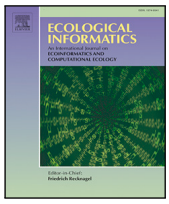
Ecological Informatics 91 (2025) 103390

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Unveiling 3D ocean biogeochemical provinces in the North Atlantic: A systematic comparison and validation of clustering methods

Yvonne Jenniges [a,b,*], Maike Sonnewald [c,d,e], Sebastian Maneth [b,**], Are Olsen [f], Boris P. Koch [a,g]

[a] Ecological Chemistry, Alfred-Wegener Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27570 Bremerhaven, Germany
[b] Faculty of Informatics, University of Bremen, 28539 Bremen, Germany
[c] Department of Computer Science, University of California, Davis, CA, USA
[d] School of Oceanography, University of Washington, Seattle, WA, USA
[e] NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[f] Geophysical Institute, University of Bergen, 5020 Bergen, Norway
[g] Department of Technology, University of Applied Sciences, 27568 Bremerhaven, Germany

ARTICLE INFO

Dataset link: Ocean regions dataset, Code, Dash board, Dashboard code

Keywords:
Ocean biogeochemistry
Ocean provinces
Machine learning
Clustering
Water masses
North Atlantic

ABSTRACT

Defining ocean regions and water masses helps to understand marine processes and can serve downstream tasks such as defining marine protected areas. However, such definitions often result from subjective decisions potentially producing misleading, unreproducible outcomes. Here, the aim was to objectively define regions of the North Atlantic through systematic comparison of clustering methods within the Native Emergent Manifold Interrogation (NEMI) framework (Sonnewald, 2023). About 300 million measured salinity, temperature, and oxygen, nitrate, phosphate and silicate concentration values served as input for various clustering methods (k-Means, agglomerative Ward, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)). Uniform Manifold Approximation and Projection (UMAP) emphasised (dis-)similarities in the data while reducing dimensionality. Based on systematic validation of clustering methods and their hyperparameters using internal, external and relative validation techniques, results showed that UMAP-DBSCAN best represented the data. Strikingly, internal validation metrics proved systematically unreliable for comparing clustering methods. To address stochastic variability, 100 UMAP-DBSCAN clustering runs were conducted and aggregated following NEMI, yielding a final set of 321 clusters. Reproducibility was evaluated via ensemble overlap (88.81±1.8%) and mean grid cell-wise uncertainty (15.49±20%). Case studies of the Mediterranean Sea, deep Atlantic waters and Labrador Sea showed strong agreement with common water mass definitions. This study revealed a more detailed regionalisation compared to previous concepts such as the Longhurst provinces through systematic clustering method comparison. The applied method is objective, efficient and reproducible and will support future research on biogeochemical differences and changes in oceanic regions.

1. Introduction

The definition of ocean regions has offered fundamental advances in our understanding of marine (eco)systems, biodiversity distributions and their variability. Since the 19th century, efforts to delineate biogeographic patterns have evolved from early taxonomic classifications (Forbes, 1856) to more comprehensive ecological and physical ocean regionalisations (Ekman, 1935; Hedgpeth, 1957; Briggs, 1974; Hayden et al., 1984; Briggs, 1995; Bailey, 1998). One of the most influential partitioning schemes, the ecological provinces by Longhurst (2007), has provided a foundational framework for investigating large-scale oceanographic patterns and processes, hotspots of biodiversity and ecological relationships. For example, the Longhurst regimes helped quantify primary production (Longhurst et al., 1995), characterise tuna movements (Logan et al., 2020) and trophic dynamics and food web structure (Arnoldi et al., 2023).

Ocean regionalisations have been used in studies addressing fields such as biogeographic realms (Costello et al., 2017), carbon flux (Gloege et al., 2017) or patterns of marine viruses (Brum et al., 2015). They also serve economic decisions in fisheries (Juan Jordá et al.,

* Corresponding author at: Ecological Chemistry, Alfred-Wegener Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27570 Bremerhaven, Germany.
** Corresponding author.
E-mail addresses: yvonne.jenniges@awi.de (Y. Jenniges), maneth@uni-bremen.de (S. Maneth).

https://doi.org/10.1016/j.ecoinf.2025.103390
Received 23 December 2024; Received in revised form 8 August 2025; Accepted 11 August 2025
Available online 1 September 2025

2022) and policy such as the designation of marine protected areas legislation (Spalding et al., 2007; Sonnewald et al., 2020; Zhao et al., 2020b; Reisinger et al., 2022).

While previous approaches often rely on predefined thresholds and subjective criteria, emerging data-driven techniques, particularly clustering algorithms, provide a more objective framework for ocean region delineation (Devred et al., 2007; Hardman-Mountford et al., 2008; Oliver and Irwin, 2008; Kavanaugh et al., 2014; Sayre et al., 2017; Sonnewald et al., 2019, 2020). The Native Emergent Manifold Interrogation (NEMI) method (Sonnewald, 2023) addresses key challenges in analysing complex geophysical data by integrating manifold learning, clustering, stochastic ensemble methods, uncertainty quantification and intuitive validation to ensure robust and interpretable clustering outcomes. This study builds upon and enhances the NEMI method through systematic comparison of clustering algorithms applied to biogeochemical data. Fundamentally, clustering results are influenced by algorithmic biases (Thrun, 2021) and, for some algorithms, the possibility of variable outcomes raising concerns about suitability, comparability and reproducibility. Two main validation strategies exist (Rui and Wunsch, 2005; Ullmann et al., 2022): (i) Internal validation, which assesses cohesion within clusters and separation between clusters using indices, such as the Calinski–Harabasz index (Caliński and Harabasz, 1974), Davies–Bouldin index (Davies and Bouldin, 1979) and Silhouette score (Rousseeuw, 1987) and (ii) external validation, which utilises knowledge not seen during model training, like comparing clustering results to established ecological classifications or visual analysis in different data spaces. For example, dimensionality reduction techniques like Uniform Manifold Approximation and Projection (UMAP, McInnes et al. (2018a)) improve the interpretability of clustering outcomes through enhancing associations between data structures and visualisation (Allaoui et al., 2020).

A key challenge for clustering is quantifying uncertainty. Some clustering and dimensionality reduction algorithms are non-deterministic and a globally optimal solution is not guaranteed. For example, UMAP uses stochasticity (McInnes et al., 2018a), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) results may change with permuted input data (Schubert et al., 2017) and k-Means is sensitive to cluster centroid initialisation (Fränti and Sieranoja, 2019). Quantifying clustering variability is crucial to assess reproducibility and reliability. One approach involves multiple runs and single-number similarity metrics, such as overlap (Manning et al., 2008). The NEMI method, which combines multiple clustering runs to form a final cluster set, has been proposed as a novel solution to represent statistical variability and quantify uncertainty.

Most prior studies on ocean partitioning focus on surface waters (e.g. Devred et al. (2007), Longhurst (2007), Hardman-Mountford et al. (2008), Oliver and Irwin (2008), Vichi et al. (2011), Reygondeau et al. (2013), Fay and Mckinley (2014), Kavanaugh et al. (2014), Reygondeau et al. (2020), Sonnewald et al. (2020)) despite the fact that critical processes such as particle export, upwelling or deep-water formation extend into deeper layers. Also, vertical mixing affects biodiversity leading to depth-variable community trends (DeLong et al., 2006; Hörstmann et al., 2022). Recent efforts to develop 3D ocean regionalisations have incorporated clustering approaches such as k-Means (Sayre et al., 2017) and hybrid methods combining k-Means, CMeans, agglomerative Ward and agglomerative full linkage (Reygondeau et al., 2017). In physical oceanography, the definition of 3D water masses, mainly defined by temperature and salinity but also e.g. oxygen, has always played a central role (e.g. Emery (2001), Tomczak and Godfrey (2003)). More recent definitions of water masses are based either on regional (Liu and Tanhua, 2021) or model data (Zika et al., 2021).

This study aims to develop an objective and reproducible 3D regionalisation of the North Atlantic ocean and its marginal seas using a data-driven clustering approach. Geographic coordinates and depth were excluded from the clustering to ensure that the regions are purely based on water properties. The analysis is based on a mostly post-industrialisation time-aggregate representing a long-term observational baseline, maximising spatial data coverage. A key novelty of this work is the systematic definition and evaluation of a marine clustering using both internal and external validation criteria, bridging statistical rigour with oceanographic knowledge. Furthermore, we increase statistical representativeness by combining multiple clustering runs within the NEMI framework that also allows for quantification of uncertainty. To contextualise the results, clustering outputs are compared to established definitions of ecological provinces and ocean regions (Longhurst, 2007; Sayre et al., 2017) in three distinct areas: the deep Atlantic, the Mediterranean and the Labrador Sea. This work specifically addressed the following research questions: (i) Which clustering method works best within the NEMI framework for the given oceanographic data? (ii) How reproducible are the clustering results? (iii) How do the results compare to existing definitions of ecological provinces and ocean regions?

The insights gained from this study have broad potential for ecological and environmental research. A thoroughly validated, data-driven ocean partitioning may refine or challenge existing frameworks, influencing our understanding of oceanic systems, climate dynamics, and marine resource management. By integrating ecological informatics methodologies, we contribute to the advancement of reproducible and objective ocean classifications, ultimately supporting more robust marine ecosystem assessments and policy applications.

The final gridded 3D set of clusters of the North Atlantic Ocean is publicly available (https://doi.org/10.5281/zenodo.15201767). It also contains auxiliary information, like the gridded oceanographic parameters. For interactive exploration of the clusters, a dashboard is available (https://ocean-cluster-dashboard.onrender.com, code: https://doi.org/10.5281/zenodo.16742244).

2. Material and methods

2.1. Native emergent manifold interrogation (NEMI) framework

This study implements the Native Emergent Manifold Interrogation (NEMI) framework, developed by Sonnewald (2023), as the methodological foundation for objective ocean regionalisation. NEMI integrates manifold learning, systematic clustering comparison, ensemble methods, and uncertainty quantification to ensure robust clustering outcomes. Specifically, we used NEMI to 1) aggregate results from multiple stochastic clustering runs (here UMAP-DBSCAN), reducing variability to yield robust final clusters, 2) quantify uncertainty by combining ensemble runs, where NEMI directly quantifies the uncertainty of cluster assignments, enhancing reproducibility and reliability, and 3) utilising a systematic framework for validation through structured approach to integrate external and relative validation strategies. The following sections detail our implementation of the NEMI protocol and systematic algorithm comparison in the context of the COMFORT dataset, described below.

2.2. Data

The dataset for this study (Korablev and Olsen, 2022) was assembled in the framework of the EU project "Our Common Future Ocean in the Earth System" (COMFORT) and combines ten observational datasets, including the World Ocean Database 2018 (WOD18) and Argo floats (for more information refer to Korablev et al. (2021)). It contains data on 47 parameters measured globally from the year 1772 to 2020 amounting to 458,724,734 values.

The focus area was the North Atlantic from −77 to 30° longitude, from 0 to 70° latitude and from 0 to 5,000 m depth. From the COMFORT dataset, the parameters temperature, salinity, as well as oxygen, nitrate, silicate and phosphate concentration were selected, which had comparatively good spatial coverage. After quality filtering,

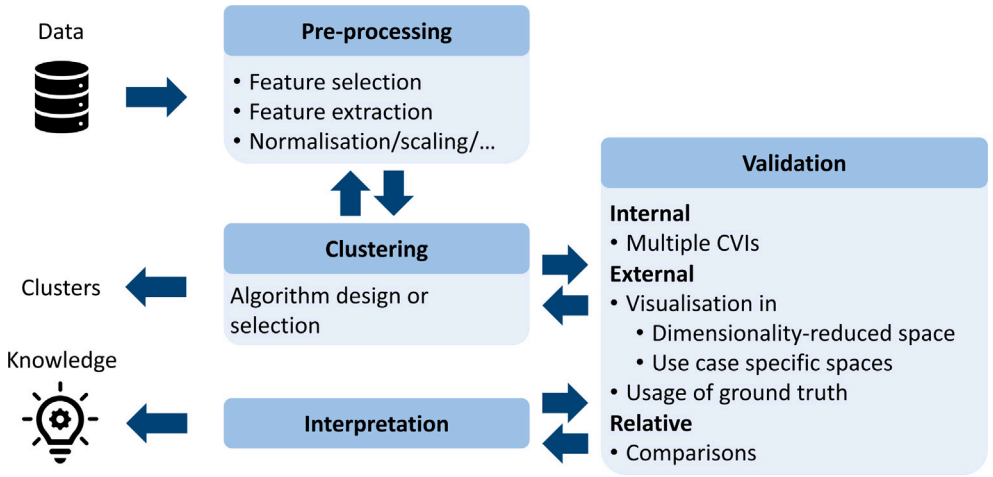


Fig. 1. The workflow was iterative and comprised (i) pre-processing, (ii) clustering, (iii) external, internal and relative validation that helped refine the selection of pre-processings, clustering methods and hyperparameter settings, and (iv) interpretation within the respective context leading to (new) knowledge.
Source: Clustering procedure adapted from Rui and Wunsch (2005).

unit conversions and averaging over all times (years 1772 – 2020), the data was mapped on a grid with a spatial resolution of 1° and 12 water depth intervals: 0 – 50 m, 50 – 100 m, 100 – 200 m, 200 – 300 m, 300 – 400 m, 400 – 500 m, 500 – 1,000 m, 1,000 – 1,500 m, 1,500 – 2,000 m, 2,000 – 3,000 m, 3,000 – 4,000 m, 4,000 – 5,000 m. Missing values were imputed using the K-Nearest Neighbours (KNN) algorithm (Python library scikit-learn, version 1.5; Pedregosa et al. (2011)). Details on data preparation and parameter distributions can be found in Supplementary Material A.

2.3. Clustering

Following a systematic clustering procedure (Fig. 1), three clustering methods were selected (Python library scikit-learn, version 1.5; Pedregosa et al. (2011)) for comparison: k-Means (the baseline), agglomerative Ward clustering (a hierarchical method) and DBSCAN (a density-based algorithm). All six parameters were scaled to a range from zero to one (MinMaxScaler, Python library scikit-learn, version 1.5) making the values comparable and usable with distance-variant methods like k-Means clustering.

