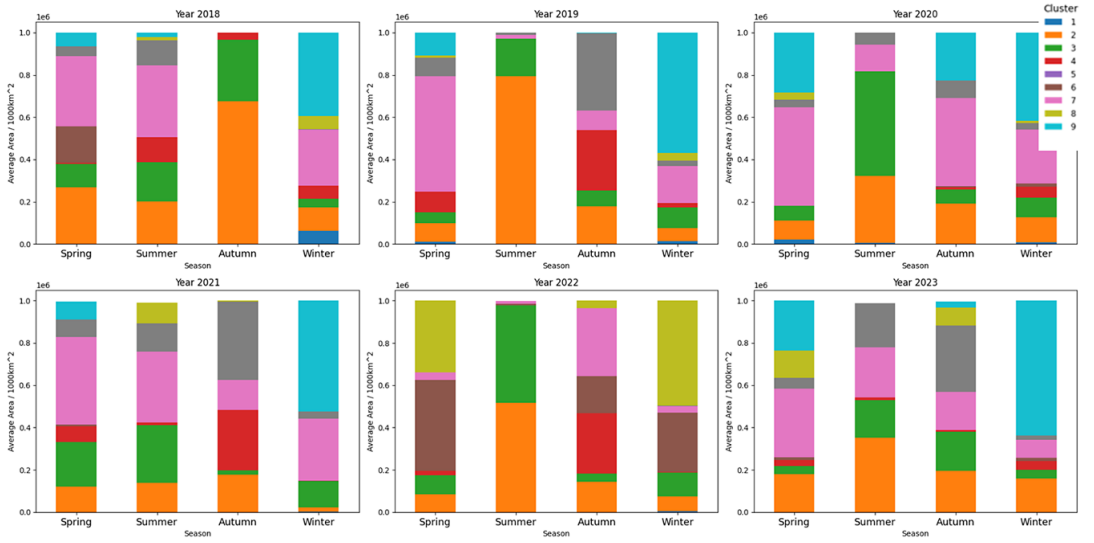
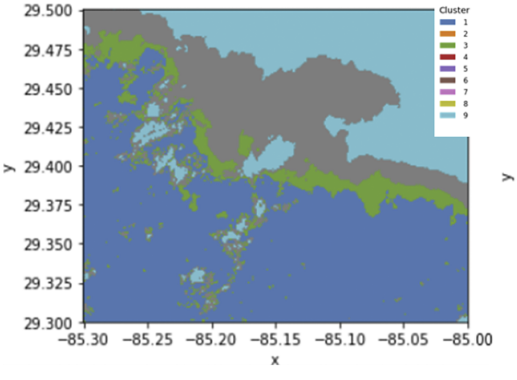
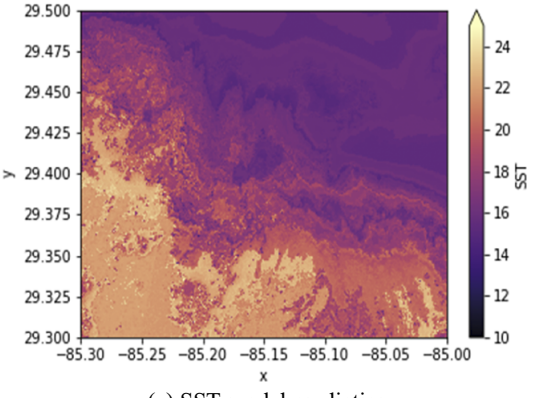


Figure 8. *Cluster size for each season in the years 2018–2023, for the Mississippi outflow region is shown in Figure 7.*

(a) Clustered region shown in black outlined box



(b) Image based cluster results.



(c) SST model prediction.

Figure 9. *Cluster class vs sea surface temperature in the Gulf of Mexico test region, adjacent to the Apalachicola River Wildlife and Environmental Area.*

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

Using a straightforward K-means algorithm and feeding its results into a deep neural network model enables complex spatial and image analysis without the computational cost and processing time associated with more sophisticated algorithms, such as convolutional neural networks. It is important to note that our segmentation method, which is based on clustering, only provides an approximation of water types such as 'river plume', 'turbid delta', 'open ocean', and so forth Therefore, employing a complex CNN for clustering would not necessarily yield a significant improvement in accuracy. Using this K-means algorithm means that the final model can be feasibly applied to large-scale datasets as seen in earth observation problems, for long-time series mapping and then applying the vectorized results for fast processing by the deep neural network model.

By reducing the computational resources required, the model becomes more accessible to countries that traditionally have limited access to supercomputers. This democratization of data analysis can lead to more globally inclusive research and findings. Furthermore, by requiring less computational power than climate models or other deep learning structures, the model helps reduce the energy consumption associated with data processing, thereby contributing to a lower carbon footprint.

