Elucidating Ecological Complexity: Unsupervised Learning determines global marine eco-provinces

Maike Sonnewald[1,2,*], **Stephanie Dutkiewicz**[1], **Christopher Hill**[1,+], **and Gael Forget**[1,+]

[2]Harvard University, Department of Earth and Planetary Sciences, Cambridge, MA 02138, USA
[1]Massachusetts Institute of Technology, Department of Earth, Atmospheric and Planetary Sciences, Cambridge, MA 02139, USA
[*]sonnewald@fas.harvard.edu
[+]these authors contributed equally to this work

ABSTRACT

An unsupervised learning framework is presented for determining marine ecosystem provinces (eco-provinces). Increasingly under anthropogenic pressure, this represents a crucial step towards understanding and monitoring marine ecosystems. Functional group summation and t-SNE reduce the dimentionality from 55 to 3D. Unsupervised learning (DBSCAN) identifies clusters, unconstrained by reliance on Voroni cells (e.g. k-means). Intended for regional studies, the 127$\pm$7 eco-provinces reveal interactions in the humanly-intractible 55D data. Reducing complexity for global applications, aggregated eco-provinces (AEPs) are determined using graph theory on eco-province ecological similarity, determining a minimal complexity of 12 AEPs. Emergent eco-provinces and AEPs are unique and interpretable, revealing environmental and physical controlling mechanisms. AEPs can refined ecosystem monitoring with recognition of regions of similar biomass but different ecological composition, and improve biomass as a predictor of higher trophic levels (e.g. fisheries). Lastly, the AEPs give context to long term monitoring and can help plan large scale monitoring efforts.

Introduction

Organizing complex biogeography into coherent and meaningful regions is important for comparing and contrasting parts of the ocean, for charaterizing observations, and for monitoring and conservations efforts . Determining such provinces is key for facilitating prediction and monitoring, as they give context for assessing changes and understanding unique regional dynamics leading to of e.g. carbon sequestration and storage as well as food-chain variability. However, work on objective determination of eco-provinces is hampered by the intractably complicated interactions, and to date provinces do not incorporate ecology. As a result bulk measurements of chlorophyll are used that mask the underlying complexity. Machine learning (ML) applications are ideally suited to objectively determine important interactions. The present paper determines which ecologicale features co-vary in a significant and unique way within even a high-dimensional feature space. The therm unique signifies its separation for other regions. The resultant insight can further be leveraged combining well ecological dissimilarity concepts, with the inherent complexity being intuitively familiar to fields such as graph theory. For useful classification, a framework needs to allow for both 1) global classification, and 2) a multiscale analysis that can be both spatially and temporally nested (Spalding *et al.*, 2007). In this study we show how defined provinces can help us understand some of the controllers of community structure. The framework can also provide insight for monitoring strategies, and be a vital tool for tracking ecosystem changes.

Terrestrial provinces are often classified according to similarity in climate (precipitation, temperature), soil, vegetation, and fauna, aiding management, biodiversity studies, and disease control. The oceans' provinces are more difficult to define, as the majority of organisms are microscopic, and the boundaries are fluid. In situ observations are sparse, and satellite measurements only capture the sea surface. Nevertheless, there have been a number of biogeographical studies and divisions of the oceans into provinces. Longhurst (Longhurst et al., 1995; 1998) provided the first global classification of marine provinces based on environmental conditions of mixing rates, stratification and irradiance, along with expert knowledge of other key conditions important to the phytoplankton at the base on the marine food-chain. The 56 Longhurst provinces have been widely used to e.g. assess carbon fluxes, aid fisheries and are even in planning in situ observational databases. Defining provinces in a more rigorous way, methods from fuzzy logic to regional unsupervised clustering have been used with the goal of identifying meaningful structures that can identify provinces in available data. Studies such as SeaScapes (Kavanaugh et al, 2014) use self organizing maps making the data less noisy, and hierarchical (tree based) clustering to define provinces on the basis of regional satellite derived Chlorophyll, photosynthetically available radiation (PAR) and mixed layer depth

(MLD). SeaScapes has been used for biogeochemical applications (Kavanaugh *et al.*, 2018) as well as more recently for coastal management (http://www.marinebon.org/seascapes.html). Previous studies have also considered the biogeography of microscopic phytoplankton in terms of the most abundant type (e.g. Alvain et al, 2008), or presence/absence of a small subset of organisms.

The use of bulk properties in previous work is largely due to the sparsity of available ecological data. This study presents a novel framework that defines global provinces based on ecological composition: Regions with similar and co-existing types and abundances of organism. Our framework offers greater utility, as it uses the entire present ecological composition not available from sparse observations. This allows analysis of new surveys and increasingly available genomic data. Specifically, efforts such as Tara Oceans feed into databases such as the Ocean Gene Atlas (Villar *et al.*, 2018) but comprehensive spatial and temporal coverage remains elusive. The presented framework can provide context for the sparse observational data.

Marine microbial ecosystems are a product of complex physical, chemical and biological interactions. Biogeography and biodiversity have manifestations on the large scale (e.g., Barton et al., 2010; Dutkiewicz et al.,2009; 2011), mesoscale (e.g., Levy et al., 2014; 2015; Perruche et al., 2011), and submesoscale (Mahadevan, 2016). Ocean currents and mixing are key in setting aspects of this biogeography (Clayton et al., 2013; Levy et al., 2014; 2015), and in supplying nutrients that support the phytoplankton. In addition, grazing pressure acts to control communities. Highly complex patterns emerge from the combination of interactions, as found by recent studies using the TARA datasets (deVagas et al., 2015, Lima-Mendez et al., 2015). Despite the importance of phytoplankton ecology for e.g. climate and fisheries, key components setting large scale ocean ecosystem remains highly uncertain. Here a framework is presented that leverages novel unsupervised machine learning techniques to provide the necessary tools to identify the mechanisms that set the ecosystem structure at different location.

Studies using Tara Ocean have made use of ML tools to help define communities (e.g. Lima-Mendez et al., 2015), but have not yet had sufficient coverage to define distinct provinces. Numerical models have global coverage, and the transformative power of data science/ML techniques allow the overwhelmingly complicated oceanographic data to reveal robust structures in the covariance of the data. Efforts to determine robust patterns and signals is a challenge even is simple systems. This is because emergent complexity can appear simply complicated/intractable until the underlying principles giving rise to the patterns are determined. The high-dimensional data is very difficult for an observer to process manually, and the goal of the presented framework is to identify emergent patterns that *are* humanly-tractable, offering immediate ecological insight.

Viewed naively, the often high-dimensional data does not allow even advanced machine learning to robustly determine meaningful and robust clusters that represent provinces. The presented framework leverages the complex nature of high-dimensional data. The distinction between complicated and complex is the later is determined by underlying simple rules. For example, to understand how fish and birds "swarm" in concerted unison, Reynolds (1987) defined three rules based on separation, alignment and cohesion. The three rules allowed the emergence of complex patterns that significantly advanced fields from behavioral biology, video games and swarm robotics (Min *et al.*, 2011).

ML techniques can be used as "black box" methods. The purpose of the presented framework is to further ecological understanding, which is achieved by returning to the original high-dimensional data after the eco-provinces and AEPs are determined. The respective geographical regions allow understanding of the key ecological variables that led to ML classification as a unique cluster. Following Sonnewald *et al.* (2019), the purpose here is to present an agile framework that leverages unsupervised learning as a tool to define distinct eco-provinces.

The present framework defines eco-provinces using output from an advanced global 3D physical/ecosystem model (DARWIN). The ecosystem is sufficiently complex (encompassing 35 phytoplankton and 16 zooplankton types) that simple diagnostics are not capable of determining coherent patterns in community structure. The method devised in this study (Fig. 1a) first reduces the dimensionality of the problem from 55 to 11 dimensions (7 plankton functional groups, 4 nutrient supply rates). Using cutting edge data science methods, these are projected into a probabilistic 3D space (t-SNE). Unsupervised clustering identifies regions of close ecological proximity (DBSCAN). The resulting eco-provinces are then back-projected onto the globe. However, the over 100 identified eco-provinces can be overwhelming. Utility is increased by robustly nesting and aggregating eco-provinces (AEPs, Fig 1b) down to the desired complexity. The minimum number complexity of robust AEPs is determined. These focus here is on the potentially groundbreaking methods and exploring the minimal complexity AEPs case to determine controls on the complex emergent community structures.