To assess the influence of pre-processing, each algorithm was applied to the scaled data and to a dimensionality-reduced version of the scaled data. For the projection from six parameters (6D) to 3D, the non-linear method UMAP (Python library umap-learn, version 0.5.5; McInnes et al. (2018b)) was applied. The resulting 3D space will be called embedding in the following. This study focused on clustering methods that map a data point to exactly one cluster/label and assumed that the number of clusters must be $K > 1$. The terms cluster and label will be used interchangeably. For methodological details see Supplementary Material B.

To analyse the importance of each input parameter, cluster assignments were replicated with a more interpretable, predictive model by using the six scaled parameters as input and the cluster labels as output. The idea is similar to translating the clustering function into a neural network (Kauffmann et al., 2024), though without explicitly transferring the formulas. For cluster assignment replication, a random forest classifier (RandomForestClassifier, Python library scikit-learn, 1.4) was trained for each clustering model and its inherent feature importance was leveraged. The hyperparameters were configured as follows: The number of trees was set to 1,000, weights were balanced and the random state was fixed.

2.4. Validation

Validation can generally be categorised into internal and external approaches. Sometimes relative validation is listed as a third option referring to the comparison of different models (Rui and Wunsch, 2005). Internal validation exploits information available during the modelling process. In particular for clustering, Cluster Validity Indices (CVIs) or "scores" provide information on how cohesive a cluster is within itself and how separate it is to other clusters. Here, distance, density, and neighbourhood based CVIs were used. The applied distance based CVIs are the Calinski–Harabasz (CH, Caliński and Harabasz (1974)), Davies–Bouldin (DB, Davies and Bouldin (1979)) and Silhouette (SH, Rousseeuw (1987)) scores (Python library scikit-learn, version 1.5). The applied density and neighbourhood based CVIs are k-Density-Based Cluster Validation (k-DBCV, (Hammer et al.); adaption of Python library kdbcv, version 1.0.0; a more efficient implementation of DBCV, (Moulavi et al., 2014)), Clustering Validation Index based on Nearest Neighbours by (Halkidi et al., 2015) (CVNNH, Python library ascvi, version 0.1.0, (Schlake and Beecks, 2024); an adaption of CVNN, (Liu et al., 2013)) and Contiguous Density Region (CDR, (Rojas-Thomas and Santos, 2021); Python library ascvi, version 0.1.0). The CVIs were computed to determine hyperparameters and compare performance and are described in detail in Supplementary Material B.3. Desirable high within-cluster cohesiveness and between-cluster separation is indicated by low DB, CDR and CVNNH and by high SH and k-DBCV as well as by a local or global maximum of CH.

External validation is achieved by additional knowledge, such as ground truth labels or domain expertise. For biogeochemical and physical clustering, basic principles can be used to evaluate the cluster sets in their 3D geographic space as well as in their temperature-salinity (TS) space. Visual cluster examination can also be conducted in a dimensionality-reduced feature space to assess compliance with feature topology.

For dimensionality reduction methods like UMAP that enhance associations between data structures, several internal validation metrics have been proposed to e.g. assess embedding quality and guide hyperparameter tuning, including reconstruction error (Zhang et al., 2021). Trustworthiness and continuity evaluate how well local neighbourhoods are preserved during the projection (Venna and Kaski, 2006). Trustworthiness measures the extent to which neighbours in the embedding were also neighbours in the original space, while continuity measures whether neighbours in the original space remain close in the embedding. Qlocal and Qglobal provide an alternative approach to quantifying neighbourhood preservation at different scales (Lee and

Verleysen, 2010). Unlike trustworthiness and continuity, which are based on binary neighbour relationships, these metrics consider relative ranking of all pairwise distances. Trustworthiness, continuity, Qglobal and Qlocal range between zero and one, with higher values indicating better structure preservation. They were computed using the Python library pydrmetrics, version 0.0.7 (Zhang et al., 2021) on a random subset of 2,000 data points for computational efficiency.

2.5. Uncertainty assessment

Uncertainty was systematically examined in four parts of the analysis. First, 100 clustering runs were performed and combined into a final cluster set following the NEMI framework since the selected clustering pipeline (UMAP-DBSCAN) contains non-deterministic components. Variability of the complete process was assessed by the Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), cluster overlap (see Section 2.6) and by NEMIs inherent grid-cell-wise uncertainty. Second, this work addressed uncertainty of the dimensionality reduction alone by computing the UMAP embedding 100 times and by comparing the similarity between runs using the root mean-squared error (RMSE). Third, for hyperparameter tuning, uncertainty was taken into account by computing each hyperparameter combination ten times for each experiment (see Section 3.2). Fourth, to quantify uncertainty of DBSCAN, it was applied 100 times to a fixed embedding.

Additional sources of uncertainty of the presented approach were data quality and coverage, as well as missing value imputation. Data was filtered using existing quality flags, which is why data quality was not further investigated in this study. It should be noted though that the COMFORT dataset comprises data measured at different times and with different instruments decreasing accuracy and precision. Another potential source of uncertainty was the imputation of non-existent measurement values in the defined geospatial grid, which was considered by flagging each imputed data point.

Following the NEMI framework, an ensemble of clustering runs was performed and aggregated, from which NEMI infers uncertainty to assess variability. The algorithm requires manual selection of a base cluster set (*base_id*), against which all other cluster sets are compared. The process begins by matching labels across ensemble members. For this, each cluster set is reassigned new labels sorted by cluster size (with label zero corresponding to the largest cluster). For every member cluster set (except the base), the algorithm visits every possible pair of a base label a_i and a member label b_j and determines the overlap as

$$\text{NEMI\_overlap}(a_i, b_j) = \frac{\text{volume}(a_i \cap b_j)}{\text{volume}(a_i \cup b_j)} \tag{1}$$

i.e., the volume of geographically shared data points divided by the joint volume of data points. The original NEMI implementation works with cell count instead of volume.

The ensemble approach also allows uncertainty quantification per grid cell, which was computed here as the number of times the cell was assigned to different clusters:

$$\text{NEMI\_uncertainty}(\text{grid\_cell}) = 100 - \frac{|\text{same\_cluster\_assignment}|}{|\text{ensemble}|} \tag{2}$$

2.6. Cluster similarity metrics

When clusters are generated, comparisons to other sets of clusters or regionalisations, i.e. relative validation, is an important step for putting results into context. Various metrics are available to assess similarity of cluster sets (cf. e.g. Vinh et al. (2010)) such as overlap (Manning et al., 2008), Adjusted Rand Index (ARI, Hubert and Arabie (1985)) and Normalised Mutual Information (NMI, Strehl and Ghosh (2002)).

2.6.1. Adjusted Rand index

The Adjusted Rand Index (ARI, Python library scikit-learn, version 1.5, Hubert and Arabie (1985)) is a symmetric measure to assess similarity between two cluster sets. It is based on the Rand Index (RI), which measures the proportion of agreeing pairs of points between two cluster sets

$$\text{RI} = \frac{a+b}{\binom{n}{2}}, \tag{3}$$

where a is the number of pairs of points that are in the same cluster in both partitions, b is the number of pairs that are in different clusters in both partitions, and n is the total number of points. The ARI adjusts this measure for chance agreement

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]} \tag{4}$$

where $\mathbb{E}[\text{RI}]$ denotes the expected RI of random labellings. The index is bound between −0.5 and one, where identical cluster sets receive a score of one, a random label matching a score of zero and complete disagreement a negative score.

2.6.2. Normalised mutual information

Normalised Mutual Information (NMI, Strehl and Ghosh (2002)) quantifies the amount of shared information between two cluster sets. In this study, the normalisation proposed by Kvalseth (1987) was applied yielding

$$\text{NMI} = \frac{2I(A,B)}{H(A)+H(B)} = \frac{2\big(H(A)+H(B)-H(A,B)\big)}{H(A)+H(B)} \tag{5}$$

where $H(A)$ and $H(B)$ denote the entropies of the two cluster sets A and B, $H(A,B)$ their joint entropy, and $I(A,B)$ the mutual information.

NMI ranges from zero to one, where one indicates perfect agreement (identical cluster sets) and zero denotes no shared information.

2.6.3. Cluster overlap

An asymmetric measure to compare two cluster sets A and B and hence quantify reproducibility is overlap or purity (Manning et al., 2008), which ranges between 0 and 1. It assesses how much the clusters of two cluster sets overlap by calculating the maximum overlap for each cluster/label in A with labels in B and vice versa. The average overlap is used as a cluster set similarity measure.

Let there be N objects that are grouped by clustering A into clusters $a_0, \ldots, a_I$ and by clustering B into clusters $b_0, \ldots, b_J$. The overlap of A with B is then defined as

$$\text{overlap}(A,B) = \frac{1}{N}\sum_i \max_j |a_i \cap b_j| \tag{6}$$

To obtain a symmetric measure, we computed the overlap as

$$\text{overlap} = \frac{\text{overlap}(A,B) + \text{overlap}(B,A)}{2} \tag{7}$$

3. Results

3.1. Embedding using UMAP

As an initial validation step, the influence of UMAP hyperparameters was tested. The parameters control if global or local associations within the data are emphasised by minimising the cross-entropy (Supplementary Material B.1). A lower number of neighbours led to a more curled structure, while a higher number unfolded it (Fig. 2, a-c). Increasing the number of neighbours to more than 20 did not further change the embedding (Fig. 2, c). As expected, a higher minimum distance caused the data points to spread apart further, blurring finer topologies (Fig. 2, e-f). For the final cluster runs, the number of neighbours was set to 20 and the minimum distance to 0 (Fig. 2, b and d) so that the embedding space, with three dimensions, exhibited a distinct topology. Over 20 iterations, the final embedding achieved a mean

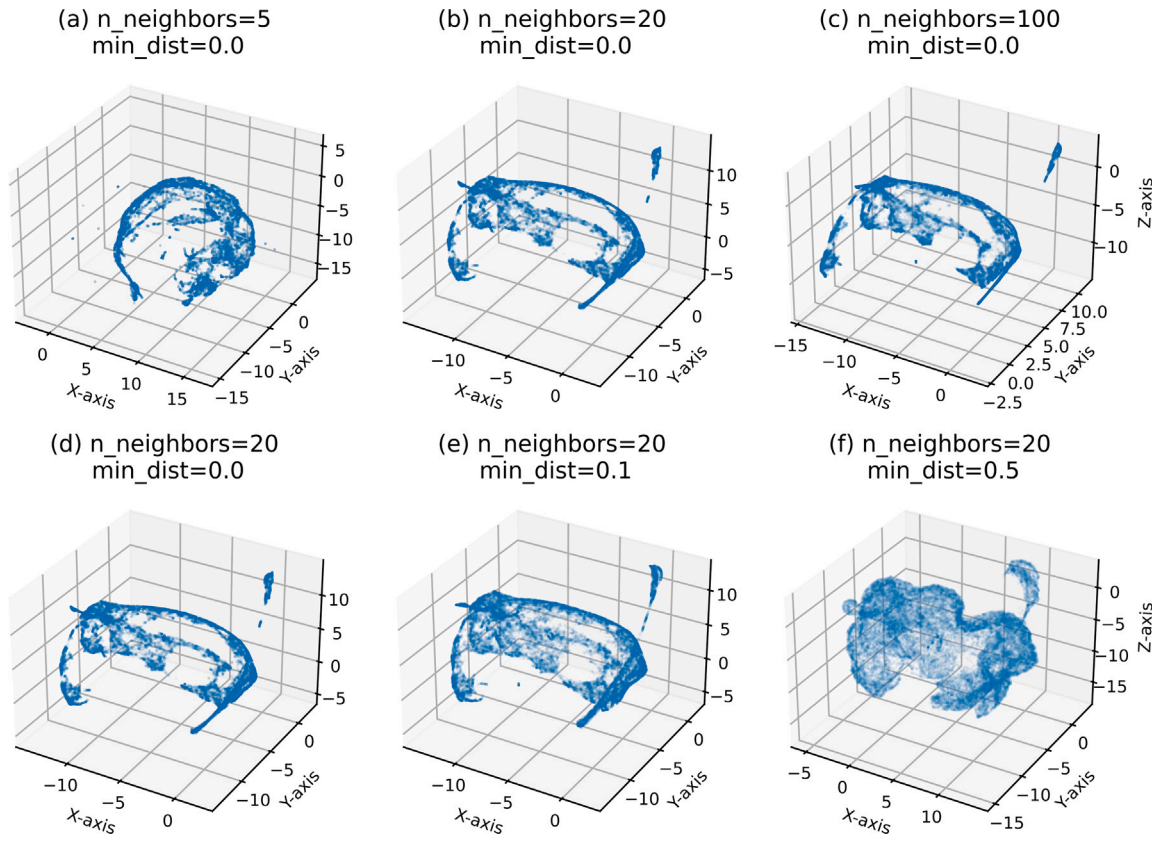


Fig. 2. Influence of the UMAP hyperparameters number of neighbours (*n_neighbors*, top) and minimum distance (*min_dist*, bottom) on the embedded space. With an increasing number of neighbours, the topology unfolded but did not substantially change for more than 20 neighbours. With a growing minimum distance, the data structure lost structural detail.

Table 1
Overview of clustering results. For external validation, the clustering was visually analysed in geographical and embedding spaces (last two columns, cf. Section 3.2). Note that the Cluster Validity Indexes (CVIs), i.e. Calinski–Harabasz Score (CH), Davies–Bouldin Score (DB), Silhouette Score (SH), k-Density Based Clustering Validation (k-DBCV), Clustering Validation Index Based on Nearest Neighbours Halkidi (CVNNH) and Contiguous Density Region (CDR), are not suitable to compare across clustering methods (see Section 4.3). High CH, SH and k-DBCV as well as low DB, CVNNH and CDR indicate a well separated clustering.