However, while this simplicity enhances accessibility, it is important to acknowledge the tension between the algorithm's broad applicability and the constraints imposed by uneven data availability. Without sufficient regional data, especially from underrepresented regions (e.g. the global south), the model's accuracy and effectiveness may be compromised in these regions.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10005.

Author contribution. Supervision: A.M., E.M., E.S., L.A., T.S.; Conceptualization: S.W.

Competing interests. The authors declare none.

Data availability statement. All Data for this research is freely available: Global in-situ can be found from CMEMS for the Global dataset (European Commission, 2023), Gulf of Mexico data from the GCOOS ocean observation site (Jochens and Watson, 2013; Gulf of Mexico Coastal Ocean Observing System (GCOOS) (unknown), 2020), and Uk data from CEFAS smart buoys (Cefas, 2024). Sentinel 2 data can be accessed from Sentinel hub or Google Earth engine (Google Earth Engine, 2023).

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. SW benefited from a Sense CDT PhD studentship with additional CASE funding from the Centre for Environment Fisheries and Aquaculture Studies.

References

Altieri AH and Gedan KB (2015) Climate change and dead zones. *Global Change Biology 21*(4), 1395–1406.

Bergsma EW and Almar R (2020) Coastal coverage of ESA' Sentinel 2 mission. *Advances in Space Research 65*(11), 2636–2644.

Breitburg D, **Levin LA**, **Oschlies A**, **Grégoire M**, **Chavez FP**, **Conley DJ**, **Garçon V**, **Gilbert D**, **Gutiérrez D**, **Isensee K**, **Jacinto GS**, **Limburg KE**, **Montes I**, **Naqvi SWA**, **Pitcher GC**, **Rabalais NN**, **Roman MR**, **Rose KA**, **Seibel BA**, **Telszewski M**, **Yasuhara M and Zhang J** (2018) Declining oxygen in the global ocean and coastal waters. *Science*, *359*(6371), eaam7240.

Brewin RJ, **Raitsos DE**, **Dall'Olmo G**, **Zarokanellos N**, **Jackson T**, **Racault MF**, **Boss ES**, **Sathyendranath S**, **Jones BH and Hoteit I** (2015) Regional ocean-colour chlorophyll algorithms for the Red Sea. *Remote Sensing of Environment 165*, 64–85.

Cael BB, **Chase A and Boss E** (2020) Information content of absorption spectra and implications for ocean color inversion. *Applied Optics 59*(13), 3971.

Caron M, **Bojanowski P**, **Joulin A and Douze M** (2018) Deep clustering for unsupervised learning of visual features. *Computer Vision and Pattern Recognition arXiv preprint arXiv:1807.05520.*

Casey KA, **Rousseaux CS**, **Gregg WW**, **Boss E**, **Chase AP**, **Craig SE**, **Mouw CB**, **Reynolds RA**, **Stramski D**, **Ackleson SG**, **Bricaud A**, **Schaeffer B**, **Lewis MR and Maritorena S** (2020) A global compilation of in situ aquatic high spectral resolution inherent and apparent optical property data for remote sensing applications. *Earth System Science Data 12*, 1123–1139.

Cefas (2024) *Cefas UK Smart Buoy Website.* Available at https://www.cefas.co.uk/data-and-publications/smartbuoys/.

CMEMS TAC Data Team (2021) *Product User Manual for Multiparameter Copernicus In Situ TAC (PUM).*

Dai Y, **Yang S**, **Zhao D**, **Hu C**, **Xu W**, **Anderson DM**, **Li Y**, **Song X-P**, **Boyce DG**, **Gibson L**, **Zheng C and Feng L** (2023) Coastal phytoplankton blooms expand and intensify in the 21st century. *Nature 615*(7951), 280–284.

Dickey TD and Bidigare RR (2005) Interdisciplinary oceanographic observations: The wave of the future. *Scientia Marina 69*(Suppl 1), 23–42.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04*. New York. Association for Computing Machinery, pp. 29.

European Commission (2023). Copernicus Marine Environment Monitoring Service. http://marine.copernicus.eu/ (accessed 08th Oct 2020).

European Space Agency (2023). Sentinel-2 MSI —User Guides https://sentiwiki.copernicus.eu/web/s2-applications (accessed 25 October 2023).

Fournier S, **Lee T and Gierach MM** (2016) Seasonal and interannual variations of sea surface salinity associated with the Mississippi River plume observed by SMOS and Aquarius. *Remote Sensing of Environment 180*, 431–439.