Model framework: DARWIN

The purpose of this study is to develop a framework for defining eco-provinces by the distinct plankton community structure.It is too early to attempt this undertaking with observational data. And as such in this study we use a complex physical/biogeochemical/ecosystem model of the global ocean. We briefly describe the model here; details can be found in the references mentioned below.

The physical component of the model comes from the Estimating the Circulation and Climate of the Ocean (ECCOv4) global state estimate described by Forget *et al.* (2015); Wunsch and Heimbach (2013), see also ECCO Consortium (2017a,b); Forget *et al.* (2018); Wunsch and Heimbach (2007, 2013); **?**. The state estimate has a nominally 1° resolution. A least-squares with Lagrange multipliers approach is used to obtain observationally adjusted initial and boundary conditions as well as internal model parameters, resulting in a *free-running* version of the MIT General Circulation Model (MITgcm, Adcroft *et al.* (2004)). The physical fields are available online (ecco.jpl.nasa.gov) and documentation described at Forget *et al.* (2018) and http://doi.org/10.5281/zenodo.2533351.

The biogeochemical/ecosystem model captures the cycling of C, N, P, Si and Fe through inorganic and organic pools. The ecosystem model is based on Dutkiewicz et al (2015) with 35 phytoplankton (2 pico-prokaryotes, 2 pico-eukaryotes, 5 coccolithophores, 5 diazotrophs, 11 diatoms, 10 mixotrophic dinoflagellates) and 16 zooplankton spanning from 0.6m to 2500 m equivalent spherical diameter. Parameters influencing phytoplankton growth, grazing, and sinking are related to size (following Ward et al., 2012) with specific differences between the 6 functional groups. Distribution of bulk properties such as Chl, phytoplankton biomass, nutrient concentrations, as well as distributions of size classes and functional groups compare well with satellite and in situ observations. Results from this 51 plankton component of the model has been used in several recent studies (e.g. Treguer et al, 2018, McParlen and Levine, 2019, and Kuhn et al submitted), though in a different physical framework.

The coupled physical/biogeochemical/ecosystem model was run for 20 years from 1992-2011. Output from the model includes the plankton biomass, nutrient concentrations, and rate of supply of the nutrients (DIN, PO4, Si, Fe). For this study we use the 20 year mean of these output as the input for the eco-provinces.

Unsupervised Machine learning: Identifying and aggregating eco-provinces

0.1 Dimensionality reduction with t-SNE

By identifying the eco-provinces, this work provides a framework for identifying important relationships between key features in biomass of species and nutrient fluxes. Obtaining robust, reproducible, provinces is not possible with the complex 55D data using learning methods based on euclidean distances as Maze *et al.* (2016); Sonnewald *et al.* (2019) employing K-means. This is because the topology of key features that define the eco-provinces *do not* inhabit shapes that are Gaussian. K-means, using Voronoi cells (straight lines), effectively partitions the space at random. Like drawing a map of a complex mountainous region, careful thought is needed so the key features remain intact, but unnecessary detail discarded. The dimensionality of the data was initially reduced by including only the sum of biomass of each of the seven functional group, and all source terms for the flux of the nutrients (Nitrogen, Iron, Phosphate and Silica). The summing into functional groups reduces the problem from a 55D space to 11D. The nutrient supply rates were included following earlier studies showing their key roles in setting community structure (e.g. Ward et al., 2012; 2013; Dutkiewicz et al 2012). The biomass and nutrient fluxes form the 11D vector $\mathbf{x}$. If $\mathbf{x}$ is a vector field on the model grid discretized sphere, where each element $\mathbf{x}_i$ represents an 11D vector on the model's horizontal grid, each index i uniquely identifies a grid point on the sphere, with (lon,lat) = (ϕ_i, θ_i). The components of the vector $\mathbf{x}_i$ are the seven functional groups and 4 nutrient fluxes. The log of the biomass data is used, and is discarded if a cell has a biomass less than $12 * 10^{-4} mgChl/m^3$or ice cover over 70%. The data is normalized and standardized such that all data exist on the same range [0 to 1]. This is done so the features (biomass and nutrient fluxes) do not get conditioned by contrasts in the ranges of possible values. The clustering should capture the variational relationships from the key probabilistic distances between features. Quantifying these distances, important features emerge while unnecessary detail is discarded. In ecological terms, this is necessary because species that may have comparatively small biomass can be key ecological players facilitating relations that only emerge when the data is normalized and standardized.

The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm is used to make existing similar regions stand out more clearly by emphasizing feature proximity in the high-dimensional space in a lower dimensional representation. Previous work aiming to build deep neural networks for remote sensing applications employed t-SNE, demonstrating its skill in separating key features (Marmanis *et al.*, 2016). This is a necessary step towards identifying robust clusters in the feature data, avoiding

non-convergent solutions (not shown). Using a Gaussian kernel, t-SNE preserves the statistical properties of the data by mapping each high-dimensional object onto a 3D point in a way that ensures a high probability of similar objects being close in both the high and low-dimensional space van der Maaten & Hinton (2008). Given a set of N high-dimensional objects x_1 ,..., x_N, the t-SNE algorithm performs a reduction by minimizing the Kullbach-Leibner (KL) divergence (Kullback, 1987) between the likelihood of association between a low dimensional rendition and the high dimensional features. If x_i is the i-th object and x_j the j-th object in the *N dimensional* space and, y_i is the i-th object and y_j is the j-th object in the *low-dimensional* space:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)},$$

and the same for a reduced dimensional set:

$$q_{i|j} = \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq i}(1+\|\mathbf{y}_j - \mathbf{y}_k\|^2)^{-1}}.$$

The KL divergence is:

$$\mathrm{KL}(P||Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

Fig. 2a illustrates the effect of reducing the variables of phytoplantion, zooplankton and nutrient fluxes. The motivation for applying t-SNE can be likened to that of principal component analysis (PCA); using variance attributes to emphasize regions/properties of the data and thus reduce the dimentionality. The t-SNE method was found to be much superior to PCA in delivering robust results for the eco-provinces (not shown). This is likely because the orthogonality assumption that underlies PCA is not appropriate for identifying key interactions between highly complicated and interacting features. Lunga *et al.* (2014) demonstrates the effect of several dimensionality reduction techniques, illustrating how complex spectral features in remote sensing data can be highlighted employing SNE in the context of manifold learning.

Clustering: Finding similar regions with DBSCAN

The points in the t-SNE scatter plot in Fig. 2a are each associated with a latitude and longitude. If two points are close to each other in Fig. 2a, this is because their biomass and nutrient fluxes are similar, not due to geographical proximity. Note that the topology/shape is arbitrary, as opposed to Gaussian methods such as k-means. The colors on Fig. 2a are the clusters found using the Density-based spatial clustering of applications with noise (DBSCAN) method proposed by Ester *et al.* (1996). Looking for densely packed observations, the DBSCAN algorithm uses the distance in the 3D representation between points (ε, here 0.39), and the number of similar points needed to define a cluster (here 100 points). The DBSCAN method makes no assumptions about the shapes or numbers of clusters in the data:

1. A random datapoint y_i is selected.
2. The number of immediately neighbouring points within distance ε of y_i is measured.
3. The cluster boundary is determined repeating step 2 iteratively for all points identified as within distance ε. If the number of points is larger than the set minimum it is designated as a cluster.
4. A new point is chosen at random from the remaining unclassified data, and the method repeated.

The data that does not meet the minimum cluster member and distance ε metric are counted as "noise", and not assigned a colour. DBSCAN is a fast and scalable algorithm, with a worst-case performance of $O(n^2)$. Setting the number of minimum points and the distance ε was done using the degree of connectednes (Fig. 6a) and how large a percentage of the global ocean is covered (Fig. 6b), discussed in the appendix.