Data	Algorithm	Hyperparameters	Internal validation						External validation	
			CH ↑	DB ↓	SH ↑	k-DBCV ↑	CVNNH ↓	CDR ↓	Geo	UMAP
Original	K-Means	$n_{clusters} = 2$	57,358	0.78	0.50	−1.00	0.24	0.57	Warm-cold separation	Straight cut
Embedded	K-Means	$n_{clusters} = 8$	49,703	0.71	0.47	−0.78	3.55	0.60	Non-separation of deep areas	Non-separations
Original	Agg. Ward	$n_{clusters} = 2$	51,825	0.84	0.48	−1.00	0.24	0.56	Hot-cold like	Intrusions
Embedded	Agg. Ward	$n_{clusters} = 25$	65,901	0.72	0.46	−0.61	1.72	0.59	Geo. connected	Intrusions, non-separations
Original	DBSCAN	$\epsilon = 0.11949153$ $min_{samples} = 11$	800	1.44	0.55	−1.00	0.98	0.54	Focus on Baltic	One big cluster
Embedded	DBSCAN	$\epsilon = 0.10661017$ $min_{samples} = 4$	1,754	1.5	−0.35	−0.41	3.31	0.56	Density-oriented	Many small clusters

trustworthiness of 0.97 ± 0.004 and a mean continuity of 0.97 ± 0.007. Global neighbourhood was better preserved than local ones ($Qglobal = 0.87 \pm 0.009$, $Qlocal = 0.66 \pm 0.01$, Fig. S4). A Shepard plot of a UMAP embedding is illustrated in Supplementary Material Fig. S4.

3.2. Clustering results

Six experiments were carried out during which three clustering methods were applied to original and embedded data. Each experiment was run ten times for every hyperparameter combination. A striking result was that little agreement between the CVIs was found (Table 1).

3.2.1. K-Means

Clustering original data. For k-Means applied to the original 6D data, all three classical distance-based scores - CH, DB and SH - reached peak performance when setting the number of clusters to two (Fig. S5). Plotting the result of k-Means with two clusters in geographic space, two coherent regions emerged (Fig. 3). They most strongly resembled the input temperature distribution, supported by temperature having the highest feature importance (47%). Structural validity scores, in contrast, showed different preferences: CVNNH reached its optimum at 12 clusters, while CDR and k-DBCV favoured higher values, continuing to improve up to 60 clusters, though k-DBCV remained negative

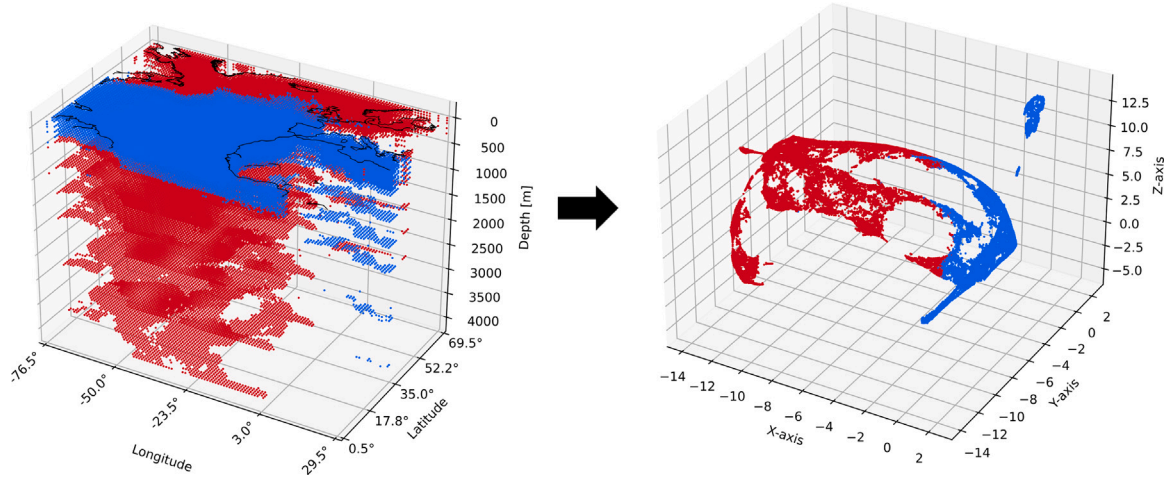


Fig. 3. K-Means applied to the original data with two clusters in geographic space (left) and its projection into the embedded space (right). The geographic separation of the clusters resembled the temperature distribution. In embedding space, the division approximated a straight cut.

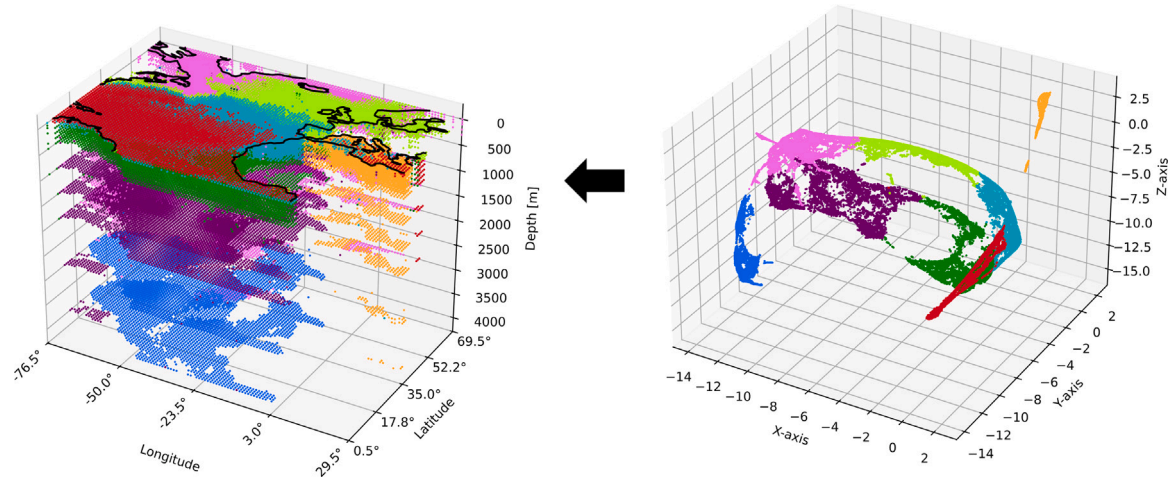


Fig. 4. K-Means applied to the embedded data with eight clusters (right) and its projection to geographic space (left). Insufficient performance is reflected in clusters that are not separated despite their clear segregation in embedding space, e.g. the blue cluster.

throughout. Note that clustering here is not performed on the embedding. The embedding is merely used to visualise how the clusters look like in this space. The visualisation in embedding space showed clear errors: Structures that are separate were not subdivided, other clusters intruded into foreign clusters.

Higher numbers of clusters resulted in more globular clusters in embedding space (Fig. S6). Across most clustering levels, temperature and oxygen emerge as the most important features, while the lower-ranked feature of phosphate slightly gained in importance for larger cluster numbers.

Clustering embedded data. For k-Means applied to the data in the previously computed embedded space (Section 3.1), internal validation (the scores) suggested numbers of clusters greater than two (Fig. S5). SH and DB both favoured eight clusters, while the other scores rose/fell beyond the selected value range with CH peaking locally at 14 clusters. Despite some improvement over clustering on original data, the resulting clusters still showed non-separated regions in embedding space (Fig. 4).

To compensate for potential inaccuracies in internal validation, cluster sets across different numbers of clusters were further evaluated externally. A higher number of clusters resulted in the embedded space adapting a chessboard pattern (Fig. S7), that was clearly not representative of the underlying co-variance space given by the data.

3.2.2. *Agglomerative Ward*

Clustering original data. Similar to k-Means on the original data, all three classical CVIs suggested two clusters as the best split using agglomerative Ward clustering (Fig. S8). The visualisation in the embedding revealed a straight cut (Fig. 5) and temperature had the highest importance (35%) followed by oxygen (34%). Structural scores (k-DBCV, CVNNH and CDR) indicated more fine-grained cluster sets, with optimal values of 27, 12 and 30 clusters, respectively. Increasing the number of clusters up to 30, regardless of the scores, resulted in unseparated clusters and intrusion issues (visible in the embedded space), though less than using k-Means. Temperature contributed most to separation followed by oxygen and salinity.

Clustering embedded data. The scores for agglomerative Ward clustering applied to the embedded data preferred a number of clusters larger than two (Fig. S8). While CH and CVNNH did not show an extreme value that could be used to optimise the number of clusters, SH suggested 24 clusters as the optimum. DB and k-DBCV preferred 24 clusters and CDR was minimal for 23 and 28 clusters. Selecting 25 reasonably divided the embedded data at areas of low data point density, except for the Labrador Sea and some smaller clusters (Fig. 6). Deeper ocean regions (two bulbous structures on the left side of embedding space) and the Mediterranean water masses were well-separated while clustering of original data was not able to find these differences.

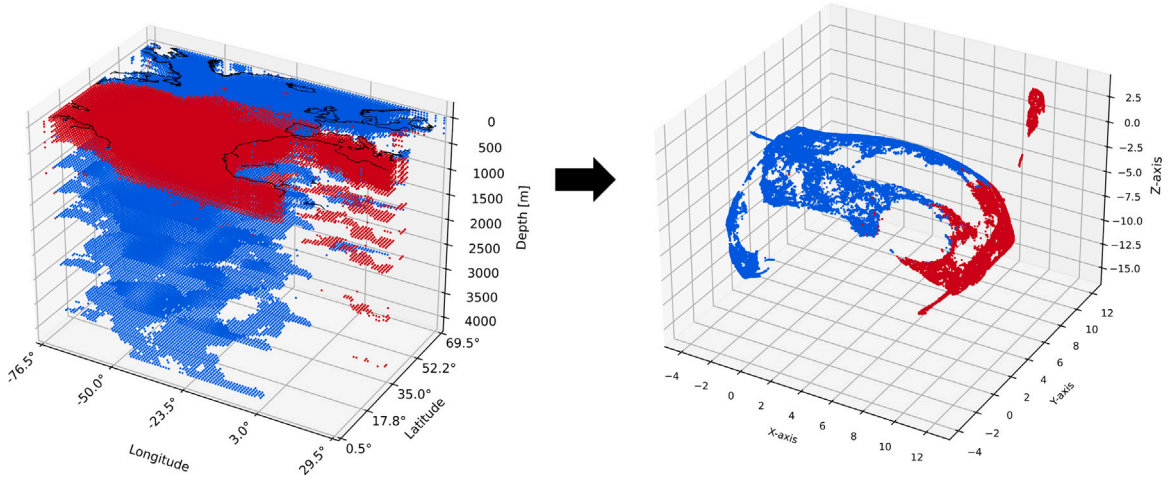


Fig. 5. Agglomerative Ward clustering applied to the original data with two clusters (left) and its projection into the embedded space (right).

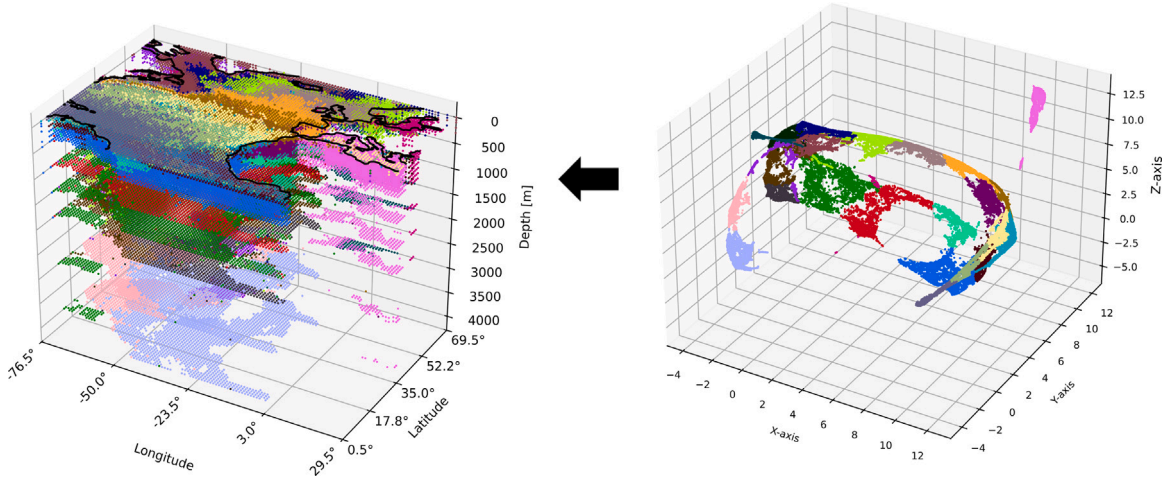


Fig. 6. Agglomerative Ward applied to embedded data with 25 clusters (right) and its projection to geographic space (left).

3.2.3. DBSCAN

In contrast to k-Means and agglomerative Ward clustering, DBSCAN requires tuning of two main hyperparameters: *epsilon*, i.e., the search radius, and *min_samples*, i.e., the minimum cluster size. DBSCAN labels data points that it could not assign to a cluster as noise. Therefore, the amount of noise was taken into account for the final choice of hyperparameter settings. For hyperparameter tuning of *min_points*, values between 2 and 11 and for *epsilon* values between 0.01 and 0.2 (20 steps with step size 0.00322) were tested for original and embedded data.

Clustering original data. All tested hyperparameter combinations for DBSCAN applied to the original 6D data resulted in similar clusterings that heavily focused on the Baltic, Black and/or Mediterranean Sea (Fig. 7) independent of the tuning criterion (Table ST1). For low *epsilon*, some surface clusters were identified along with 40–60% DBSCAN noise. The number of small clusters decreased with increasing *epsilon* while higher *min_samples* led to less small clusters and more coherent, but few (mostly below 10) regions.

CH favoured a high *min_samples* producing a low-noise clustering with coherent regions in the Baltic Sea. DB preferred fewer *min_samples* and generated a clustering with low noise but spatially less coherent clusters. SH and CVNNH suggested high values for *epsilon*, SH also requiring high *min_samples*. K-DBCV quantified all points as noise in 99% of the tested combinations. CDR suggested low *epsilon* producing cluster sets with around 10% noise.

Clustering embedded data. The distribution of the number of clusters and the noise fraction was similar for DBSCAN applied to the embedded data compared to original data. However, the number of clusters rose above 6,500 and the noise reached proportions of 99%. Low *epsilon* and high *min_samples* generally resulted in highest noise. Low *epsilon* and low *min_samples* led to a large number of small clusters, while an *epsilon* value above 0.1 generated larger, coherent structures.

The CH (Fig. S9) formed a clear ridge that rose linearly with increasing *min_samples*. It separated very coarse cluster sets (above the ridge, i.e., larger *epsilon* values) from cluster sets consisting of smaller regions (below the ridge, i.e., smaller *epsilon*). The score hence delineated a compromise between coarse and fine cluster sets. Its optimal/maximal value (at $epsilon = 0.15491525$, $min\_samples = 11$) produced a cluster set that incorporated clearly delineated regions in geographic and embedded space. In the embedded space, some larger structures were not subdivided despite only thin connections, such as a geographic region in the north. The SH resembled the CH only favouring lower *epsilon* and *min_samples* resulting in more small clusters. In contrast, the DB preferred hyperparameter combinations producing much noise. The structural CVIs (k-DBCV, CVNNH and CDR) generally favoured configurations that produced many small clusters with small differences in the few larger clusters and noise proportion.