Frouin RJ, **Franz BA**, **Ibrahim A**, **Knobelspiesse K**, **Ahmad Z**, **Cairns B**, **Chowdhary J**, **Dierssen HM**, **Tan J**, **Dubovik O**, **Huang X**, **Davis AB**, **Kalashnikova O**, **Thompson DR**, **Remer LA**, **Boss E**, **Coddington O**, **Deschamps PY**, **Gao BC**, **Gross L**, **Hasekamp O**, **Omar A**, **Pelletier B**, **Ramon D**, **Steinmetz F and Zhai PW** (2019) Atmospheric correction of Satellite Ocean-color imagery during the PACE era. *Frontiers in Earth Science 7*, 1–43.

Giannini MFC, **Garcia CAE**, **Tavano VM and Ciotti ÁM** (2013) Effects of low-salinity and high-turbidity waters on empirical ocean colour algorithms: An example for Southwestern Atlantic waters. *Continental Shelf Research 59*, 84–96.

Google Earth Engine (2023). Google Earth Engine Documentation. https://developers.google.com/earth-engine/guides/get started/ (accessed 28th November 2023).

Gulf of Mexico Coastal Ocean Observing System (GCOOS) (unknown) (2020) GCoos Mexico Quality Control Flags. Available at https://ntl.gcoos.org/data/GCOOS_QC_Flags.pdf (accessed 9th May 2025).

Hadjal M, **Medina-López E**, **Ren J**, **Gallego A and McKee D** (2022) An artificial neural network algorithm to retrieve chlorophyll a for northwest european shelf seas from top of atmosphere ocean colour reflectance. *Remote Sensing 14*(4), 3553.

Jin X and Han J (2008) *K-Means Clustering*. Encyclopedia of Machine Learning, Springer Nature. pp. 563–564.

Jochens A and Watson S (2013) The Gulf of Mexico coastal ocean observing system: An integrated approach to building an operational regional observing system. *Marine Technology Society Journal 47*, 118–133.

Kirillov A, **Mintun E**, **Ravi N**, **Mao H**, **Rolland C**, **Gustafson L**, **Xiao T**, **Whitehead S**, **Berg AC**, **Lo W-Y**, **Dollár P and Girshick R** (2023) Segment anything.

Lee T and Gentemann CL (2017) Satellite SST and sss observations and their role to constrain ocean models. in *new frontiers in operational oceanography*. https://doi.org/10.17125/gov2018.ch11.

Li G, **Cheng L**, **Zhu J**, et al. (2020) Increasing ocean stratification over the past half-century. *Nature Climate Change 10*, 1116–1123.

Li M, **Ren Y**, **Aw ZQ**, **Chen B**, **Yang Z**, **Lei Y**, **Cheng L**, **Liang Q**, **Hong J**, **Yang Y**, **Chen J**, **Wong YH**, **Wei J**, **Shan S**, **Zhang S**, **Ge J**, **Wang R**, **Dong JZ**, **Chen Y**, **Shi X**, **Zhang Q**, **Zhang Z**, **Chu JJH**, **Wang X and Zhang L** (2022) Broadly neutralizing and protective nanobodies against sars-cov-2 omicron subvariants bA.1, bA.2, and bA.4/5 and diverse sarbecoviruses. *Nature Communications 13*(1), 250.

Matsushita B, **Yang W**, **Chang P**, **Yang F and Fukushima T** (2012) A simple method for distinguishing global Case-1 and Case-2 waters using SeaWiFS measurements. *ISPRS Journal of Photogrammetry and Remote Sensing 69*, 74–87.

Mayo M (2022) Centroid initialization methods for k-means clustering. *KDnuggets*.

Medina-Lopez E (2019) High-resolution sea surface temperature and salinity in coastal areas worldwide from raw satellite data. *Remote Sensing 11*(19).

Medina-López E, **Navarro G**, **Santos-Echeandía J**, **Bernárdez P and Caballero I** (2023) Machine learning for detection of macroalgal blooms in the mar menor coastal lagoon using sentinel-2. *Remote Sensing 15*(5), 1208.

Melgani F and Pasolli E (2013) *Multiobjective PSO for Hyperspectral Image Clustering*. Springer, pp. 265–280.

Moore TS, **Dowell MD**, **Bradt S and Ruiz Verdu A** (2014) Optical water type classification based on ocean color radiometry: Application to the southern ocean. *Remote Sensing of Environment 152*, 336–349.