Back-projecting onto the globe

The 115 clusters identified in Fig. 2a are presented in their geographical extent in Fig. 2b. The colour corresponds to a geographically coherent combination of biogeochemical factors identified by DBSCAN. Strikingly, globally coherent regions are identified in Fig. 2a. Once the clusters are determined, the association of each point in Fig. 2a to a specific latitude and longitude is used to project clusters back to the geographical domain. Fig. 2b illustrates this, with colours of clusters still the same as in Fig. 2a. Similar colours should not be interpreted as ecological similarity, as they are assigned by the order in which

the algorithm discovers clusters.

Familiar regions appear in Fig. 2b. The clusters in the Southern Ocean are zonally symmetric, the oligotrophic gyres emerge, sharp transitions suggest the influence of the trade winds, and distinct regions associated with upwelling are seen e.g. in the equatorial Pacific. Many of these regions correspond to the Longhurst provinces.

Ecological similarity: BC-dissimilarity

To understand the ecological context of the eco-provinces, the intra-cluster ecology is assessed using the Bray-Curtis Dissimilarity metric (BC Bray and Curtis (1957)) is applied to the 51 species of phyto- and zooplankton. BC is a *dissimilarity* metric defined as:

$$BC_{n_i n_j} = 1 - \frac{C_{n_i n_j}}{S_{n_i} + S_{n_j}},$$

where BC compares one species assemblage to another. $BC_{n_i n_j}$ refers to the dissimilarity of assemblage n_i compared to assemblage n_j, where the $C_{n_i n_j}$ is the minimum of biomass of individual species present in both assemblages n_i and n_j while S_{n_i} refers to the sum over all the biomass present in both assemblages n_i and S_{n_j}. The BC-dissimilarity is similar to a distance metric (e.g. Manhattan distance), but operates in a non-euclidean space for ease of ecological interpretation.

For each cluster identified in Fig. 2b, the intra- and inter-province BC-dissimilarity can be assessed. The intra-province BC-dissimilarity refers to the dissimilarity between the province mean and each point in it. The inter-province BC-dissimilarity refers to how similar one province is to each other province. Fig. 3a illustrates the symmetric BC matrix where 0 (black: perfect correspondence) and 1 (white: completely dissimilar). Each line in this plot demonstrates patterns in the data. Fig. 3b demonstrates the geographical implications of the BC results from Fig. 3a for individual provinces. For a province in the low nutrient oligotrophic region, Fig. 3b demonstrates that large areas are reasonably similar symmetrically around the equator and in the Indian Ocean, but the higher latitudes are markedly different along with upwelling areas.

The intra-province BC-dissimilarity within each province from Fig. 2b is illustrated in Fig. 4a. Determined using the mean area averaged assemblage within one cluster, and determining the BC-dissimilarity of each gridpoint within the province to the mean, it illustrates how well the machine learning framework is able to separate the 51 species of the model data according to ecological similarity. The global mean intra-cluster BC-dissimilarity is 0.102 ± 0.0049. Using the sum of the 7 functional groups mean intra-seasonal BC-dissimilarity instead gives 0.105 ± 0.004. This suggests the presented framework, based on cell size, is appropriate for the high dimensional case although it was trained on the biomass sum of the 7 functional groups. However, if the temperature-types in the model were used the utility could be different.

The Longhurst intra-province BC-dissimilarity is presented in Fig. 4b using the biomass of the 51 species, with a global mean of 0.227. This is significantly larger than for the clusters identified in Fig. 1b. Using the sum of the 7 functional groups mean intra-seasonal BC-dissimilarity of the Longhurst provinces increases to 0.232. The maps of the global provinces in Fig. 2 offer intricate detail of ecological interactions that are unique. Regional studies that target particular areas of the ocean can use the presented provinces to assess the ecological context of the area of interest.

Leveraging chaos: Defining Aggregated Eco-Provinces (AEPs)

While the provinces presented in Fig. 2a offer a considerable refinement of the Longhurst provinces, dealing with over hundred provinces can become overwhelming. One of the uses of provinces is to facilitate understanding of where they are and how they are governed. To identify emergent properties the method in Fig. 1a is developed further to allow a nesting of ecologically similar provinces. An adjustable level of "complexity" is set as the number of provinces that will be considered. For defining meaningful aggregation, the intra-province BC-dissimilarity from the Longhurst provinces of 0.227 is used as a benchmark.

The eco-provinces are coherent across the globe as Fig. 3b demonstrates. Some configurations are very "common" as seem using the inter-province BC-dissimilarity. Inspired by methods from genetics and graph theory such as "connectivity graphs" we can sort the > 100 provinces according to which province they are most similar to as measured by the inter-province BC-dissimilarity (Costanzo *et al.*, 2016; Diestel *et al.*, 2005). For a chosen number AEPs, the P most dominant/highly connected eco-provinces are used to aggregate the remainder of the eco-provinces; each eco-province is assigned to which of the P most dominant/highly connected eco-province they are most similar to. This aggregation determined by the BC-dissimilarity allows

a nested approach to global ecology.

The chosen P can be anything from 1 to the full complexity from Fig. 2a. However, because of the probabilistic dimentionality-reduction step (t-SNE), low complexities can be degenerate. Degeneracy implies that significant variations in the geographical area covered by the AEPs is possible. In addition, a benchmark is used based on improved performance over the intra-province BC-dissimilarity of the Longhurst provinces. Fig. 4c illustrates the spread of the intra-province BC-dissimilarity in the AEPs of increasing complexity across ten realizations (Illustration in Fig. 1b). In Fig. 4c the 2σ (blue area) is a measure of the degeneracy within the ten realizations, and the green line represents the Longhurst benchmark. A complexity of 12 is demonstrated to keep the intra-province BC-dissimilarity both below the Longhurst benchmark in all realizations, and a relatively small 2σ/ degeneracy. In sum, the minimum recommended complexity is 12 AEPs, for which the mean intra-seasonal BC-dissimilarity assessed using the 51 species is 0.198 $\pm$0.013, as seen in Fig. 4d. Using the sum of the 7 functional groups mean intra-seasonal BC-dissimilarity 2σ changes 0.198 $\pm$0.004. This similarity suggests the presented framework is appropriate for the 55D case although it was trained on the biomass sum of the 7 functional groups.

Utility of Aggregated Eco-Provinces: Community structure and their controls

The minimum complexity 12 AEPs, are useful to explore the controls on the emergent community structure. With the knowledge of the geographical extent of these 12 AEPs, the overwhelming complexity in the 55D DARWIN data is revisited to gain insight into the present ecology. The familiar province distinctions of biomass rich upwelling, picoplankton dominated oligotrophic gyres and diatom rich polar regions is apparent. Fig. 5 illustrates the ecological insights grouped by AEPs (names A to L): The geographical extent (Fig. 5c), functional group biomass composition (Fig. 5a) and nutrient supply (Fig. 5b) scaled by N in the stoichiometric Redfield ratio (N:Si:P:Fe, 1:1:16:16$\times 10^3$), meaning that P is multiplied by 16 and Fe by 16×10^3 so the bars are comparable and variability visible.

The identified AEPs are all unique. There is some symmetry around the equator in the Atlantic and Pacific ocean, and similar, but augmented regions exist in the Indian ocean. Coherence with global physical regimes as presented in Sonnewald *et al.* (2019) are seen in features such as the Western Boundary Currents (WBCs) seen in AEPs hugging the western sides of continents, the Antarctic Circumpolar Current (ACC) is seen as a large zonal feature. Known upwelling regions are seen on the Eastern side of the ocean basins such as the California upwelling and the upwelling in the equatorial Pacific. The subtropical gyres stand out as complex series of oligotrophic AEPs. The subtropical North Atlantic is interesting as the Fe rich dust input from the Sahara cuts across the otherwise Sverdrupian gyre.