Final hyperparameters were selected based on CH as well as external validation, i.e., the visual inspection of embedded and geographic space. CH was used as orientation since it reflected the trade-off between small clusters versus larger coherent regions. The other scores were ignored due to their lack of agreement with visual clustering

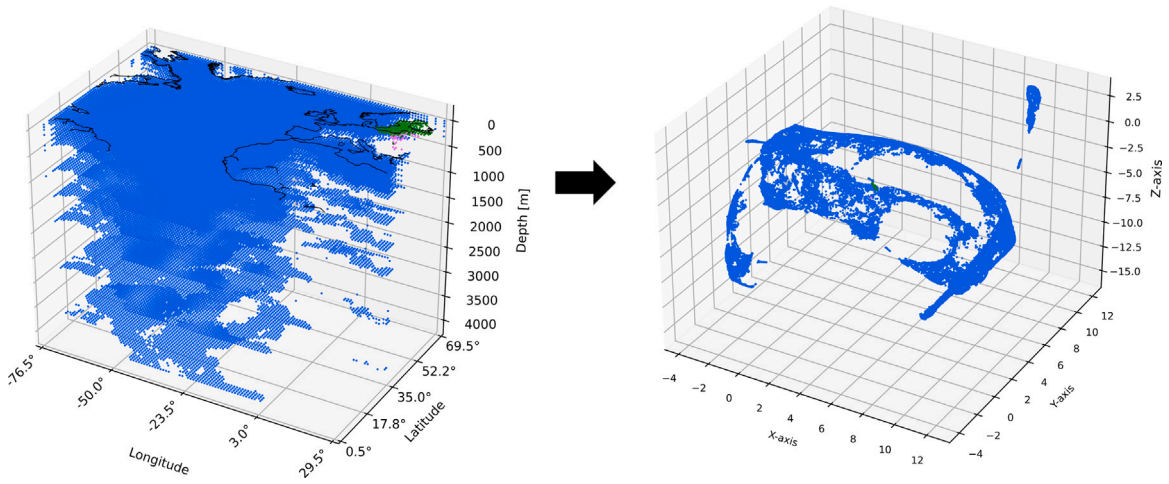


Fig. 7. DBSCAN applied to original data (left) and its projection to embedding space (right). The hyperparameters were selected through the Calinski–Harabasz Score (CH). The clustering was heavily focused on the Baltic.

quality, i.e., they favoured cluster sets with more noise and/or many small clusters. The visual inspection led to the decision of $epsilon = 0.10661017$ and $min\_samples = 4$ (Fig. S9, red square). *Epsilon* smaller than the selected value resulted in more small clusters while for larger *epsilon*, clusters tended to merge. A *min_samples* smaller than four at the selected *epsilon* formed a larger cluster in the Labrador Sea and more small clusters. The clustering for a minimum of five samples was similar, but had more noise (5.17% versus 3.82%). Increasing the *min_samples* up to 11 resulted in regions of smaller size.

3.3. UMAP-DBSCAN: Uncertainty and post-processing

Based on the above analysis, DBSCAN applied to the data embedded by UMAP best represented the given data structure. Due to stochasticity of UMAP-DBSCAN, each of the 100 runs produced slightly different results. To assess reproducibility, three sources of uncertainty were examined:

First, UMAP was applied 100 times with the same $n\_neighbors$ and min_dist to the scaled input data. It was not possible to directly compute the Root Mean Squared Error (RMSE) between any two embeddings since they were not always spatially aligned. Therefore, the point clouds of the embeddings were aligned through manual translation, rotation and/or mirroring followed by the application of the Iterative Closest Point (ICP, Python library simpleicp, version 2.0.14) algorithm to maximise alignment. ICP calculates a translation and rotation matrix for one point cloud trying to minimise distances to another point cloud. The RMSE was computed for the embeddings before and after applying ICP. The minimal deviation, i.e. the error when two embeddings aligned best, was stored, resulting in an average RMSE of 0.22 ± 0.06 (or about $1.3 \pm 0.36\%$ of the value range).

Second, variability of DBSCAN was assessed by applying it 100 times to a fixed, pre-computed embedding while keeping hyperparameters fixed. For each run, the rows of the training data were shuffled as DBSCAN is sensitive to the sequence of input data. The mean overlap between the resulting cluster sets was $99.99 \pm 0.003\%$.

Third, variability of the combined UMAP-DBSCAN pipeline was assessed by running it 100 times, i.e., in each iteration both, the embedding and the clustering, were recomputed. On average, the ARI between the resulting cluster sets was 0.78 ± 0.05, the NMI was 0.91 ± 0.01 and the overlap was $88.81 \pm 1.8\%$ (Fig. S10).

Final cluster assignments. To obtain a final cluster set, following the NEMI framework, the individual members of the 100 UMAP-DBSCAN runs were combined. As *base_id*, the cluster set with lowest mean uncertainty was chosen. The final clustering (Fig. 8) had 321 clusters and 3.92% ($n = 1920$) of the grid cells were assigned as noise and were subsequently excluded.

Table 2
Similarity of this study's final cluster set with marine provinces (Longhurst, 2007) and Ecological Marine Units (EMUs, Sayre et al. (2017)).

Similarity score	Longhurst provinces	EMUs
Overlap	0.56	0.62
Normalised Mutual Information (NMI)	0.43	0.51
Adjusted Rand Index (ARI)	0.16	0.18

The average uncertainty (Fig. 9) was $15.49 \pm 20\%$ (min: 0%, max: 96%) and 50% of the uncertainties were $\leq 5\%$. Lowest mean uncertainties (0%) were found in 6 clusters, three of which were geographically scattered but well separated in embedding space (labels 209, 244, 288). The three certain, geographically cohesive clusters were located in the Black Sea (100 – 2,000 m; label 93), in the southern Baltic Sea (100 – 400 m; label 102) and in the Gulf of Saint Lawrence (200 – 1,000 m; label 112). Uncertainties $\geq 50\%$ were geographically scattered with the exception of a cluster stretching from the Strait of Gibraltar until −64.5° West (0 – 300 m; label 28) and a cluster off the coast of North West Africa (200 – 1,500 m; label 53).

3.4. Case studies

Three ocean regions from the final clustering, which were determined based on the combination of 100 UMAP-DBSCAN runs using the NEMI framework (Fig. 8), were inspected. Overall, the final cluster set shows higher similarity to the EMUs than to Longhurst provinces (Table 2). In both cases, overlap and NMI are moderate, while ARI values are relatively low.

3.4.1. Mediterranean Sea

In the presented regionalisation, the Mediterranean Sea, delineated by longitude $\in [-6, 30]$ and latitude $\in [30, 48]$ omitting the Black Sea and Bay of Biscay, had a total of 2,509 grid cells (Fig. 10). The region was subdivided into one large cluster occupying 83% of all cells (label 6), the second largest cluster covering 7% of the grid cells (label 35) and 25 smaller clusters (< 80 cells) that were mainly found at the surface. Only one larger clusters extended downwards until 100 m (label 35, eastern basin).

The main Mediterranean clusters (labels 6, 32, 35, 138) were partly related to the North Atlantic representing the Mediterranean in- and outflow (labels 0, 10, 19, 28) at the Strait of Gibraltar. In TS space, the division between the Mediterranean Sea and North Atlantic waters

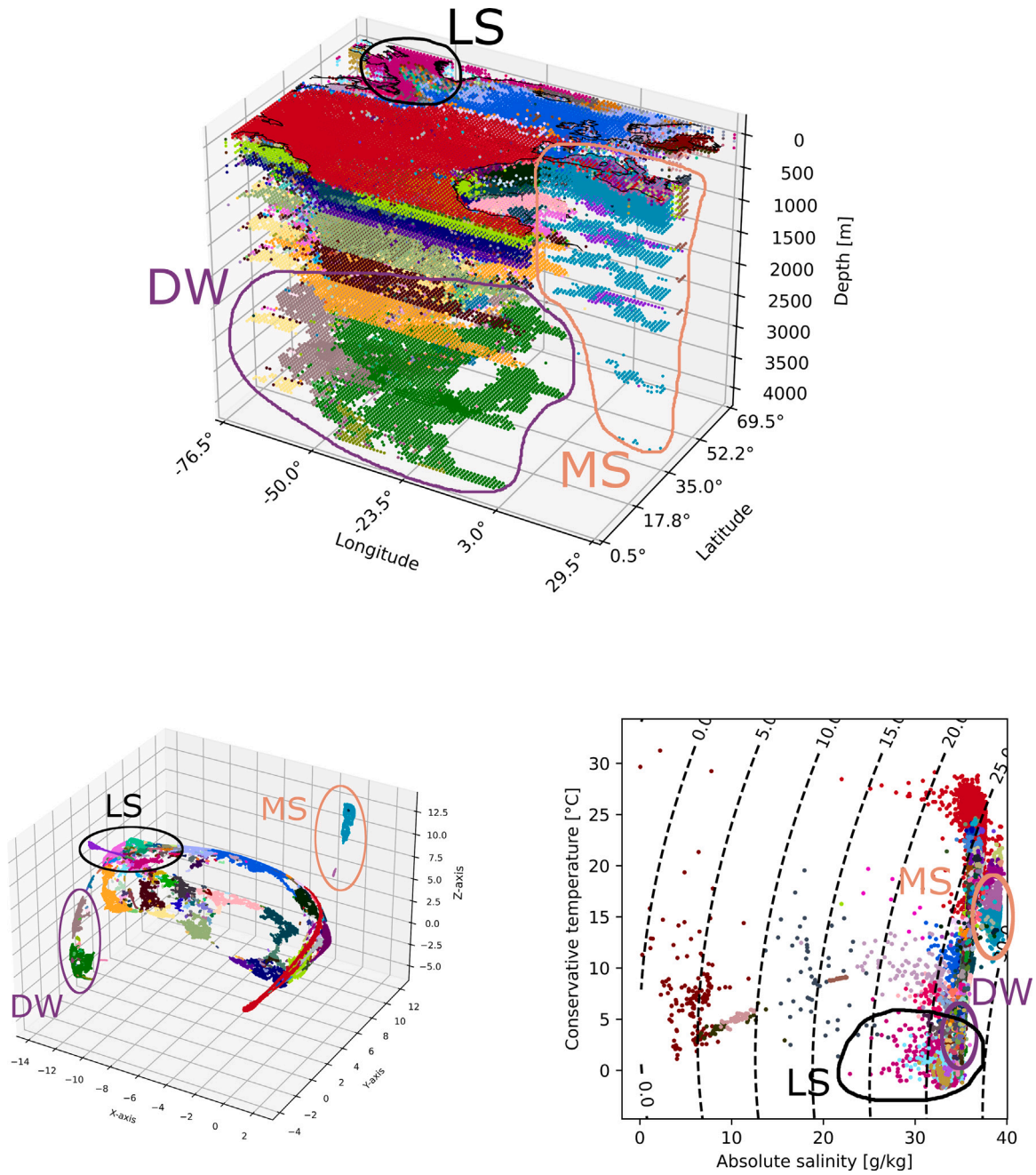


Fig. 8. Final UMAP-DBSCAN clustering melted from 100 runs using Native Emergent Manifold Interrogation (NEMI) in geographic (top), embedded (bottom left) and temperature-salinity (TS) space (bottom right). The three use cases (discussed in Section 3.4) are highlighted: Deep Atlantic waters (DW), Labrador Sea (LS) and Mediterranean Sea (MS).

was clearly visible around the 27-isopycnal. In embedding space, it was even more pronounced by the large distance between the main Mediterranean cluster (label 6) and the remaining data structure. The clusters representing the Mediterranean outflow were sorted by increasing depth in embedding space. Highest uncertainties were found in this outflow area (Fig. S11), while the remaining clusters of the Mediterranean Sea had low uncertainty.

3.4.2. Deep Atlantic waters

The deep waters of the North Atlantic extended from 3,000 m to 5,000 m stretching over the complete range of considered latitudes and longitudes excluding the Mediterranean Sea. The presented clustering detected 43 clusters (Fig. 11), two of which accounted for the assignment of 79% of the grid cells in that area (labels 2 and 11). 38 clusters had less than 100 cells. Despite having been generally separated by the North Atlantic Ridge, the eastern water mass was also present at the west side of the ridge between 10 and 30°. In TS space, these water masses were not distinguishable. For further analysis, six clusters (labels 2, 11, 31, 42, 51, 52) were chosen that had uncertainties below 10%, except one in the south-west (27%, label 52, Fig. S12).

The east–west separation (label 2 and 11 in Fig. 11) was not clear in TS space but very pronounced in embedding space (which also considered oxygen and nutrients). Taking a closer look at parameter distributions (Fig. 12) revealed that the western region has notably less silicate and more oxygen as well as lower nitrate and phosphate values.

3.4.3. Labrador Sea and Davis Strait

The Labrador Sea is located between the Labrador Peninsula and Greenland. The Davis Strait between 60 and 70°N connects it with Baffin Bay forming a shallower water pathway. 4,170 grid cells were located

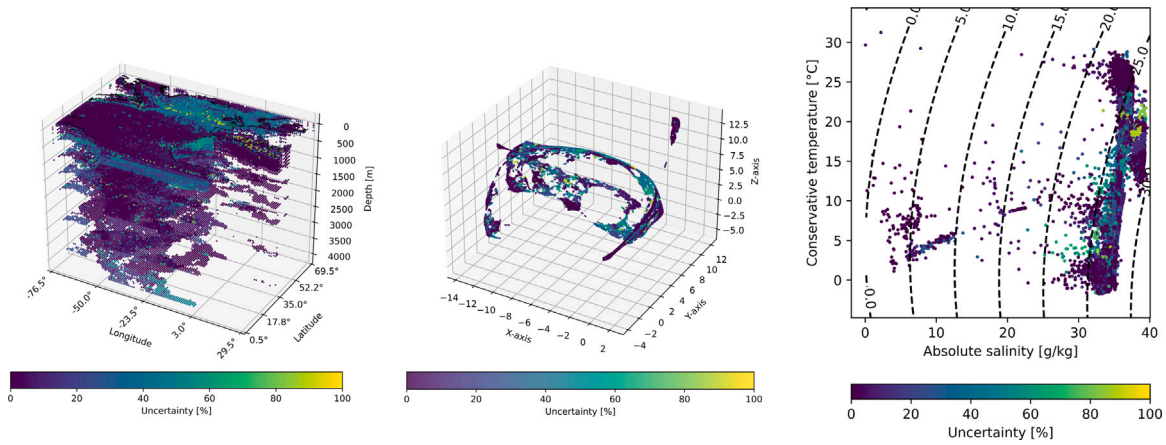


Fig. 9. NEMI uncertainty of a UMAP-DBSCAN clustering melted from 100 runs using NEMI in geographic (left), embedded (middle) and temperature-salinity (TS) space (right). The mean uncertainty was $15.49 \pm 20\%$ with a median of 5%. The high standard deviation and low median indicate a few highly uncertain grid cells skewing the mean.

