Decoding Ecological Complexity: Machine Learning reveals 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

The complex nature of ecology in the global ocean is leveraged to present the emergent biogoechemical provinces. A global framework based on unsupervised machine learning is presented that allows a robust and nested classification of biogoechemical provinces spanning global and regional scales. Similar ecology is identified in 20 year mean surface data from state of the art biogeochemical model DARWIN. Initially, a large number of global eco-provinces are identified, using 51 species with a range of trait-based functional types and nutrient fluxes of N, P, Si and Fe coupled to physics from the ECCO State Estimate. Provinces are determined by nesting the provinces according to their ecological similarity, and a minimum complexity is determined by comparison to the Longhurst provinces. The presented framework is applicable both globally and regionally as the degree of nesting can be adjusted according to the region of interest. A dynamic ocean ecology emerges, with gradual and sharp transitions between eco-provinces. Expected compositions are found in oligotrophic gyres or regions of upwelling and strong seasonality.

Introduction

Biogeography is used in terrestrial and marine studies for organizing the complex biogeochemical realm into regions based on similarity. Such regions are often called biomes or provinces. Here we will use the term "provinces" as defined as areas where specific set of variables (e.g. environment, flora) are distinguishable and unique at global scale. Terrestrial provinces are often classified according to similarity in climate (precipitation, temperature), soil, vegetation, and fauna. Such biogeography has been important for variety of issues, including management, biodiversity studies, and disease control. The oceans are more difficult to define into provinces given that the majority of organisms are microscopic, the environmental similarity is more difficult to determine, in situ observations are sparse, and satellite measurements only capture the 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 such as surface mixing rates, stratification and irradiance, along with expert knowledge of other conditions that are important to the phytoplankton (the photosynthesizing organism at the base on the marine foodweb). The 56 Longhurst provinces have been widely used in studies looking at biogeochemistry of the oceans (e.g. carbon fluxes), fisheries and are even used as criteria in adding in situ observations to certain databases. Fuzzy logic, machine learning, OTHERS, have since been used to define provinces in more rigorous ways (REFS). For instance, SeaScapes (Kavanaugh et al, REF) uses self organizing maps to make the data more smooth, and hierarchical (tree based) clustering to define provinces on the basis of satellite derived Chlorophyll, photosynthetically available radiation and mixed layer depth. SeaScapes has been used for biogeochemical applications (REF) as well as more recently for coastal management (REF).

These previously studies defined "biogeochemical provinces", such they considered bulk properties of the system. There has not been an as much attempt to define the oceans in terms of eco-provinces: Regions with similarity in the types and abundances of organism that co-exist. This is largely due to sparsity of available data of ecological characteristics. Studies that have considered the biogeography of the microscopic phytoplankton have done so in terms of the most abundant type (e.g. Alvain et al, 2008) or presence/absence of a small subset of organisms. New surveys and greater ease of genomic data analysis will begin to improve our coverage of the marine ecosystem. Observations from efforts such as Tara Oceans feed into databases such as the Ocean Gene Atlas OGA. Studies using Tara Ocean have made use of machine learning tools to help define communities (e.g. Lima-Mendez et al., 2015), but have not yet had sufficient coverage to define distinct provinces.

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. Grazing pressures are also important in controlling communities. The combination of interactions result in highly complex patterns, as found by recent studies using the TARA datasets (deVagas et al., 2015, Lima-Mendez et al., 2015). Despite the importance of phytoplankton ecology in the climate and for fisheries, identifying key components setting ecosystem structure in the pelagic ocean is lacking. Here a framework is presented that leverages novel unsupervised machine learning techniques to provide the necessary tools to identify the underlying mechanisms that set the ecosystem structure at different location.
Following Sonnewald *et al.*, 2019, the purpose here is proof-of-concept study that leverages unsupervised learning as a tool to define distinct eco-provinces. Determining a useful classification, such a scheme needs to allow for both 1) global classification, and 2) a multiscale analysis that can be both spatially and temporally nested. In this study we show how these provinces can help us understand some of the controllers of community structure, but such as system could also provide insight into monitoring strategies and as a way to track ecosystem changes.