AEPs with very similar phytoplankton biomass can have very different community structure (e.g. D, H and K), and cover very different geographical areas. AEP H is present mainly in the equatorial Indian ocean and has a larger population of diazotrophs. AEP D is found in several basins, but is prominent in the Pacific surrounding the very highly productive region around the Equatorial upwelling. The shape of this province in the Pacific is reminiscent of planetary wavetrains. AEP D has very few diazotrophs but more cocolithophores. AEP K is found only in the high Arctic ocean, and is dominated by diatoms. The nutrient levels in the D, H, and K province are somewhat similar, with AEP K having more Si, and AEP H and D having very little. There is less N compared to P in AEP H. If only Chlorophyll were used to define AEPs then D, H and K could not be distinguished. It is notable that the zooplankton biomass in the three regions are very different, with AEP K having very little, but AEP D and H having relatively similar levels. The phytoplankton biomass is similar, but is not sufficient to predict the zooplankton, in this manner using just Chlorophyll to define provinces would not capture this.

It is apparent that some AEP that have very different biomass are very similar in terms of their ecological community structure. This is seen in AEP D and E for example. These are close to each other, notably in the Pacific, where AEP E is close to the highly productive AEP J. There is again not a clear connection between phytoplankton biomass and zooplankton abundance. In this manner, monitoring Chlorophyll is again not a sufficient predictor of zooplankton biomass and does not translate to knowledge of higher trophic levels.

Diatoms only exist where there is silica supply; generally the higher the silica the higher the diatom biomass. However, the proportion of diatom biomass relative to other phytoplankton is dictated by how much more N, P, Fe are supplied, relative to the diatoms demands. This is because diatoms are the fastest growers, but are limited by silicon supply. If this limitation was not present, diatoms would dominate in all but the lowest nutrient supply regions. Diatoms are seen in the AEPs A, J, K, and L together with silicon supply. As expected, diatoms are found in the polar regions. AEP L spans the high latitude Southern Ocean and AEP K is a region in the Arctic Ocean. The very productive AEP J has diatoms, and AEP A can be interpreted as an exten-

sion of AEP J outside of the Equatorial regions. Province L has the most even supply of nutrients relative to the Si supply: here diatoms dominate. In province J, there is relatively more N,P,Fe than Si, so diatoms are only a small fraction of the total biomass.

Diazotrophs have the ability to fix N, but grow slowly. They coexist with other phytoplankton where there is excess of Fe and P relative the the demands of the non-diazotrophs. It is notable that there is higher diazotroph biomass where the amount of Fe and P supply are relatively large. In this manner, the diazotroph biomass is larger than AEP H than in J, although the overall biomass in AEP J is larger. It is worth noting that AEP J and H are very different, with H located in the equatorial Indian Ocean.

The insight gained from patterns in the minimum complexity of 12 AEPs would be much less clear if the biomass data were not separated into provinces. The AEPs facilitate the coherent and simultaneous comparison of the plethora of global maps from DARWIN in Fig. 5. The AEPs effectively highlight why and where chlorophyll is not a good proxy for productivity. This information can in future work be used to improve uses of chlorophyll as a bulk measure, by filling in the relevant ecological information on the basis of knowledge gained from the AEPs. In this manner, the AEPs can be leveraged to make chlorophyll a better proxy for biomass in higher trophic levels. Fig 5a illustrates that zooplankton biomass can not be inferred from total phytoplankton biomass,as it is a product of the AEP species assemblage. A detailed analysis is the topic of an ongoing study beyond scope of this paper. However the technique presented here provides a way to explore other mechanisms in a more tractable way than looking from point to point.

Discussion and Conclusion

A framework for decoding the overwhelmingly complicated ecological data from DARWIN is presented. Global eco-provinces are determined by summation of biomass across functional groups and application of the t-SNE probabilistic dimensionality reduction. The unsupervised ML method DBSCAN is applied to determine the eco-provinces, and an inter-province BC-dissimilarity/graph theoretic method for nesting is applied to arrive at the AEPs. Both the eco-provinces and AEPs are unique by construction. The nesting can be adjusted between the full complexity of the original provinces and a minimum threshold of 12 AEPs set using the intra-province BC-dissimilarity of the Longhurst provinces as a benchmark. The nesting and determination of minimal complexity of the eco-provinces into AEPs is seen as a crucial step, as the probabilistic t-SNE makes the <12 complexity AEPs degenerate. The presented nesting method delivers robust AEPs, as well as eco-provinces with significantly higher ecological utility than the Longhurst provinces. The framework is global, and can span a range of complexity from >100 AEPs to 12. Here, the 12 global AEPs were mostly discussed. However, regional studies could focus on a smaller subset of the global map, and perform the nesting within this smaller region, easily leveraging the same ecological insight.

The eco-provinces and AEPs can be used to identify governing mechanisms. Assessing the 12 AEPs, the presence of similar biomass but significantly different ecological composition highlights regions where using chlorophyll as a bulk measure is not a sufficient characterization of ecology (e.g. D and E). In contrast, AEPs such as D and K have very different biomass but similar ecological composition. With the knowledge of the relationship between AEP composition a relationship to determine the associated zooplankton biomass could be determined, as seen in e.g. AEP D+E or K+L. This has implications for using satellite based chlorophyll measurements to assess the base of the food-chain for fisheries, allowing more accurate resource management. This paves the way for satellite based monitoring of the health of the ecosystem and forcasting danger of throphic cascades in fisheries/the food-chain (Sommer, 2008).

The presented framework provides a convenient way to assess the mechanisms that control the features in the provinces (e.g. biomass/chlorophyll, NPP and community structure). For example, the relative amount of diatoms is set by the imbalance in the Si to N,P, Fe supplies. With balanced supply rates, the community is diatom dominated (L) and where they are less balanced diatoms comprise only a smaller fraction (K). The diazotrophs survive where the Fe and P supplies are in excess of the N supplies (e.g. E and H).

The present framework can critically inform how geographically general samples from one location may be, allowing more confidence in comparisons between different geographical regions. Comparisons with less complicated numerical models could also be facilitated through knowing how general an in situ sample is likely to be. Long term observational datasets will continue to be invaluable, and the presented framework can be seen as a map of regions that are key to sample, as well as a means to retrospectively assess how generally informative the region is. For example, the timeseries from 137°E and BATS(Fig. 5c, nr 1 and 4) are both in quire oligotrophic regions, but 137°E is further from the boundaries while BATS is just at the border between somewhat similar AEPs (C and F). ALOHA (Fig. 5c, nr 2) is located quite close a boundary of a very different AEP, suggesting that it should not be considered representative of the entire AEP it belongs to. The P Station (Fig. 5c, nr 3) is quite far from the

boundaries, and could therefore be more representative. The provinces and AEPs identified could help establish a monitoring framework suitable for assessing global change, as the provinces allow assessment of where in-situ sampling should be done. The framework can be repeated for climatological data to assess temporal variability (not shown).

The success of the framework is achieved through careful application of data scinece/ML methods, together with domain specific knowledge. Specifically, the dimensionality reduction step allows the high-dimensional data to be visualized. The data is arranged in streaks and sheats of covariance, clearly indicating that purely distance based metrics such as k-means are inappropriate as they often assume a Gaussian (round) distribution. The DBSCAN method is appropriate for arbitrary topologies, provided careful attention is given to setting the parameters. The regions that fail to be identified as within a province by the developed method can be seen as the remaining black dots in Fig. 2. These are seen to be highly seasonal, as they are arranged as "streaks" in the 3D space, or associated with very few datapoints, likely failing the "minimum" criterion of DBSCAN. Geographically, these regions are in highly seasonal areas. This suggests that capturing the time evolving eco-provinces would provide better coverage.