Muller-Karger FE, **Hestir E**, **Ade C**, **Turpie K**, **Roberts DA**, **Siegel D**, **Miller RJ**, **Humm D**, **Izenberg N**, **Keller M**, **Morgan F**, **Frouin R**, **Dekker AG**, **Gardner R**, **Goodman J**, **Schaeffer B**, **Franz BA**, **Pahlevan N**, **Mannino AG**, **Concha JA**, **Ackleson SG**, **Cavanaugh KC**, **Romanou A**, **Tzortziou M**, **Boss ES**, **Pavlick R**, **Freeman A**, **Rousseaux CS**, **Dunne J**, **Long MC**, **Klein E**, **McKinley GA**, **Goes J**, **Letelier R**, **Kavanaugh M**, **Roffer M**, **Bracher A**, **Arrigo KR**, **Dierssen H**, **Zhang X**, **Davis FW**, **Best B**, **Guralnick R**, **Moisan J**, **Sosik HM**, **Kudela R**, **Mouw CB**, **Barnard AH**, **Palacios S**, **Roesler C**, **Drakou EG**, **Appeltans W and Jetz W** (2018) Satellite sensor requirements for monitoring essential biodiversity variables of coastal ecosystems. *Ecological Applications 28*(3), 749–760.

Naeini AA, **Jamshidzadeh A**, **Saadatseresht M and Homayouni S** (2014) An efficient initialization method for k-means clustering of hyperspectral data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

O'Reilly JE and Werdell PJ (2019) Chlorophyll algorithms for ocean color sensors—OC4, OC5 & OC6. *Remote Sensing of Environment 229*, 32–47.

Salisbury J, **Vandemark D**, **Campbell J**, **Hunt C**, **Wisser D**, **Reul N and Chapron B** (2011) Spatial and temporal coherence between Amazon River discharge, salinity, and light absorption by colored organic carbon in western tropical Atlantic surface waters. *Journal of Geophysical Research: Oceans 116*(7), 1–14.

SeaDataNet. Available at https://www.seadatanet.org/ (accessed 9 May 2025).

Spyrakos E, **Jackson O and Hunter M**, **Claire M** (2017) Optical water typologies in European waters. *Remote Sensing of Environment 201*, 99–116.

https://doi.org/10.1017/eds.2025.10005 Published online by Cambridge University Press

Spyrakos E, **O'Donnell R**, **Hunter PD**, **Miller C**, **Scott M**, **Simis SG**, **Neil C**, **Barbosa CC**, **Binding CE**, **Bradt S**, **Bresciani M**, **Dall'Olmo G**, **Giardino C**, **Gitelson AA**, **Kutser T**, **Li L**, **Matsushita B**, **Martinez-Vicente V**, **Matthews MW**, **Ogashawara I**, **Ruiz-Verdú A**, **Schalles JF**, **Tebbs E**, **Zhang Y and Tyler AN** (2018) Optical types of inland and coastal waters. *Limnology and Oceanography 63*(2), 846–870.

Sun D, **Hu C**, **Qiu Z**, **Cannizzaro JP and Barnes BB** (2014) Influence of a red band-based water classification approach on chlorophyll algorithms for optically complex estuaries. *Remote Sensing of Environment 155*, 289–302.

Thakur KK, **Vanderstichel R**, **Barrell J**, **Stryhn H**, **Patanasatienkul T and Revie CW** (2018) Comparison of remotely-sensed sea surface temperature and salinity products with in situ measurements from British Columbia, Canada. *Frontiers in Marine Science 5*.

Umargono E, **Suseno J and Gunawan S** (2020) K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. https://doi.org/10.2991/assehr.k.201010.019.

Wang F, **Franco-Penya H-H**, **Kelleher J**, **Pugh J and Ross R** (2017) An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-62416-7_21.

Wei J, **Wang M**, **Ondrusek M**, **Gilerson A**, **Goes J**, **Hu C**, **Lee Z**, **Voss KJ**, **Ladner S**, **Lance VP and Tufillaro N** (2023) Chapter 20—Satellite ocean color validation. In Nalli NR (ed), *Field Measurements for Passive Environmental Remote Sensing*. Elsevier, pp. 351–374.

White S, **Lopez EM**, **Silva T**, **Spyrakos E**, **Martin A and Amoudry L** (2025) Exploring the link between spectra, inherent optical properties in the water column, and sea surface temperature and salinity. *Remote Sensing Applications: Society and Environment 37*, 101454.

White S, **Silva T**, **Amoudry LO**, **Spyrakos E**, **Martin A and Medina-Lopez E** (2024) The colours of the ocean: Using multispectral satellite imagery to estimate sea surface temperature and salinity on global coastal areas, the gulf of Mexico and the UK. *Frontiers in Environmental Science 12*, 1426547.

Yuan C and Yang H (2019) Research on K-value selection method of K-means clustering algorithm. *Journal of Engineering 2*(2), 226–235.

Cite this article: White S, Medina-Lopez E, Silva T, Spyrakos E, Amoudry L and Martin A (2025). Classification-informed estimation: the role of water-type clustering to improve neural network generalization for salinity and temperature estimation in coastal waters. *Environmental Data Science*, 4: e32. doi:10.1017/eds.2025.10005