The present work uses output from an existing global 3-dimensional physical/ecosystem model (DARWIN) to explore a method to define eco-provinces. The ecosystem is sufficiently complex (encompassing 35 phytoplankton and 16 zooplankton types) that simple diagnostics are not capable of defining similarity of community structure. The method (Figure 1a) first reduces the dimensionality of the problem by first aggregating the plankton into 7 functional groups, and then by using a method to projecting a 11-dimensional data (7 functional groups and the 4 nutrient supply rates which are hypothesized to control community structure) into a 3-dimensional configuration (t-SNE). A clustering algorithm then connects regions of similarity (DBSCAN). The resulting eco-provinces are then back-projected onto the globe. There provinces are coherent and unique. However, the over 100 eco-provinces that this method identifies are initially overwhelming. To reduce the problem to something tractable we then Aggregate Eco-Provinces (AEPs, Fig 1b). The minimum number of sensible provinces is determined. Following the purpose of the paper as a proof of concept, we then provide a short example how these classifications can help us to explore controllers on the complex community structures.

Model framework: DARWIN

The complexity at the heart of ecological province construction makes it a problem well suited for machine learning applications. The approach here assumes the modeled ecology is correct, using the DARWIN model and treating the model data as truth. An Ecosystem-Optical model uses the physical model. The eco-provinces are defined by the proportional presence of phytoplantion, zooplankton and nutrients (Nitrogen, Iron, Phosphate and Silica). The 51 plankton types (2 pico-prokaryotes, 2 pico-eukaryotes, 5 coccolithophores, 5 diazotrophs, 11 diatoms, 10 mixotrophic dinoflagellates, and 16 zooplankton) range from 0.6μm to 2500 μm equivalent spherical diameter. Parameters influencing growth, grazing, and sinking are related to size (following Ward et al., 2012) with specific differences between functional groups. The plankton distributions in this model compare well with both observations based on functional types as well as size distributions.

The physical component comes from the Estimating the Circulation and Climate of the Ocean (ECCOv4) global state estimate described by Wunsch and Heimbach, 2013, Forget *et al.*, 2015, see also ECCO Consortium, 2017a, ECCO Consortium, 2017b. The state estimate has a nominally 1° resolution, available at: ecco.jpl.nasa.gov. A least-squares with Lagrange multipliers approach is used to obtain the state estimate. The result is a *free-running* version of the MIT General Circulation Model (MITgcm, Adcroft *et al.*, 2004), with adjusted initial and boundary conditions and internal model parameters. The ECCO state satisfies basic conservation laws for enthalpy, energy, salt, volume, and momentum while remaining largely within error estimates of a diverse set of global data Wunsch and Heimbach, 2007, Wunsch and Heimbach, 2013, **?**. Regions without data are filled in a dynamically consistent way using the dynamics, still relying on parameterizations but avoiding the use of untested statistical hypotheses e.g., Reynolds *et al.*, 2013.

Unsupervised Machine learning: Identifying and aggregating eco-provinces

0.1 Dimensionality reduction with t-SNE

By identifying the eco-provinces in the 20 year average of the surface field, this work aims to provide a framework for looking a important relations between key features such as biomass of species. Like drawing a map of a complex mountainous region, careful thought needs to be given so the key features remain intact. 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 55-dimensional space to all 11. 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; 2014, in prep). The biomass and nutrient fluxes leave an 11 dimensional vector $\mathbf{x}$. If $\mathbf{x}$ is a vector field on the model grid discretized sphere, where each element $\mathbf{x}_i$ represents an 11-dimensional 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 sever functional groups and 4 nutrient terms. The log of the biomass data is used, and if a cell has less than $12 * 10^{-4} mgChl/m^3$ is is discarded. 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 features with wider range of possible values when computing distances that map the key relations between features. In ecological terms, this step 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 their proximity in the high-dimensional parameter 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 sateliteDNN. This is a necessary step to identify robust clusters in the biogeochemical data. 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 and vice versa[van der Maaten & Hinton, 2008]. Given a set of N high-dimensional objects x_1 ,..., x_N, the t-SNE does a reduction by minimizing the Kullbach-Leibner distance[Kullback, 1987] between the likelihood of association between a low dimensional rendition and the high dimensional data. If x_i is the i-th object in the N dim space and y_i is the i-th object in the low-dim 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:

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

This is done as:

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

Figure 2a illustrates the effect of reducing the variables of phytoplantion, zooplankton and nutrients. While different in practice, the motivation for applying t-SNE is similar to a principal component analysis (PCA) decomposition, using a method to highlight attributes in the data and reduce the dimentionality. The t-SNE method was found to be much superior to PCA in delivering robust results. PCA did very little to make the applied clustering more able to identify robust and repeatable clusters. 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 demonstrating 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