To construct the present framework, ideas from complex systems/data science have been leveraged. We exploit the ability to determine aggregations of "events" (high probability of close proximity in a 55D space), and determine "provinces". These provinces, present throughout the ocean, describe a specific volume in our 3D phase space. In a manner similar to the Henon and Heilies (1964) system, where Poincaré section are used to reduce the dimensionality and the volume occupied demine "regular" or "chaotic" areas, we organize the 55D DARWIN data in 3D space using t-SNE and find provinces that are related by the minimization of the Kullbach-Leibner distance. The relation between geographical area and the area in 3D space is not simple. The relation between the volume in 3D space can be interpreted in terms of ecological similarity, although it is not straightforward to do so. The more conventional BC-dissimilarity metric was preferred for this reason. The BDSCAN algorithm scales well in most implementations. The t-SNE algorithm scales well in theory as the Kullbach-Leibmer distance is readily parallelisable (van der Maaten, 2014).

Future work will repeat the presented framework for climatological data, to assess the spatial variability in the identified provinces and AEPs. A future goal of the AEPs is to develop methodology that allows detection of ecology using the traditionally used bulk properties (chlorophyll, MLD and PAR). This would allow remote sensing assessments of ecological composition and highly agile monitoring of the eco-provinces and their variability.

Acknowledgements

This work was supported by grant NASA-IDS (80NSSC17K0561), ECCO Consortium funding via the Jet Propulsion Laboratory.

Author contributions statement

M. Sonnewald conceived the experiment(s), developed the method, ran the analysis and wrote the main body of the text. S. Dutkiewicz contributed with ecological expertise, biochemical model development and writing. C. Hill contributed to discussions and reviewed the manuscript. G. Forget ran the model and reviewed the manuscript.

References

: This list of references is not complete.

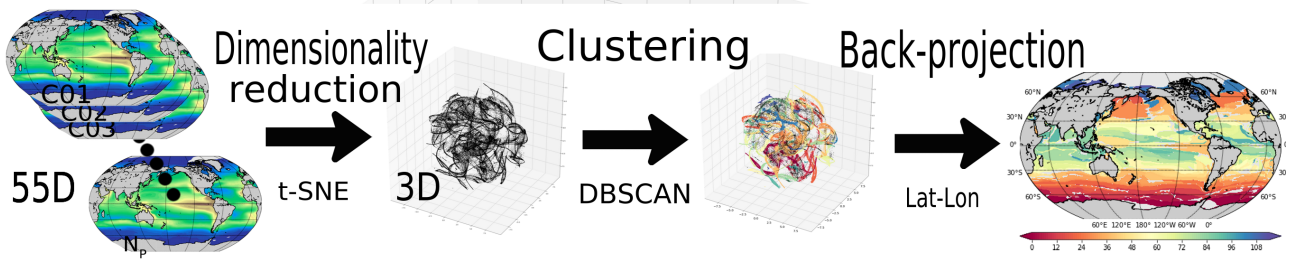
Adcroft, A., Hill, C., Campin, J. M., Marshall, J., & Heimbach, P. (2004). Overview of the formulation and numerics of the MIT GCM (pp. 139-150). Presented at the ECMWF Conference Proceedings, Shinfield Park, Reading, UK.

Bray, J. R. and J. T. Curtis. 1957. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27:325-349.

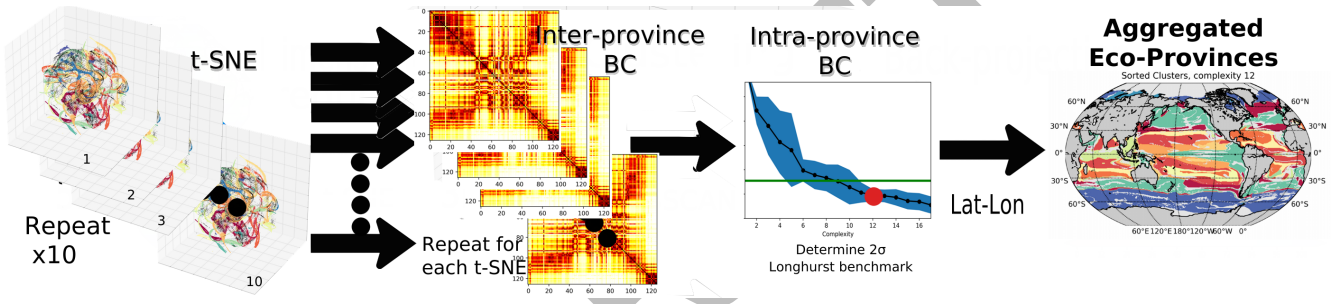
Costanzo M. , Van der Sluis B. , Koch E.N. , Baryshnikova A. , Pons C. , Tan G. , Wang W. , Usaj M. , Hanchard J. , Lee S.D. et al. . A global geneticA global genetic interaction network maps a wiring diagram of cellular function. Science . 2016; 353:aaf1420.

Diestel, R. (2005). "Graph Theory, Electronic Edition". p. 12.

ECCO Consortium, 2017a, A Twenty-Year Dynamical Oceanic Climatology: 1994-2013. Part 1: Active Scalar Fields: Temperature, Salinity, Dynamic Topography, Mixed-Layer Depth, Bottom Pressure.

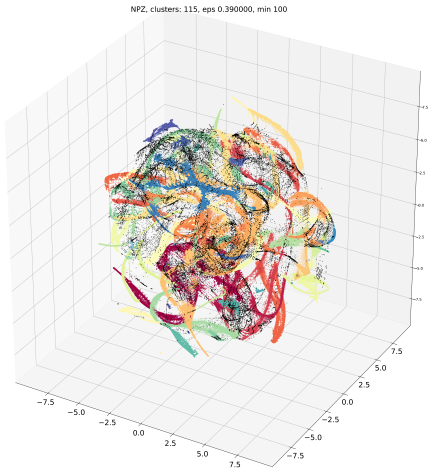


(a) Work flow repeated ten times

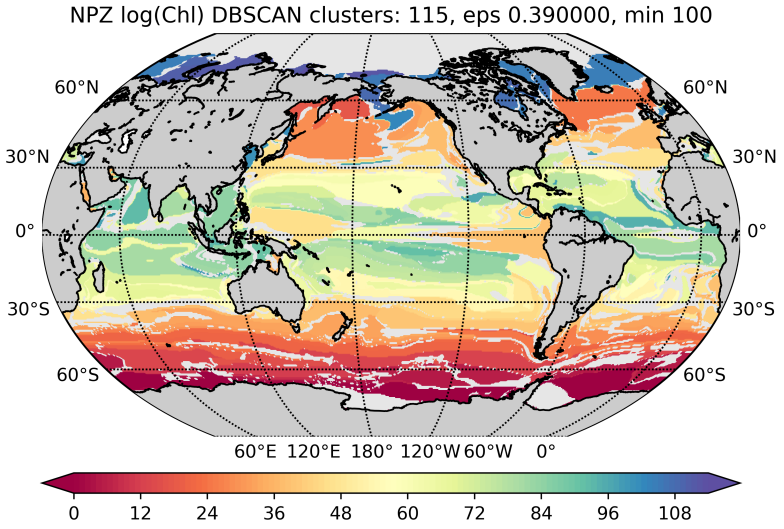


(b) Determining the Aggregated Eco-Provinces (AEPs)

Figure 1. Upper panel is Fig. 1a is a sketch of the work-flow to determine the provinces; The raw 55D model data has sum of biomass of functional groups taken. Negligible values and persistent ice cover areas are discarded. Data is normalized and standardized. The 11D data is given to the t-SNE algorithm to highlight statistically similar feature combinations. DBSCAN selects the clusters carefully setting parameter values. The data is finally projected back onto a lat-long projection. Note this is repeater 10 times as a slight stochastic element is possible through the application of t-SNE. Lower panel Fig. 1b illustrates how the AEPs are arrived at by repeating the work-flow in Fig. 1a ten times. For each of the ten realisations, the inter province BC-dissimilarity matrix is determined based on the 51 species. The BC-dissimilarity within the aggregated provinces is determines increasingly allowing more provinces until a benchmark set by Longhurst is reached.