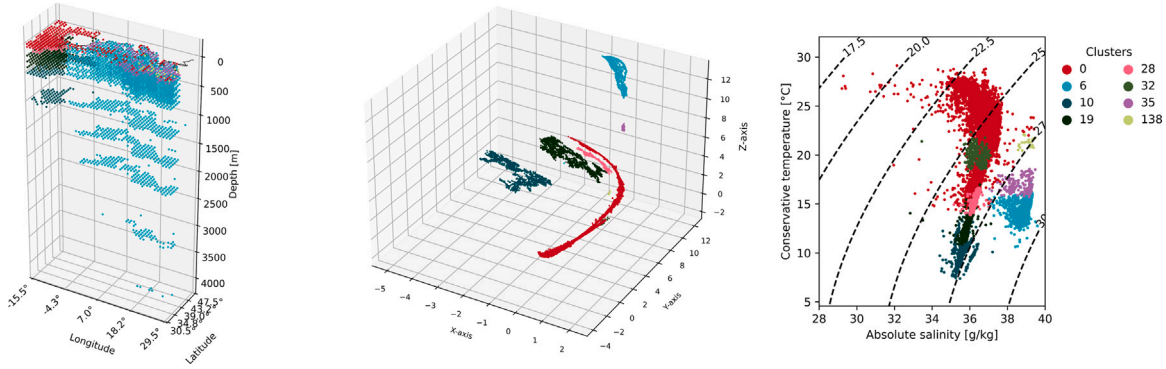


Fig. 10. Selected Mediterranean Sea clusters (labels 6, 32, 35, 138) and related in-/outflow regions in the North Atlantic (labels 0, 10, 19, 28) in geographic (left, zoomed into the geographic area of the Mediterranean Sea), embedded (middle) and temperature-salinity (TS) space (right).

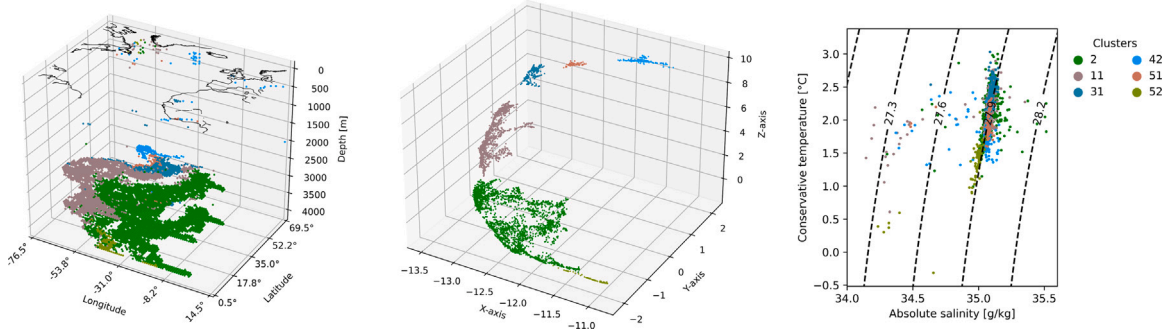


Fig. 11. Selected clusters in the deep North Atlantic (labels 2, 11, 31, 42, 51, 52) in geographic (left), embedded (middle) and temperature-salinity (TS) space (right). The clusters were indistinguishable in TS space (except label 50 in the south), but well separable in embedding space. In geographic space, a division along the Mid-Atlantic ridge was visible.

between 45 and 70°N and 40 and 77°W, which were assigned to 124 different clusters. The six largest clusters (labels 3, 17, 18, 24, 30, 42, Fig. 13) were analysed, each comprising more than 100 grid cells and the largest having 780 cells (label 17).

The clusters exhibited a strong vertical structure. One cluster (label 17) formed a wide stretch around the surface coasts vanishing at about 300 m, another one appeared centrally in the Labrador Sea at 50 m depth stretching down to 1,000 m (label 18). It borders the largest cluster (label 3), which reaches until 3,000 m where it is replaced until the bottom by another cluster (label 42). Cluster label 30 filled Baffin Bay from 100 to 400 m and was separated from the central cluster (label 18) by an intermediate region (label 24). The latter reached form the surface until 3,000 m.

The cluster assignment uncertainty (Fig. S13) in this region was on average $10.73 \pm 15.4\%$, the median was 4% indicating that uncertainty was inclined towards the lower end. Uncertainties mainly increased along the edges of the embedded data structure, which corresponded

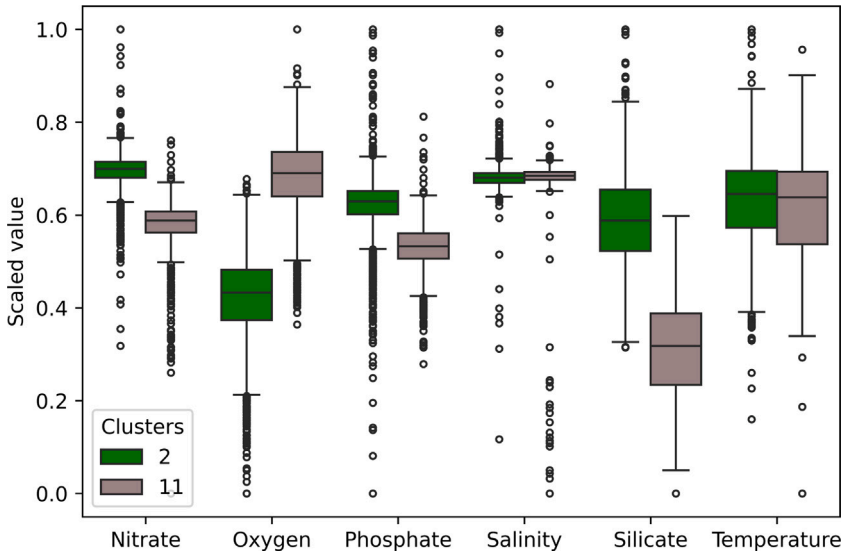


Fig. 12. Comparison of parameter statistics of the eastern (label 2) and western deep Atlantic (label 11). Note that the western region had lower nutrient and higher oxygen concentration. The per-parameter differences were statistically significant (p_value = 0, Mann–Whitney U test, Python library scipy, version 1.11.4).

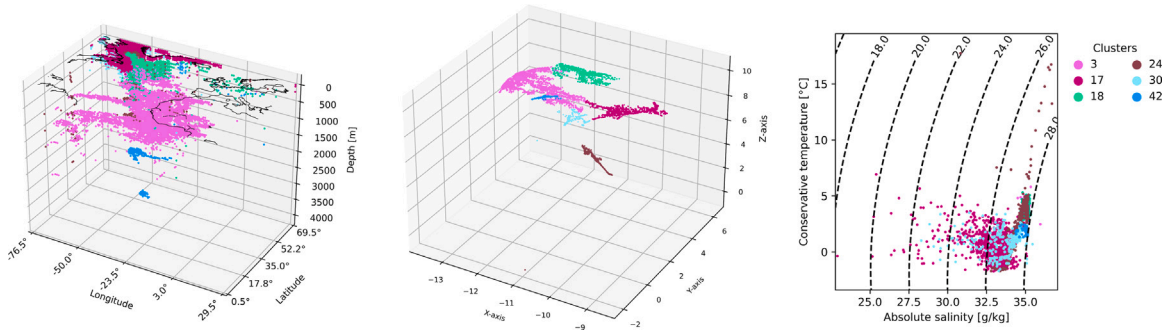


Fig. 13. Selected clusters in the Labrador Sea and Davis Strait (labels 3, 17, 18, 24, 30, 42) in geographic (left), embedded (middle) and temperature-salinity (TS) space (right).

geographically to the central area (0 – 500 m) and an area along the American east coast (1,500 m).

4. Discussion

4.1. Comparison of clustering algorithm performance

DBSCAN applied to embedded space created through UMAP best suited the input data. It outperformed k-Means, agglomerative Ward clustering and DBSCAN on original data, as seen in external validation, specifically the visualisation in geographic and embedding spaces. K-Means and (to a smaller extent) Ward clustering were not able to distinguish small data structures in the embedded space, where clusters were non-separated or merged (Section 3.2). We assess that the main reason for the superiority of DBSCAN is its ability to detect clusters of any shape since it operates on data density (Ahmad and Dang, 2015), i.e., it identifies areas where points are concentrated. Also, DBSCAN proved to work well in a similar use case, where it is applied to a dimensionality-reduced data space using t-Stochastic Neighbour Embedding (t-SNE) (Sonnewald et al., 2020). Clusters with varying densities pose challenges for DBSCAN (Ahmad and Dang, 2015), which may explain the occurrence of small clusters. Erroneous data could also contribute to this issue. A key result was that CVIs based on similarity, density and neighbourhood structures were inconsistent and thus not helpful. This has wide implication for studies relying on only one or a few CVIs as is common in the geosciences.

DBSCAN on original data led to unsatisfactory results, as none of the tested hyperparameter combinations yielded a clustering that reflected the actual data structure, neither in embedded nor in geographic space. The focus on the Baltic Sea, whose salinity levels are well below oceanic average (Tomczak and Godfrey, 2003), is consistent with the feature importances revealing salinity (35%) as the most important feature followed by silicate (19%).

K-Means is a common and fast clustering method and therefore used as a baseline in this work. However, k-Means was not able to reflect the data structure. When applied to the original data, the CVIs agreed on two as the optimal number of clusters. However, the scores cannot be computed for less than two clusters and there is no clustering for less than two clusters. Hence, this result indicated that either (i) two clusters was indeed the best number of clusters or (ii) k-Means could not separate the data structure into relevant clusters or (iii) the scores were not meaningful. For only two clusters, the clustering is mainly a trivial hot-cold separation as seen by the high feature importance of temperature. Mapping the clusters into the embedded space highlighted how k-Means operates, i.e., drawing straight cuts through data structures. Further visual investigation with higher numbers of clusters revealed a tendency to form globular clusters and no increase in clustering quality, i.e., better representation of the embedded data structure.

When applying k-Means to the embedding, the CVIs reached their optimum beyond two clusters. Compared to clustering on the original data, this indicated that k-Means was now better able to detect structures that were also recognised by the scores. However, the scores did not agree: While SH reached its optimum for eight and DB for ten clusters, the DB rose beyond the selected value range. This disagreement emphasised the need for selecting scores carefully and consulting

not only one but multiple scores. Still, the clustering exhibited obvious flaws as it could e.g. not distinguish clearly separate structures in embedding space, such as the deep water. Again, a likely explanation is that k-Means detects spherical clusters best (Ahmad and Dang, 2015; Jain, 2010; Harris and De Amorim, 2022), which were not observed in the embedded space. This clearly indicated that k-Means was not appropriate for the data and might not be for other non-linear use cases either, which are frequent in environmental sciences. Moreover, k-Means is sensitive to the initialisation of cluster centroids (Jain, 2010; Harris and De Amorim, 2022) introducing a variance not investigated in this study.

In contrast to k-Means and DBSCAN, hierarchical clustering provides a complete hierarchy of clusters. In this work, agglomerative Ward detected the overall data structures in embedding space well (Section 3.2.2) but neglected smaller data point collections. Ward linkage is mathematically close to a k-Means algorithm in a hierarchical context since both methods try to minimise the same objective function, namely the within-cluster-sum-of-squares (Murtagh and Legendre, 2014). It is therefore reasonable that they also had similar score curves.

Within the NEMI framework, the combination of 100 ensemble runs showed sensitivity to the *base_id* parameter. However, initial experiments showed that the standard deviation of mean uncertainties across runs with different *base_id*s was only 1%. An alternative to NEMI for combining cluster sets from an ensemble are averages over the proximity matrices (whose ij-entry is one if ith and jth point are in the same cluster, else zero) of a UMAP-clustering pipeline (Bollon et al., 2022). This average is then partitioned using spectral clustering.

In the final DBSCAN clustering, some clusters were not geographically contiguous and appear as geographically disjoint regions with similar water mass properties. While spatial coherence can be desirable for certain applications, first experiments with a spatially constrained Ward clustering (using a 52-nearest-neighbour graph) showed that enforcing geographic continuity led to lower similarity with established oceanic classifications such as the EMUs (Sayre et al., 2017) and Longhurst provinces (Longhurst, 2007), quantified by NMI and ARI. This reveals the trade-off between feature homogeneity and geospatial contiguity (cf. e.g. (Yuan et al., 2015; Wang et al., 2024)). Spatial constraints may obscure meaningful biogeochemical patterns. Not enforcing spatial constraints, on the other hand, may reveal physically or biogeochemically similar water masses across distant regions or alternatively, highlight limitations in the feature set's ability to distinguish regional differences. Future work may investigate this aspect in more detail, e.g. by a binarised spatially-constrained spectral clustering as suggested by Yuan et al. (2015).

In this study, min–max scaling was applied prior to embedding and clustering to ensure equal contributions of features with varying ranges. Since the distribution of data used in this work did contain skew, a robust scaler that is less affected by outliers may be preferable. Downstream evaluation using UMAP quality metrics (Qlocal, Qglobal, trustworthiness and continuity) revealed that min–max scaling consistently outperformed robust scaling (RobustScaler, Python library scikit-learn, version 1.5) across multiple hyperparameter settings. This effect may be explained by UMAP relying by default on a distance metric to construct its neighbourhood graph (McInnes et al., 2018a) rendering the method sensitive to how features are scaled. When extreme values represent meaningful physical or biogeochemical conditions rather than noise, scaling methods that preserve the full value range, such as min–max scaling, may therefore help maintain relevant relationships.