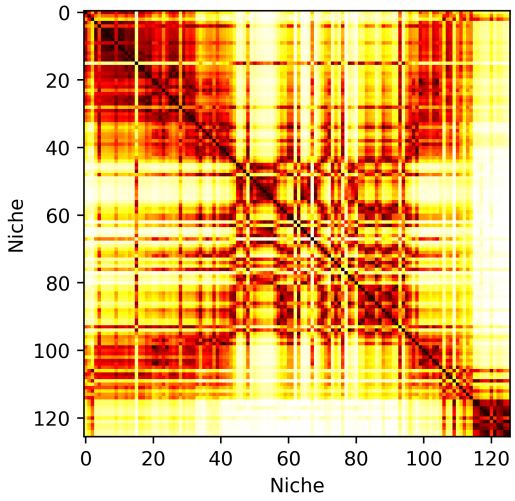


(a) t-SNE projection with provinces in colour

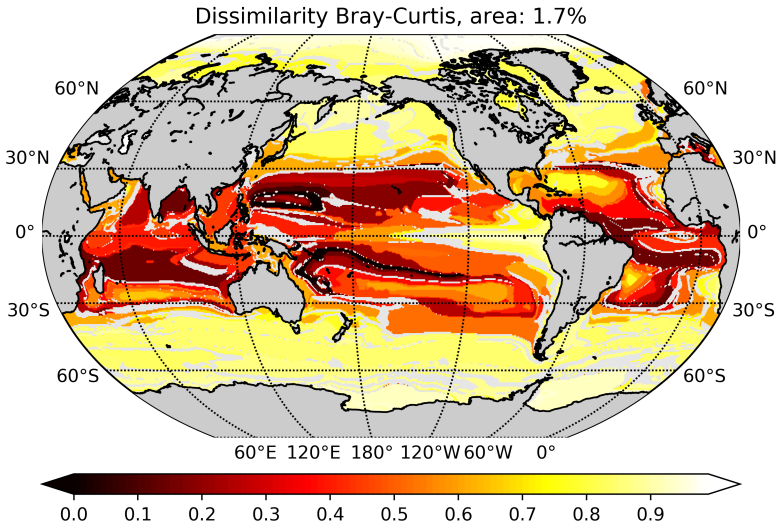


(b) Spatial representation provinces in Fig. 2a BC-dissimilarity

Figure 2. The left Fig. 2a showing nutrient, phytoplankton and zooplankton data as rendered by the t-SNE algorithm, and colored by province using DBSCAN. Each point represents one point in the high dimensional space, with the vast majority of points captured as is demonstrated in the appendix. Axes refer to the "t-SNE" dimensions 1, 2 and 3. The right Fig. 2b shows the geographical projection of the provinces discovered by DBSCAN onto the origin lat-lon grid. Colours should be considered arbitrary but correspond to Fig. 2a.

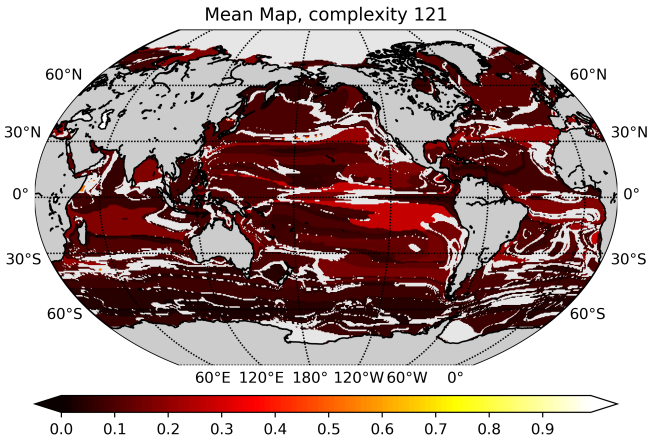


(a) Symmetric BC matrix

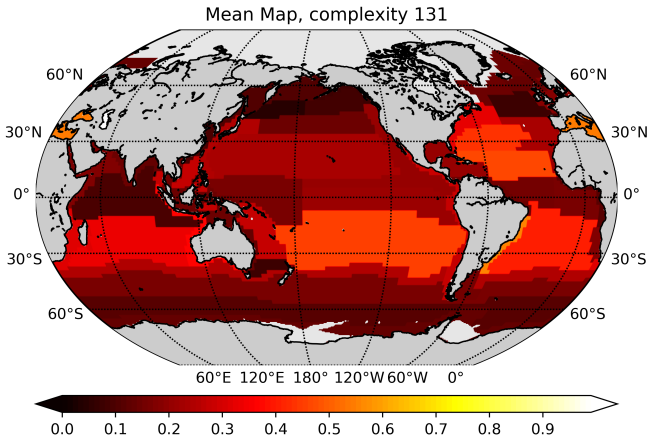


(b) Spatial representation of one column in Fig. 3a BC-dissimilarity

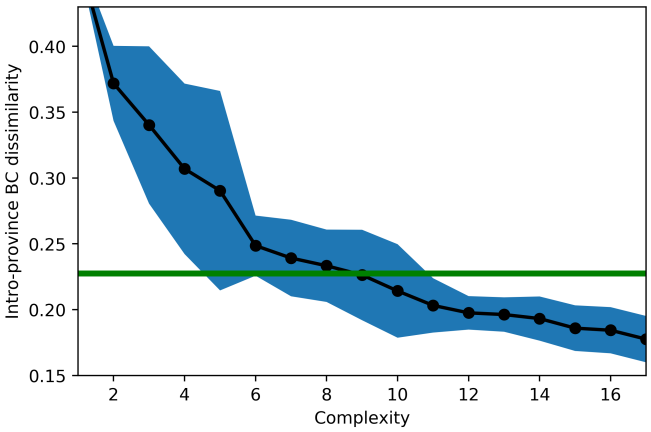
Figure 3. The left Fig. 3a Bray-Curtis Dissimilarity metric evaluated for **every province compared to every other** for the global surface 20 year mean. Note the expected symmetry of the values. The spatial projection of one column (or row) is illustrated in the right Fig. 3b. The global distribution of Bray-Curtis Dissimilarity metric evaluated for one province compared to every other for the global surface 20 year mean. Note the example shows an oligotrophic gyre example where other basins have similar regions and there is some symmetry across the equator. Black (Bray-Curtis = 0) denotes an identical region, while white (Bray-Curtis = 1) denotes no similarity.



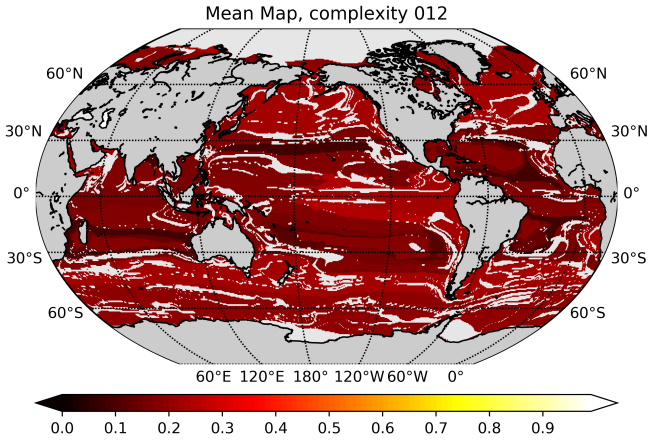
(a) Full complexity intra-province BC-dissimilarity