A focus of future research is the data preparation, especially with regard to the imputation. Also, density was set to a constant value of $1.025\,\mathrm{kg\,m^{-3}}$ for unit conversions as suggested previously (Korablev et al., 2021) when temperature or salinity values are unavailable. Feature importances computed by random forests were only used to get a first impression on the influence of parameters on cluster sets. For a thorough analysis, the models require further tuning and validation.

4.2. The importance of UMAP embedding for clustering performance

A key result was that all clustering algorithms performed better when applied to the embedding despite the relatively low dimensional original data space, confirming previous findings (Allaoui et al., 2020; Herrmann et al., 2023; Sonnewald et al., 2020). Additionally, the embedded data shows a more balanced feature contribution across all parameters (Fig. S14), reflecting a potentially less biased clustering result. This may lead to clusters that are influenced by multiple factors rather than being driven by a single dominant feature. A possible reason for the performance gain using UMAP is that it enhances separability and thus the ability to cluster data (Herrmann et al., 2023). Moreover, working on a space with fewer dimensions accelerates computation and optimises memory consumption of the given down-stream clustering tasks, and supports external validation by visualisation in the 3D space, where the six original dimensions could not have been used. As noted above, relying on CVIs would have resulted in misleading clusters that did not fairly represent the data. Working on the embedding helps to overcome the "curse of dimensionality" (Ayesha et al., 2020) that encompasses all phenomena related to higher dimensional data resulting in challenges for learning algorithms. For example, the amount of necessary training data grows exponentially with the number of dimensions to prevent overfitting. Also, Euclidean distances, which are used in all three clustering algorithms, become less discriminative in higher dimensional spaces (Verleysen and François, 2005).

Similar to clustering, dimensionality reduction is an unsupervised task with hyperparameters that need to be tuned using internal and/or external validation. In this study, embedding quality was externally evaluated based on visual clusterability, i.e. how clearly distinct and compact the clusters appear in the 3D embedded space and by assessing if the uncertainty of UMAP over multiple runs was within acceptable bounds. Alternatively, hyperparameters can be optimised using various internal metrics. Here, optimising for Qlocal resulted in a more dispersed embedding and notably impaired subsequent cluster sets. This suggests a potential trade-off between faithful Euclidean structure preservation and clear cluster formation. The discrepancy may also be explained by misaligned objectives: While the Q-metrics optimise distance-based rank preservation, UMAP optimises information preservation using cross-entropy. Another approach to tune UMAP hyperparameters would be to use CVIs as proxies for clusterability (e.g. Jouilili et al. (2024)). However, performing a joint grid search over UMAP and clustering parameters increases computational cost substantially and CVIs are not inherently reliable; they may not consistently reflect meaningful structure or lead to improved embedding or clustering outcomes (see Section 4.3). The final embedding preserved global structure well, as indicated by a high Qglobal, and also achieved high trustworthiness and continuity, suggesting good preservation of local neighbourhood membership. A lower Qlocal, however, indicated that the fine-grained local neighbourhood rank ordering was more distorted. For the given clustering tasks, this level of distortion is acceptable, as precise ranks are typically less critical than maintaining broader neighbourhood consistency. This study used UMAP's default Euclidean distance and comparative tests with cosine, Mahalanobis, Chebyshev and Manhattan distances based on Qlocal and Qglobal supported this choice. The superiour performance of Euclidean distance is likely due to its ability to preserve meaningful absolute differences across the scaled oceanographic features and its alignment with UMAP's local neighbourhood assumptions. It should be noted, however, that the Q-metrics are based on Euclidean distances and may therefore introduce bias.

Due to the non-linear nature of the used environmental data, linear dimensionality reduction techniques were discarded such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Singular Value Decomposition (SVD), Latent Semantic Analysis (LSA), Locality Preserving Projections (LPP), Independent Component Analysis (ICA) and Project Pursuit (PP) (Nanga et al., 2021). Popular non-linear methods include Kernel Principal Component Analysis (KPCA),

Multi-Dimensional Scaling (MDS), Isomap, Locally Linear Embedding (LLE), Self-Organizing Map (SOM), Latent Vector Quantization (LVQ), t-Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (Nanga et al., 2021). With the perspective to apply the method to a larger dataset in the future (e.g. by considering a larger geographic area or time as an additional dimension), slow dimensionality reduction methods that do not scale well were omitted (KPCA, ISOMAP, LVQ, t-SNE) (Nanga et al., 2021). Further informing our choice, SOMs are computationally demanding (Vesanto et al., 2000) and MDS and LLE are sensitive to noise (Nanga et al., 2021) that could be present in the input data. Therefore, UMAP was favoured because it is able to preserve non-linear structures and to scale well. As noted in Section 3.1 and detailed in Supplementary Material B.1, the cross-entropy method UMAP uses for optimisation was also seen as highly advantageous through its ability to strengthen associations between the data, which facilitated subsequent clustering.

4.3. Choice and interpretability of cluster validity indices (CVIs)

CVIs for comparing clustering methods showed limited agreement with external validation. Despite clearly being the best choice, DBSCAN received worst ranks according to the classical CVIs (CH, DB and SH) and CVNNH, except for SH, which assigned DBSCAN on original data highest rank despite it being the visually poorest subdivision. These scores clearly favoured k-Means and Ward on the original data, reflecting their bias towards globular, convex clusters (details on convexity in Supplementary Material B.3). Conversely, CDR preferred DBSCAN for clustering original and embedded data, likely due to its different notion of high clustering quality, defined by small local density variations better captured by DBSCAN. Since each score has its own bias, i.e. imposes different assumptions on data and clusters, they are not useful for performance comparison across clustering algorithms. Thrun (2021) support this by arguing that instead of selecting a clustering algorithm based on a CVI, the same result would be achieved by directly optimising for that CVI (Thrun, 2021). Nonetheless, the CVIs are useful for tuning hyperparameters since their bias is constant throughout the experiments.

Evaluating the impact of embedding with CVIs requires caution, as results did not always align with external validation. Most CVIs, including SH, CH applied to k-Means, DB applied to DBSCAN, CVNNH and CDR for all clustering methods, scored worse on the embedding than on original data, which conflicted with the previous visual finding that the clusterings benefited from the preceding dimensionality reduction. A potential explanation may be the scale sensitivity of the CVIs (except SH and k-DBCV): UMAP inflated the maximum pairwise distanced from about 1.75 to around 25 (Fig. S4) increasing absolute distances. This could be further evaluated in future work by scaling embedded dimensions before score computation. K-DBCV, a scale-independent score, could not be computed for clustering original data suggesting that the original feature space lacked a clear density-based structure. After UMAP, k-DBCV scores slightly increased but remained negative indicating weak density separability. Visual inspection supported this, revealing few distinct, arbitrary-shaped clusters embedded in a more continuous structure with irregular boundaries. This aligns with SH deteriorating post-embedding since it is incompatible with non-convex geometry.

These results highlight the need for careful CVI selection and interpretation. The choice of CVIs depends on the context and structure of the data and the nature of clusters, as each index offers a unique perspective on the clustering. For example, despite assuming convexity, CH did reflect DBSCAN cluster quality to some extent, indicating some flexibility of the scores. Generally, Arbelaitz et al. (2013) found that presence of overlapping clusters or noise significantly impaired performance of the 30 CVIs they investigated. Another impact factor is cluster shape: Some CVIs are more suitable for globular clusters, while others are better equipped to handle arbitrarily shaped clusters, like DBCV or CDR (Schlake and Beecks, 2024). Other scores could be tested to tune hyperparameters of the clustering methods, like the Dunn index (Dunn, 1973), WB index (a weighted ratio of sum-of-squares within and sum-of-squares between clusters, Zhao and Fränti (2014)), I index (Maulik and Bandyopadhyay, 2002), Cluster Validity index based on Density-involved Distance (CVDD, Hu and Zhong (2019)) or Distance-based Separability Index (DSI, Guan and Loew (2020)). Each has its own mathematical assumptions about the data (Thrun, 2021). CVDD e.g. claims that it can deal with both spherical and non-spherical clusters. To assess ecological similarity, indices such as the Jaccard and Bray-Curtis indices have been utilised (e.g. Carteron et al. (2012), Sonnewald et al. (2020)).

4.4. Uncertainty quantification

The uncertainty and reproducibility of the best-performing clustering method (UMAP-DBSCAN) was evaluated using overlap (between all UMAP-DBSCAN runs) and RMSE as a measure. DBSCAN is sensitive to the sequence of input samples (Tran et al., 2013), which was here determined to be negligible (overlap: $99.99 \pm 0.003\%$). UMAP uses randomness as it implements stochastic gradient descent for an efficient optimisation (McInnes et al., 2018a). With an average RMSE between the data points of the 100 embeddings of 0.22, or 1.3% of the value range, the procedure was assessed reproducible on the given data. Consistency across multiple runs is supported by low standard deviations of the four computed dimensionality reduction scores. The combination of UMAP followed by DBSCAN had a mean overlap of $88.81 \pm 1.8\%$, corresponding to about 11% uncertainty. Besides this high point-level compliance, the cluster sets also exhibited strong consistency in information content, as reflected by the high NMI (0.91 ± 0.01). The ARI (0.78 ± 0.05) indicates some variability, confirmed by the grid cell-wise uncertainty ($15.49 \pm 20\%$), which is likely caused by the sensitivity of the pipeline to UMAP. In summary, both, UMAP and DBSCAN, yielded robust and reproducible results, both individually and in combination, with only minor variations in the latter.

Uncertainty can be a factor for deciding on a clustering method since a reproducible clustering is often desired. Variance of k-Means and agglomerative Ward was not further investigated, though both methods can differ over multiple runs (Harris and De Amorim, 2022; Gordon, 1987).

4.5. Relevance for ecological interpretations

The biogeochemical clustering approach in our study has strong connections to previous approaches. In comparison to the ecological and biogeochemical regionalisations by Longhurst (2007) and Sayre et al. (2017), the clustering of this study resulted in similar but more detailed clusters with some variation in the spatial extents (Figs. 14, S15, S16) with stronger similarities to EMUs (Table 2). While moderate NMI values for both subdivisions indicate shared information content and a degree of structural correspondence, relatively low ARI values suggest larger differences in exact partitioning. It is obvious that the physical and biogeochemical conditions are closely connected to the characteristics of marine biomes, such as primary production (e.g. Taylor et al. (2011)), microbial diversity (e.g. Friedline et al. (2012)) or cycling of organic matter (e.g. Koch and Kattner (2012), Schmitt-Kopplin et al. (2012), Hertkorn et al. (2013)). For example, the Labrador Sea showed strong similarities across the three clustering sets, though (Longhurst, 2007) suggested a coarser subdivision. Similarly, Longhurst (2007) and Sayre et al. (2017) identified only one or few regions in the Mediterranean Sea, whereas this study resulted in a total of 27 clusters. This higher number of regions may arise from the presented method being data-driven in contrast to Longhurst's knowledge-guided approach, more sensitive to local structure and not imposing constraints on cluster number, shape, or size. Moreover, the

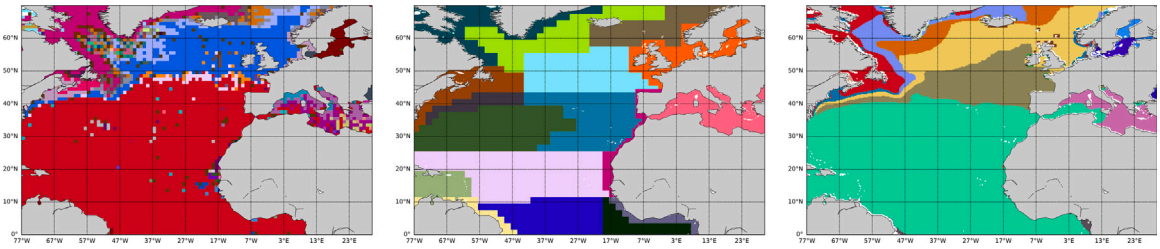


Fig. 14. Surface water masses as defined in this study (UMAP-DBSCAN, left), by Longhurst (2007) (middle) and Sayre et al. (2017) (right). While similarities e.g. in the Labrador Sea stand out, there were also differences, such as (Longhurst, 2007)'s subdivision of the central North Atlantic.

COMFORT dataset includes the data used in Sayre et al. (2017) and incorporates additional measurements. Previous studies are based on *a priori* decisions on the total number of clusters — based on the data or the parameters applied for the clustering method. For example, k-Means (as used by Sayre et al. (2017)) uses a predefined number of clusters and optimises for the entire dataset. This can result in smoothing over local variations, with consequences for the ecological interpretations of the regions. DBSCAN in our study does not enforce a specific number of clusters but prescribes the connectivity conditions, i.e., how far apart data points in feature space may be to form a cluster, which promotes fine-grained subdivisions. In global ecological studies, the Mediterranean is often described as one region (e.g. Longhurst (2007), Costello et al. (2017)) though works like (Sayre et al., 2017; Zhao et al., 2020a) and this study suggest a higher diversity especially at the surface in comparison, for example, to the North Atlantic. This is likely caused by complex upper ocean currents (Tomczak and Godfrey, 2003) and small-scale patterns of seasonal primary production (as represented in https://www.grida.no/resources/5937). A short analysis of the proportion of endemic species per region using OBIS data (Ocean Biodiversity Information System (OBIS) (2025), data not shown) revealed 12% of occurring species in the largest and in the second largest Mediterranean clusters (labels 6, 35) to be endemic suggesting the ecological uniqueness of the main biogeochemical water masses in the Mediterranean Sea.

An example for varying region extent in different clustering approaches is the western deep North Atlantic (cluster 11). In this study using DBSCAN and UMAP, the region extended further south along the American coast compared to the Ecological Marine Units by Sayre et al. (2017). By using k-Means, we were able to reproduce the spatial extend in the previous work. The visualisation in embedding space revealed that k-Means struggled to adequately separate this area. This might be attributable to the complexity of the data, visible in embedding space as an irregularly connected, curved structure, not resembling normal distributions for which k-Means is optimised. Despite the intrinsic bias of k-Means, the clustering by Sayre et al. (2017) and the DBSCAN clustering presented here exhibited many similarities (e.g. surface regions, Fig. 14). A possible reason is that Sayre et al. (2017) use a higher depth resolution: In total, 102 depth intervals were defined and the very variable first 100 m of the ocean column are subdivided into 5 m steps. This higher resolution might enable a better separation of data points in feature space and thus a more precise clustering.

Generally, the presented cluster set picked up well-known oceanographic features, like the outflow of warm, saline Mediterranean water through the Strait of Gibraltar (Pinardi et al., 2023) that is traceable at the 2,000 m level across the North Atlantic Ocean (Tomczak and Godfrey, 2003). Due to the time-averaging of data, small-scale and dynamic oceanographic features such as eddies were not sufficiently represented in this cluster set. Another well-represented feature is the subdivision of deep Atlantic waters along the north-west axis. Compared to the east, the western waters were characterised by lower silicate, nitrate and phosphate and higher oxygen concentration (Fig. 12). This is in good agreement with the fact that the western part of the deep North Atlantic is more influenced by relatively young North Atlantic deep water, while there is more influence of Southern Ocean deep waters in the east (Johnson, 2008). This emphasises that while temperature and salinity remain key parameters in defining water masses, the inclusion of additional parameters such as oxygen and nutrients is crucial for a comprehensive and detailed analysis of water mass properties and dynamics in the ocean.

The clustering results for the case study in the Labrador Sea and Davis Strait, an area of deep-water formation, were generally in very good agreement with pertinent oceanographic literature. Those clusters that represented freshly formed deep water (particularly labels 3 and 24) were characterised by high salinity, low temperature and slightly depleted oxygen values, as shown previously e.g. by Tomczak and Godfrey (2003). The clustering did not always yield spatially coherent clusters, for example a cluster in the deep Atlantic near the equator (label 7). Despite fairly different temperatures, the same cluster label was also assigned to water in the Labrador Sea, because of a high similarity in the inorganic nutrient concentrations. A possible reason is the exclusion of geographic coordinates (or proximity to coasts, such as in Longhurst (2007)) from the clustering process or that additional parameter(s) are required for the distinction.

5. Conclusion and outlook

By comparing pre-processings, clustering methods and various validation techniques, this study found a clustering that adequately reflected the embedded data structure of North Atlantic physical and biogeochemical properties. Such a methodological approach to clustering is of high importance for quality and hence potential downstream tasks. Thrun (2021) formulated this precisely referring to their medical use case:

> "[...] only the combination of empirical medical knowledge and an unbiased, structure-based choice of the optimal cluster analysis method w.r.t. the data will result in precise and reproducible clustering with the potential for knowledge discovery of high clinical value".
>
> [— (Thrun, 2021)]

DBSCAN applied to a dimensionality-reduced space using UMAP best reflected the data structure, outperforming k-Means and agglomerative Ward. When validating the results, it was imperative to not rely on single criteria, e.g. to compute multiple CVIs for hyperparameter tuning. The presented results moreover discourage using CVIs for the comparison between clustering methods.

For reproducibility purposes, analysis of uncertainty is an important aspect to consider when non-deterministic algorithms are applied. The variability of the presented method was quantified using ensemble analysis revealing low variabilities of the individual methods (UMAP, DBSCAN) and slight deviations in the clustering when combined (overlap of 88.81 ± 1.8%). By combining the clustering results over the ensemble following the NEMI framework, reproducibility and representativeness of the statistical co-variance space was further increased (uncertainty of 15.49 ± 20%).

There were several aspects that could be further explored related to data, pre-processing, method and post-processing. Regarding the data, further parameters like dissolved (in-)organic carbon or biogeochemical tracers such as Apparent Oxygen Utilisation (AOU) could be added. Future work aims to scale the clustering up to global coverage and add the temporal component to increase oceanographic utility. For this, data sparsity could be a limiting factor machine learning is especially suited to overcome. Also, other clustering methods that are able to deal with varying densities, such as HDBSCAN, are worth exploring. Further, self-organising maps (Kohonen, 1990) could increase interpretability of results by providing meaningful maps of the classes while preserving data topology (Yonggang and Weisenberg, 2011). To further improve performance, other hyperparameters can be explored. Distance metrics other than Euclidean could be investigated for Ward and DBSCAN clustering to ensure the optimal strategy for the given data. Oceanographically, the cluster set can be compared more extensively to existing definitions to potentially extract new knowledge.

CRediT authorship contribution statement

Yvonne Jenniges: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Maike Sonnewald:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Sebastian Maneth:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Data curation, Conceptualization. **Are Olsen:** Writing – review & editing, Data curation. **Boris P. Koch:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our gratitude to the COMFORT project group for their invaluable contribution in providing the dataset that formed the foundation of this research. We are thankful to Claire Monteleoni and Anastase Alexandre Charantonis for their insightful discussions that helped shape the direction of this study. Also, we wish to thank Marlo Bareth for being a valuable discussion partner, whose perspectives helped refine the analysis. YJ was funded through the Helmholtz School for Marine Data Science, Germany (MarDATA, Grant No. HIDSS-0005). This work was supported by a fellowship of the German Academic Exchange Service (DAAD), allowing for a research stay of YJ with MS at Princeton University and University of Washington, Seattle. We acknowledge support by the Open Access publication fund of Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Germany. MS acnowledges support from the Award NA24OARX431C0058-T1-01 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce and the Award RISE-2425906 from the U.S. National Science Foundation.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103390.

Data availability

The cluster set and auxiliary information is publicly available on Zenodo (Oceanregionsdataset) along with the code base used to conduct and analyse the presented experiments (Code) and the dashboard to explore the final cluster set (Dashboard, Dashboardcode). Column descriptions can be found in Supplementary Material D. The COMFORT dataset is publicly available online (Korablev and Olsen, 2022).

References

Ahmad, P.H., Dang, S., 2015. Performance evaluation of clustering algorithm using different datasets. Int. J. Adv. Res. Comput. Sci. Manag. Stud. 3 (1), 167–173.

Allaoui, M., Kherfi, M.L., Cheriet, A., 2020. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (Eds.), Image and Signal Processing. Springer International Publishing, pp. 317–325. http://dx.doi.org/10.1007/978-3-030-51935-3_34.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., PéRez, J.M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recognit. 46 (1), 243–256. http://dx.doi.org/10.1016/j.patcog.2012.07.021.

Arnoldi, N.S., Litvin, S.Y., Madigan, D.J., Micheli, F., Carlisle, A., 2023. Multi-taxa marine isoscapes provide insight into large-scale trophic dynamics in the north Pacific. Prog. Oceanogr. 213, 103005. http://dx.doi.org/10.1016/j.pocean.2023.103005.

Ayesha, S., Hanif, M.K., Talib, R., 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf. Fusion 59, 44–58. http://dx.doi.org/10.1016/j.inffus.2020.01.005.

Bailey, R.G., 1998. Ecoregions, first ed. Springer New York, NY, http://dx.doi.org/10.1007/978-1-4612-2200-2.

Bollon, J., Assale, M., Cina, A., Marangoni, S., Calabrese, M., Salvemini, C.B., Christille, J.M., Gustincich, S., Cavalli, A., 2022. Investigating how reproducibility and geometrical representation in UMAP dimensionality reduction impact the stratification of breast cancer tumors. http://dx.doi.org/10.3390/app12094247.

Briggs, J.C., 1974. Marine Zoogeography. McGraw-Hill, New York.

Briggs, J.C., 1995. Global biogeography. In: Developments in Palaeontology and Stratigraphy, vol. 14, Elsevier, p. 452. http://dx.doi.org/10.1016/S0920-5446(06)80051-8.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., Gorsky, G., Gregory, A.C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B.T., Schwenck, S.M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans, C., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., Sullivan, M.B., 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348 (6237), 1261498. http://dx.doi.org/10.1126/science.1261498.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3 (1), 1–27. http://dx.doi.org/10.1080/03610927408827101.

Carteron, A., Jeanmougin, M., Leprieur, F., Spatharis, S., 2012. Assessing the efficiency of clustering algorithms and goodness-of-fit measures using phytoplankton field data. Ecol. Inform. 9, 64–68. http://dx.doi.org/10.1016/j.ecoinf.2012.03.008.

Costello, M.J., Tsai, P., Wong, P.S., Cheung, A.K.L., Basher, Z., Chaudhary, C., 2017. Marine biogeographic realms and species endemicity. Nat. Commun. 8 (1), 1057. http://dx.doi.org/10.1038/s41467-017-01121-2.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2), 224–227. http://dx.doi.org/10.1109/TPAMI.1979.4766909.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., Chisholm, S.W., Karl, D.M., 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311 (5760), 496–503. http://dx.doi.org/10.1126/science.1120250.

Devred, E., Sathyendranath, S., Platt, T., 2007. Delineation of ecological provinces using ocean colour radiometry. Mar. Ecol. Prog. Ser. 346, 1–13. http://dx.doi.org/10.3354/meps07149.

Dunn, J., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. 3 (3), 32–57. http://dx.doi.org/10.1080/01969727308546046.

Ekman, S., 1935. Tiergeographie Des Meeres. Akademische Verlagsgesellschaft, Leipzig (Germany), p. 542.

Emery, W.J., 2001. Water types and water masses. In: Steele, J.H. (Ed.), Encyclopedia of Ocean Sciences. Academic Press, Oxford, pp. 3179–3187. http://dx.doi.org/10.1006/rwos.2001.0108.

Fay, A.R., Mckinley, G.A., 2014. Global open-ocean biomes: mean and temporal variability. Earth Syst. Sci. Data 6 (2), 273–284. http://dx.doi.org/10.5194/essd-6-273-2014.

Forbes, E., 1856. Map of the distribution of marine life. In: Johnston, A.K. (Ed.), The Physical Atlas of Natural Phenomena, second ed. William Blackwood and Sons, Edinburgh (Scotland), pp. 99–102.