(b) Longhurst intra-province BC-dissimilarity

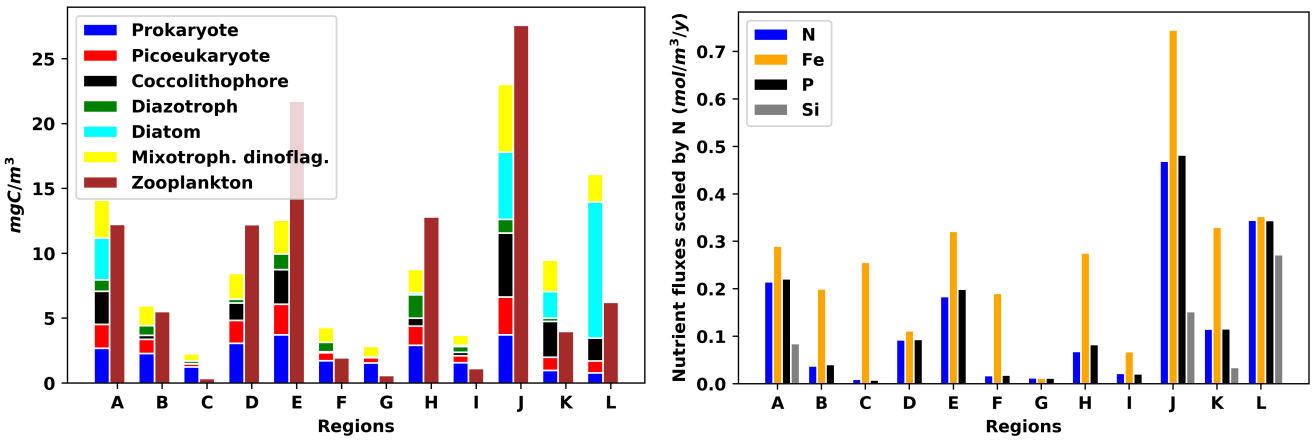


(c) Longhurst criterion



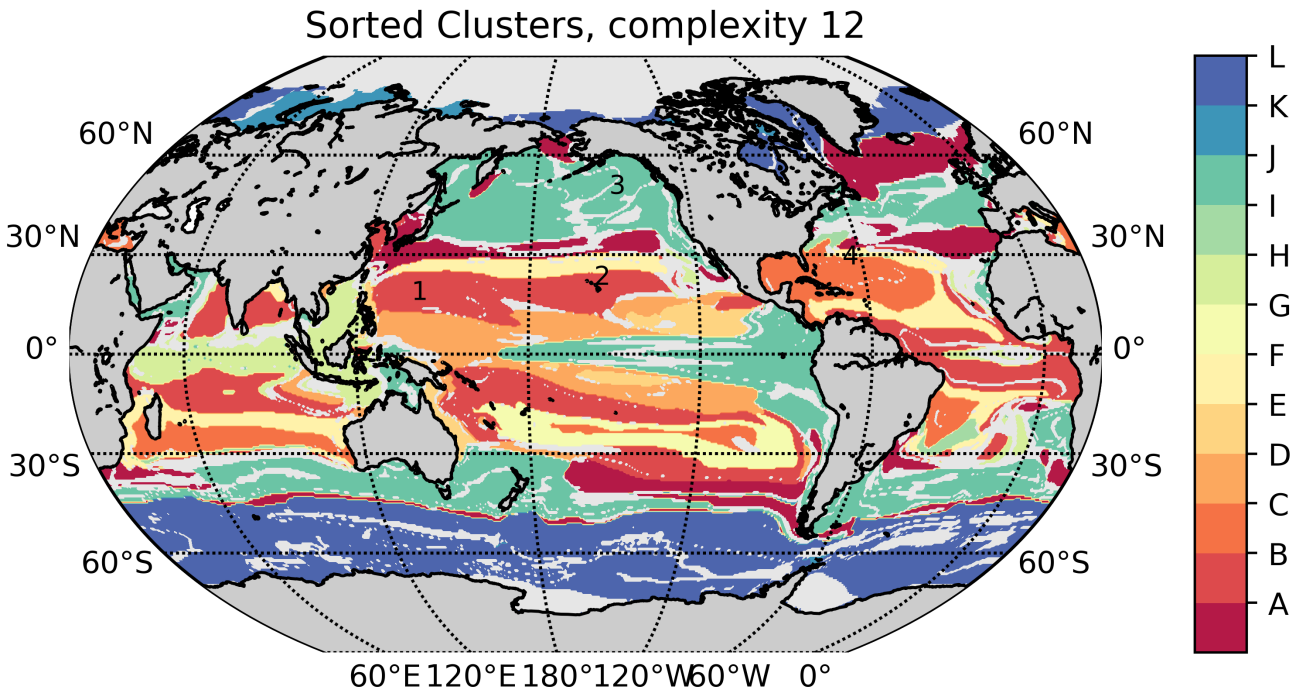
(d) Complexity 12 intra-province BC-dissimilarity

Figure 4. How complex should the biogeochemistry be? The top left Fig. 4c shows the intra-province BC-dissimilarity of the Longhurst provinces (green line). The black line illustrates the intra-province BC-dissimilarity of increasing complexity. The 2σ is from 10 repeats of the eco-province recognition process. For Fig. 4b, 4d and 4a the intra-province BC-dissimilarity is assessed as the mean BC dissimilarity of the individual gridpoint communities compared to the mean province with no reduction in complexity. For Fig. 4b, the global mean intra-province BC-dissimilarity is 0.227. This is the benchmark for the ecologically motivated sorting presented in this work (green line in Fig. 4c). For the full complexity in the provinces discovered by DBSCAN, Fig. 4a illustrates that an intra-province BC-dissimilarity of 0.099 is reached, while sorting into a complexity of 12 as suggested by Fig. 4c gives an intra-province BC-dissimilarity of 0.200 is reached as demonstrated in Fig. 4d.



(a) Ecological ensemble

(b) Nutrient fluxes



(c) Global provinces

Figure 5. Sorting the provinces into the 12 most dominant provinces named from A to L. Top left Fig. 5a showing ecological ensemble in the 12 provinces. Top right Fig. 5b the nutrient fluxes. Bottom panel Fig. 5c. Note the distinction between Polar, subtropical gyres and dominantly seasonal/upwelling regions in the bottom panel. Monitoring stations marked are 1: 137°E, 2: ALOHA, 3: Station P, and 3: BATS.

ECCO Consortium, 2017b, A Twenty-Year Dynamical Oceanic Climatology: 1994-2013. Part 2: Velocities, Property Transports, Meteorological Variables, Mixing Coefficients.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

Flanders Marine Institute (2009). Longhurst Provinces. Available online at http://www.marineregions.org/. Consulted on 2019-03-15.

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M Ponte, and C. Wunsch, ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation, Geo. Sci. Model Dev., 8, 2015

Forget, G., Ferreira, D. and Liang, X. (2015) On the observability of turbulent transport rates by Argo: supporting evidence from an inversion experiment. Ocean Science, 11 (5). pp. 839-853. https://doi.org/10.5194/os-11-839-2015

Forget, G. (2018). gaelforget/ECCOv4: Documentation updates (Version v1.8). Zenodo. http://doi.org/10.5281/zenodo.1211363

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch, 2016: ECCO Version 4: Second Release, MIT D-space http://hdl.handle.net/1721.1/102062

Fukumori, I., P. Heimbach, R. M. Ponte, C. Wunsch, A dynamically-consistent ocean climatology. Bull. Am. Met. Soc., doi:10.1175/BAMS-D-17-0213.1, in press, 2018

Hénon, M.; Heiles, C. (1964). "The applicability of the third integral of motion: Some numerical experiments". The Astronomical Journal. 69: 73–79. Bibcode:1964AJ.....69...73H. doi:10.1086/109234.

Kavanaugh, Maria J. Church, Matthew O. Davis, Curtiss M. Karl, David Letelier, Ricardo & Doney, Scott. (2018). ALOHA From the Edge: Reconciling Three Decades of in Situ Eulerian Observations and Geographic Variability in the North Pacific Subtropical Gyre. Frontiers in Marine Science. 5. 130. 10.3389/fmars.2018.00130.

Kavanaugh, M.T., B. Hales, M. Saraceno, Y.H. Spitz, A.E. White, and R.M. Letelier. 2014. Hierarchical and dynamic seascapes: A quantitative framework for scaling pelagic biogeochemistry and ecology. Progress in Oceanography 120:291–304, https://doi.org/10.1016/j.pocean.2013.10.013.

Kullback, S. (1987). "Letter to the Editor: The Kullback–Leibler distance". The American Statistician. 41 (4): 340–341. doi:10.1080/00031305.1987.10475510. JSTOR 2684769.

Longhurst, Ecological Geography of the Sea (ISBN 0124555217), Academic Press, 2007, San Diego, 560p.