Fränti, P., Sieranoja, S., 2019. How much can k-means be improved by using better initialization and repeats? Pattern Recognit. 93, 95–112. http://dx.doi.org/10.1016/j.patcog.2019.04.014.
Friedline, C.J., Franklin, R.B., McCallister, S.L., Rivera, M.C., 2012. Bacterial assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal biogeographic signatures. Biogeosciences 9 (6), 2177–2193. http://dx.doi.org/10.5194/bg-9-2177-2012.
Gloege, L., McKinley, G.A., Mouw, C.B., Ciochetto, A.B., 2017. Global evaluation of particulate organic carbon flux parameterizations and implications for atmospheric pCO2. Glob. Biogeochem. Cycles 31 (7), 1192–1215. http://dx.doi.org/10.1002/2016GB005535.
Gordon, A.D., 1987. A review of hierarchical classification. J. R. Stat. Soc. Ser. A (General) 150 (2), 119–137. http://dx.doi.org/10.2307/2981629.
Guan, S., Loew, M., 2020. An internal cluster validity index using a distance-based separability measure. In: 32nd International Conference on Tools for Artificial Intelligence. IEEE, pp. 827–834. http://dx.doi.org/10.1109/ICTAI50040.2020.00131.
Halkidi, M., Vazirgiannis, M., Hennig, C., 2015. Method-independent indices for cluster validation and estimating the number of clusters. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. Chapman and Hall/CRC, p. 24.
Hammer, J.L., Devanny, A.J., Kaufman, L.J., 2024. in review. Density-based optimization for unbiased, reproducible clustering applied to single molecule localization microscopy. BioRxiv. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/2024.11.01.621498.
Hardman-Mountford, N.J., Hirata, T., Richardson, K.A., Aiken, J., 2008. An objective methodology for the classification of ecological pattern into biomes and provinces for the pelagic ocean. Remote Sens. Environ. 112 (8), 3341–3352. http://dx.doi.org/10.1016/j.rse.2008.02.016.
Harris, S., De Amorim, R.C., 2022. An extensive empirical comparison of k-means initialization algorithms. IEEE Access 10, 58752–58768. http://dx.doi.org/10.1109/ACCESS.2022.3179803.
Hayden, B.P., Ray, G.C., Dolan, R., 1984. Classification of coastal and marine environments. Environ. Conserv. 11 (3), 199–207. http://dx.doi.org/10.1017/S0376892900014211.
Hedgpeth, J.W., 1957. Marine biogeography. In: Hedgpeth, J.W. (Ed.), Treatise on Marine Ecology and Paleoecology. Geological Society of America, http://dx.doi.org/10.1130/MEM67V1-p359.
Herrmann, M., Kazempour, D., Scheipl, F., Kröger, P., 2023. Enhancing cluster analysis via topological manifold learning. Data Min. Knowl. Discov. http://dx.doi.org/10.1007/s10618-023-00980-2.
Hertkorn, N., Harir, M., Koch, B.P., Michalke, B., Schmitt-Kopplin, P., 2013. High-field NMR spectroscopy and FTICR mass spectrometry: powerful discovery tools for the molecular level characterization of marine dissolved organic matter. Biogeosciences 10 (3), 1583–1624. http://dx.doi.org/10.5194/bg-10-1583-2013.
Hörstmann, C., Buttigieg, P.L., John, U., Raes, E.J., Wolf-Gladrow, D., Bracher, A., Waite, A.M., 2022. Microbial diversity through an oceanographic lens: refining the concept of ocean provinces through trophic-level analysis and productivity-specific length scales. Environ. Microbiol. 24 (1), 404–419. http://dx.doi.org/10.1111/1462-2920.15832.
Hu, L., Zhong, C., 2019. An internal validity index based on density-involved distance. IEEE Access 7, 40038–40051. http://dx.doi.org/10.1109/ACCESS.2019.2906949.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classification 2 (1), 193–218. http://dx.doi.org/10.1007/BF01908075.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31 (8), 651–666. http://dx.doi.org/10.1016/j.patrec.2009.09.011.
Johnson, G.C., 2008. Quantifying Antarctic Bottom Water and North Atlantic Deep Water volumes. J. Geophys. Res.: Ocean. 113 (C5), http://dx.doi.org/10.1029/2007JC004477.
Jouilili, A., Hantouti, H., Ouazzani, R.E.L., 2024. Optimizing dimensionality reduction in SDN: A metaheuristic approach of UMAP parameter tuning. In: 2024 5th International Conference on Communications, Information, Electronic and Energy Systems. CIEES, pp. 1–6. http://dx.doi.org/10.1109/CIEES62939.2024.10811181.
Juan Jordá, M.J., Nieblas, A.E., Hanke, A., Tsuji, S., Andonegi, E., Di Natale, A., Kell, L., Diaz, G., Alvarez Berastegui, D., Brown, C., Die, D., Arrizabalaga, H., Yates, O., Gianuca, D., Niemeyer Fiedler, F., Luckhurst, B., Coelho, R., Zador, S., Dickey-Collas, M., Pepin, P., Murua, H., 2022. Report of the ICCAT Workshop on the Identification of Regions in the ICCAT Convention Area for Supporting the Implementation of the Ecosystem Approach to Fisheries Management. Technical Report, International Commission for the Conservation of Atlantic Tunas, URL: https://www.iccat.int/en/pubs_CVSP.html.
Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., Müller, K.R., 2024. From clustering to cluster explanations via neural networks. IEEE Trans. Neural Netw. Learn. Syst. 35 (2), 1926–1940. http://dx.doi.org/10.1109/TNNLS.2022.3185901.
Kavanaugh, M.T., Hales, B., Saraceno, M., Spitz, Y.H., White, A.E., Letelier, R.M., 2014. Hierarchical and dynamic seascapes: A quantitative framework for scaling pelagic biogeochemistry and ecology. Prog. Oceanogr. 120, 291–304. http://dx.doi.org/10.1016/j.pocean.2013.10.013.
Koch, B.P., Kattner, G., 2012. Sources and rapid biogeochemical transformation of dissolved organic matter in the Atlantic surface ocean. Biogeosciences 9, 2597–2602. http://dx.doi.org/10.5194/bg-9-2597-2012.
Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78 (9), 1464–1480.
Korablev, A., Olsen, A., 2022. COMFORT dataset. http://dx.doi.org/10.11582/2022.00039, Norstore.
Korablev, A., Olsen, A., Jones, S.D., 2021. Report on COMFORT Dataset Compilation. Technical Report, University of Bergen.
Kvalseth, T.O., 1987. Entropy and correlation: Some comments. IEEE Trans. Syst. Man Cybern. 17 (3), 517–519. http://dx.doi.org/10.1109/TSMC.1987.4309069.
Lee, J.A., Verleysen, M., 2010. Scale-independent quality criteria for dimensionality reduction. Pattern Recognit. Lett. 31 (14), 2248–2257. http://dx.doi.org/10.1016/j.patrec.2010.04.013.
Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., Wu, S., 2013. Understanding and enhancement of internal clustering validation measures. IEEE Trans. Cybern. 43 (3), 982–994. http://dx.doi.org/10.1109/TSMCB.2012.2220543.
Liu, M., Tanhua, T., 2021. Water masses in the Atlantic ocean: characteristics and distributions. Ocean. Sci. 17 (2), 463–486. http://dx.doi.org/10.5194/os-17-463-2021.
Logan, J., Pethybridge, H., Lorrain, A., Somes, C., Allain, V., Bodin, N., Choy, C., Duffy, L., Goñi, N., Graham, B., Langlais, C., Ménard, F., Olson, R., Young, J., 2020. Global patterns and inferences of tuna movements and trophodynamics from stable isotope analysis. Deep. Sea Res. Part II: Top. Stud. Ocean. 175, 104775. http://dx.doi.org/10.1016/j.dsr2.2020.104775, Oceanic biodiversity under climate change: shifts in natural and human systems.
Longhurst, A.R., 2007. Ecological Geography of the Sea, second ed. Academic Press, Burlington, MA, USA, http://dx.doi.org/10.1016/B978-0-12-455521-1.X5000-1.
Longhurst, A., Sathyendranath, S., Platt, T., Caverhill, C., 1995. An estimate of global primary production in the ocean from satellite radiometer data. J. Plankton Res. 17 (6), 1245–1271. http://dx.doi.org/10.1093/plankt/17.6.1245.
Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.
Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. 24 (12), 1650–1654. http://dx.doi.org/10.1109/TPAMI.2002.1114856.
McInnes, L., Healy, J., Melville, J., 2018a. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. http://dx.doi.org/10.48550/arXiv.1802.03426, ArXiv E-Prints.
McInnes, L., Healy, J., Saul, N., Großberger, L., 2018b. UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. 3, 861. http://dx.doi.org/10.21105/joss.00861.
Moulavi, D., Jaskowiak, P.A., Campello, R.J.G.B., Zimek, A., Sander, J., 2014. Density-based clustering validation. In: Proceedings of the 2014 SIAM International Conference on Data Mining. SDM, pp. 839–847. http://dx.doi.org/10.1137/1.9781611973440.96.
Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? J. Classification 31 (3), 274–295. http://dx.doi.org/10.1007/s00357-014-9161-z.
Nanga, S., Bawah, A.T., Acquaye, B.A., Billa, M.I., Baeta, F.D., Odai, N.A., Obeng, S.K., Nsiah, A.D., 2021. Review of dimension reduction methods. J. Data Anal. Inf. Process. 9 (3), 189–231. http://dx.doi.org/10.4236/jdaip.2021.93013.
Ocean Biodiversity Information System (OBIS), 2025. OBIS occurrence data. http://dx.doi.org/10.25607/obis.occurrence.b89117cd, Dataset.
Oliver, M.J., Irwin, A.J., 2008. Objective global ocean biogeographic provinces. Geophys. Res. Lett. 35 (15), http://dx.doi.org/10.1029/2008gl034238.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12 (85), 2825–2830. http://dx.doi.org/10.5555/1953048.2078195.
Pinardi, N., Estournel, C., Cessi, P., Escudier, R., Lyubartsev, V., 2023. Chapter 7 - dense and deep water formation processes and Mediterranean overturning circulation. In: Schroeder, K., Chiggiato, J. (Eds.), Oceanography of the Mediterranean Sea. Elsevier, pp. 209–261. http://dx.doi.org/10.1016/B978-0-12-823692-5.00009-1.
Reisinger, R.R., Brooks, C.M., Raymond, B., Freer, J.J., Cotté, C., Xavier, J.C., Trathan, P.N., Bornemann, H., Charrassin, J.B., Costa, D.P., Danis, B., Hückstädt, L., Jonsen, I.D., Lea, M.A., Torres, L., Van de Putte, A., Wotherspoon, S., Friedlaender, A.S., Ropert-Coudert, Y., Hindell, M., 2022. Predator-derived bioregions in the Southern Ocean: Characteristics, drivers and representation in marine protected areas. Biol. Cons. 272, 109630. http://dx.doi.org/10.1016/j.biocon.2022.109630.
Reygondeau, G., Cheung, W.W.L., Wabnitz, C.C.C., Lam, V.W.Y., Frölicher, T., Maury, O., 2020. Climate change-induced emergence of novel biogeochemical provinces. Front. Mar. Sci. 7 (657), http://dx.doi.org/10.3389/fmars.2020.00657.
Reygondeau, G., Guieu, C., Benedetti, F., Irisson, J.O., Ayata, S.D., Gasparini, S., Koubbi, P., 2017. Biogeochemical regions of the Mediterranean Sea: An objective multidimensional and multivariate environmental approach. Prog. Oceanogr. 151, 138–148. http://dx.doi.org/10.1016/j.pocean.2016.11.001.
Reygondeau, G., Longhurst, A., Martinez, E., Beaugrand, G., Antoine, D., Maury, O., 2013. Dynamic biogeochemical provinces in the global ocean. Glob. Biogeochem. Cycles 27 (4), 1046–1058. http://dx.doi.org/10.1002/gbc.20089.
Rojas-Thomas, J.C., Santos, M., 2021. New internal clustering validation measure for contiguous arbitrary-shape clusters. Int. J. Intell. Syst. 36 (10), 5506–5529. http://dx.doi.org/10.1002/int.22521.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Rui, X., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Trans. Neural Netw. 16 (3), 645–678. http://dx.doi.org/10.1109/TNN.2005.845141.

Sayre, R.G., Wright, D.J., Breyer, S.P., Butler, K.A., Van Graafeiland, K., Costello, M.J., Harris, P.T., Goodin, K.L., Guinotte, J.M., Basher, Z., Kavanaugh, M.T., Halpin, P.N., Monaco, M.E., Cressie, N., Aniello, P., Frye, C.E., Stephens, D., 2017. A three-dimensional mapping of the ocean based on environmental data. Oceanography 30 (1), 90–103. http://dx.doi.org/10.5670/oceanog.2017.116.

Schlake, G.S., Beecks, C., 2024. Validating arbitrary shaped clusters - a survey. In: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics. DSAA, pp. 1–12. http://dx.doi.org/10.1109/DSAA61799.2024.10722773.

Schmitt-Kopplin, P., Liger-Belair, G., Koch, B.P., Flerus, R., Kattner, G., Harir, M., Kanawati, B., Lucio, M., Tziotis, D., Hertkorn, N., Gebefügi, I., 2012. Dissolved organic matter in sea spray: a transfer study from marine surface water to aerosols. Biogeosciences 9 (4), 1571–1582. http://dx.doi.org/10.5194/bg-9-1571-2012.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Trans. Database Syst. 42 (3), http://dx.doi.org/10.1145/3068335.

Sonnewald, M., 2023. A hierarchical ensemble manifold methodology for new knowledge on spatial data: An application to ocean physics. http://dx.doi.org/10.22541/essoar.168056792.25480169/v1, ESS Open Archive.

Sonnewald, M., Dutkiewicz, S., Hill, C., Forget, G., 2020. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. Sci. Adv. 6 (22), eaay4740. http://dx.doi.org/10.1126/sciadv.aay4740.

Sonnewald, M., Wunsch, C., Heimbach, P., 2019. Unsupervised learning reveals geography of global Ocean Dynamical Regions. Earth Space Sci. 6 (5), 784–794. http://dx.doi.org/10.1029/2018ea000519.

Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M., Halpern, B.S., Jorge, M.A., Lombana, A., Lourie, S.A., Martin, K.D., McManus, E., Molnar, J., Recchia, C.A., Robertson, J., 2007. Marine ecoregions of the world: A bioregionalization of coastal and shelf areas. BioScience 57 (7), 573–583. http://dx.doi.org/10.1641/B570707.

Strehl, A., Ghosh, J., 2002. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617.

Taylor, B.B., Torrecilla, E., Bernhardt, A., Taylor, M.H., Peeken, I., Röttgers, R., Piera, J., Bracher, A., 2011. Bio-optical provinces in the eastern Atlantic Ocean and their biogeographical relevance. Biogeosciences 8 (12), 3609–3629. http://dx.doi.org/10.5194/bg-8-3609-2011.

Thrun, M.C., 2021. Distance-based clustering challenges for unbiased benchmarking studies. Sci. Rep. 11 (1), 18988. http://dx.doi.org/10.1038/s41598-021-98126-1.

Tomczak, M., Godfrey, J.S., 2003. Regional Oceanography: An Introduction, second ed. Astral International Pvt Ltd, Delhi.

Tran, T.N., Drab, K., Daszykowski, M., 2013. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometr. Intell. Lab. Syst. 120, 92–96. http://dx.doi.org/10.1016/j.chemolab.2012.11.006.

Ullmann, T., Hennig, C., Boulesteix, A.L., 2022. Validation of cluster analysis results on validation data: A systematic framework. WIREs Data Min. Knowl. Discov. 12 (3), e1444. http://dx.doi.org/10.1002/widm.1444.

Venna, J., Kaski, S., 2006. Local multidimensional scaling. Neural Netw. 19 (6), 889–899. http://dx.doi.org/10.1016/j.neunet.2006.05.014.

Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction. In: Cabestany, J., Prieto, A., Sandoval, F. (Eds.), Computational Intelligence and Bioinspired Systems. Springer Berlin Heidelberg, pp. 758–770.

Vesanto, J., Himberg, J., Alhoniemi, E., Parkankangas, J., 2000. SOM Toolbox for Matlab 5. pp. 1–60, Report A57.

Vichi, M., Allen, J.I., Masina, S., Hardman-Mountford, N.J., 2011. The emergence of ocean biogeochemical provinces: A quantitative assessment and a diagnostic for model evaluation. Glob. Biogeochem. Cycles 25 (2), http://dx.doi.org/10.1029/2010gb003867.

Vinh, N.X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 11 (95), 2837–2854, URL: http://jmlr.org/papers/v11/vinh10a.html.

Wang, H., Song, C., Wang, J., Gao, P., 2024. A raster-based spatial clustering method with robustness to spatial outliers. Sci. Rep. 14 (1), 4103. http://dx.doi.org/10.1038/s41598-024-53066-4.

Yonggang, L., Weisenberg, R.H., 2011. A review of self-organizing map applications in meteorology and oceanography. In: Josphat Igadwa, M. (Ed.), Self Organizing Maps. IntechOpen, Rijeka, p. Ch. 13. http://dx.doi.org/10.5772/13146.

Yuan, S., Tan, P.N., Cheruvelil, K.S., Collins, S.M., Soranno, P.A., 2015. Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity. In: 2015 IEEE International Conference on Data Science and Advanced Analytics. DSAA, pp. 1–10. http://dx.doi.org/10.1109/DSAA.2015.7344878.

Zhang, Y., Shang, Q., Zhang, G., 2021. pyDRMetrics - A python toolkit for dimensionality reduction quality assessment. Heliyon 7 (2), e06199. http://dx.doi.org/10.1016/j.heliyon.2021.e06199.

Zhao, Q., Basher, Z., Costello, M.J., 2020a. Mapping near surface global marine ecosystems through cluster analysis of environmental data. Ecol. Res. 35 (2), 327–342. http://dx.doi.org/10.1111/1440-1703.12060.

Zhao, Q., Fränti, P., 2014. WB-index: A sum-of-squares based index for cluster validity. Data Knowl. Eng. 92, 77–89. http://dx.doi.org/10.1016/j.datak.2014.07.008.

Zhao, Q., Stephenson, F., Lundquist, C., Kaschner, K., Jayathilake, D., Costello, M.J., 2020b. Where Marine Protected Areas would best represent 30% of ocean biodiversity. Biol. Cons. 244, 108536. http://dx.doi.org/10.1016/j.biocon.2020.108536.

Zika, J.D., Gregory, J.M., Mcdonagh, E.L., Marzocchi, A., Clément, L., 2021. Recent water mass changes reveal mechanisms of ocean warming. J. Clim. 34 (9), 3461–3479. http://dx.doi.org/10.1175/jcli-d-20-0355.1.