Longhurst, A.R et al. (1995). An estimate of global primary production in the ocean from satellite radiometer data. J. Plankton Res. 17, 1245-1271

Longhurst, A.R. (1995). Seasonal cycles of pelagic production and consumption. Prog. Oceanogr. 36, 77-167

Longhurst, A.R. (1998). Ecological Geography of the Sea. Academic Press, San Diego. 397p. (IMIS)

D. Lunga, S. Prasad, M. M. Crawford and O. Ersoy, "Manifold-Learning-Based Feature Extraction for Classification of Hyperspectral Data: A Review of Advances in Manifold Learning," in IEEE Signal Processing Magazine, vol. 31, no. 1, pp. 55-66, Jan. 2014. doi: 10.1109/MSP.2013.2279894

D. Marmanis, M. Datcu, T. Esch and U. Stilla, "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks," in IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 1, pp. 105-109, Jan. 2016. doi: 10.1109/LGRS.2015.2499239

Maze, G., et al. Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. Prog. Oceanogr. (2017), http://dx.doi.org/10.1016/j.pocean.2016.12.008

Min, Hongkyu; Wang, Zhidong (2011). Design and analysis of Group Escape Behavior for distributed autonomous mobile robots. IEEE International Conference on Robotics and Automation (ICRA).

Reynolds, Craig (1987). Flocks, herds and schools: A distributed behavioral model. SIGGRAPH '87: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. Association for Computing Machinery. pp. 25–34. CiteSeerX 10.1.1.103.7187. doi:10.1145/37401.37406. ISBN978-0-89791-227-3.

Reynolds, R.W., D.B. Chelton, J. Roberts-Jones, M.J. Martin, D. Menemenlis, and C.J. Merchant, 2013: Objective Determination of Feature Resolution in Two Sea Surface Temperature Analyses. J. Climate, 26, 2514-2533,

Sommer, U. (2008), Trophic Cascades in Marine and Freshwater Plankton. International Review of Hydrobiology, 93: 506-516. doi:10.1002/iroh.200711039

Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions. Earth and Space Science, 6. https://doi.org/10.1029/2018EA000519

Spalding, Mark D., Helen E. Fox, Gerald R. Allen, Nick Davidson et al. "Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas". Bioscience Vol. 57 No. 7, July/August 2007, pp. 573–583.

van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE" (PDF). Journal of Machine Learning Research. 9: 2579-2605.

L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

E. Villar, T. Vannier, C. Vernette, M. Lescot, M. Cuenca, A. Alexandre, P. Bachelerie, T. Rosnet, E. Pelletier, S. Sunagawa, P. Hingamp. (2018), The Ocean Gene Atlas: exploring the biogeography of plankton genes online. Nucleic Acids Research.

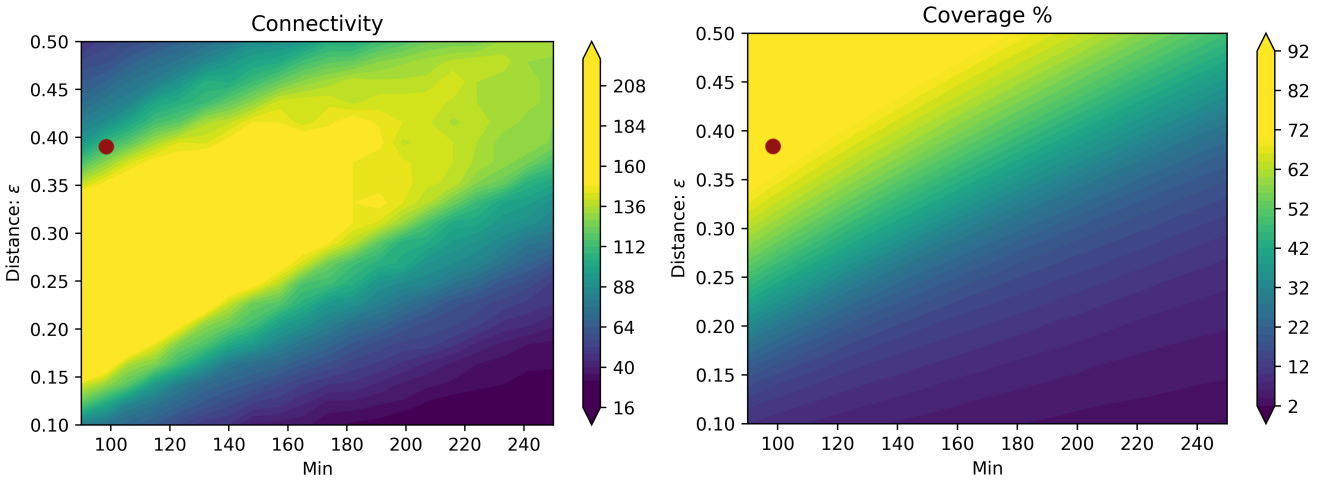
Wunsch, C., and P. Heimbach, 2007: Practical global oceanic state estimation. Physica D, 230, 197-208.

Wunsch, C. and P. Heimbach, 2013, Dynamically and kinematically consistent global ocean circulation and ice state estimates. In Ocean Circulation and Climate 2nd Edition, Siedler et al., Eds.

Appendix A: Importance of parameter selection

The careful determination of the DBSCAN parameters is necessary to obtain results that consistently give meaningful and robust results. Degeneracy of clusters is likely if the parameters are not chosen with care, meaning that the regions that the t-SNE+DBSCAN method find determine significantly different regions. To set the ε and minimum parameters for DBSCAN Fig. 6a and 6b were used. In Fig. 6a a sharp increase is seen to a plateau (yellow, $>$ 200 clusters), followed by a sharp decrease (green, 100 clusters) up to a minimum of ca 130, surrounded by regions of very few clusters (blue, $<$ 60 clusters). In the blue regions for a minimum of 100, either one cluster largely dominates the whole ocean ($\varepsilon < 0.42$), or most of the ocean is classified as noise ($\varepsilon > 0.99$). The yellow region has a highly variable, non-reproducible, cluster distribution, with increasing noise as ε is reduced. The green region of sharp increase is referred to as the "elbow", as a line-plot would show it as a sharp kink. This is the optimal region, where robust clusters are be identified, as determined using the intra-province BC-dissimilarity, despite the probabilistic t-SNE. The minimum number is set in the range below 130 using expert guidance with knowledge of what is apropriate for e.g. the modeling framework and resolution. Where there is a more "gradual" decrease in the connectedness ε is hard to determine and needs evaluation on a case-by-case basis for each t-SNE realisation. A number of 100 gridpoints was chosen, denoted by the location of the red dot in Fig. 6a and 6b. A distance ε was chosen to be at the "elbow"/kink where the connectedness increases suddenly. This is symmetric, and the point allowing maximal area coverage was preferred. When the minimum number of points to be designated as a cluster is chosen, the lower limit is set for what spatial scales the identified provinces and aggregated eco-provinces are relevant for. This step should be guided by the application at hand, which in the present case was set to 100 points.

Note that using a method like k-means would *not* be able to determine robust provinces. Using the DARWIN data, the robustness measure based on the Akaike and Baysean Incormation Criteria utilized in Maze *et al.* (2016); Sonnewald *et al.* (2019) fails to converge even with $K > 1000$. This number of eco-provinces is clearly impractical, and the provinces are assigned at random. The arbitrary assignment is based on the topology of the variance in the data. Visualized using the 3D t-SNE, the data are shaped like strings or sheats (Fig. 2). The K-means algorithm searches for Gaussian (round) areas of variance, drawing straight lines (Voroni cells). The DARWIN data does not have this type of topology in its 55D variance space, making k-means a poor choice.



(a) Connectedness

(b) Coverage

Figure 6. Setting the parameters for t-SNE the resultant number of found clusters is used as a measure of the connectedness (Fig. 6a) and the percentage of the data assigned to a cluster (Fig. 6b). The red dot illustrates the optimal combination of coverage and connectedness. The minimum number was set on the basis of minimum number relevant for ecology.