Applying t-SNE, the result is a 3D representation of the 11D feature vector. Fig. 2a illustrates this space, where the points are the 3D representation of every latitude and longitude point. 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. The colors on Fig. 2a are the clusters found using the Density-based spatial clustering of applications with noise (DBSCAN) proposed by[Ester *et al.*, 1996]. Looking for densely packed observations, the DBSCAN algorithm uses the distance in the 3D representation between points, setting a minimum distance for associated points (ε), and the number of similar points needed to be interesting. 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 itteratively 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 distance metric are counted as "noise". DBSCAN is a fast and scalable algorithm, with a worst-case performance of O(n). 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). A minimum number of points was set to an area of 100 gridpoints denoted by the black line 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.

Each cluster $j = 1,...,N$ is represented by the 11-dimensional characterizing vector of the cluster. The DBSCAN classification attributes each vector $\mathbf{x}_i$ to a unique cluster if it is not "noise". The 115 clusters identified in Fig. 2a are presented in their geographical extent in Fig. 2b. Each colour corresponds to a geographically coherent combination of biogeochemical factors picked out by the unsupervised learning algorithm DBSCAN after processing by t-SNE. Striking features stand out in Fig. 2a in that globally coherent regions are identified.

Back-projecting onto the globe

Once the clusters in the 3D data are determined, each point is still associated with a specific latitude and longitude. Keeping track of this, the data are projected back into the more familiar geographical domain. Figure 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 them. In Fig. 2b familiar regions appear, such as the zonally symmetric Southern Ocean, the oligotrophic gyres, sharp transitions reminiscent of the trade winds, and distinct regions associated with upwelling e.g. in the equatorial Pacific. Many of these regions are also described by the Longhurst provinces.

Ecological similarity: BC-dissimilarity

To understand the ecological context of the provinces, the intra-cluster ecology is assessed using the Bray-Curtis Dissimilarity metric (BC[Bray and Curtis, 1957]). BC is defined as:

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

where BC is always measured comparing 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 a certain 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} represents the sum over all the biomass present in assemblage n_i. For each cluster identified in Fig. 2b, the coherence within each province is illustrated in Fig. 4c. This is determined using the mean area averaged assemblage within one cluster, and determining the BC-dissimilarity of each gridpoint within the province to the mean. The global mean intra-cluster BC-dissimilarity is 0.102 ± 0.0049.

The Longhurst intra-province BC-dissimilarity is presented in Fig. 4b using the biomass data from DARWIN, with a global mean of 0.227. This is significantly larger than for the clusters identified by the framework presented in Fig. 1a. For defining aggregated eco-provinces, the intra-province BC-dissimilarity from the Longhurst provinces of 0.227 is used as a benchmark.

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 improvement on the Longhurst provinces, the utility of over hundred provinces is limited for global applications as the complexity if still overwhelming. One of the uses of provinces is to facilitate understanding of how global provinces are set and governed. To leverage the emergent properties, the framework summarized in Fig. 1a is developed further to allow a nesting of ecologically similar provinces. An adjustable level of

"complexity" is determined between a minimum benchmark set by the Longhurst provinces and the maximal number of the original provinces from the full complexity in Fig. 2a.

The distinction between intra- and inter-province BC-dissimilarity is used to mean that the intra-province BC-dissimilarity refers to the mean dissimilarity within a province itself. The inter-province BC-dissimilarity refers to how similar one province is to each other province. Fig. **??** 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 emergent provinces are globally coherent across the globe as Fig. 3b demonstrates. Some configurations are very "common", and using methods from genetics such as "connectivity" we can sort the > 100 provinces according to which province they are most similar to con. Using a set number of "provinces" P, the P most dominant ones are used to organize the remainder of the provinces into the province most similar to them. This organization according to the BC-dissimilarity allows the nested approach to global ecology. The complexity of the regions can be anything up to the full complexity from Fig. 2a.

What level of complexity is appropriate for the global ocean? The Longhurst provinces are leveraged as our benchmark in Fig. 4a (green line). Moving from a complexity of 1, the provinces are aggregated, and the intra-province BC dissimilarity assessed. The entire operation represented schematically in Fig. 1a is repeated ten times, as a slight difference could arise from using t-SNE. The computational cost associated limits the number, with the mean (black line) and 2σ (blue area) in Fig. 4a representing the presented method. A complexity of 12 is demonstrated to keep the intra-province BC-dissimilarity both below the Longhurst benchmark and and shows a small 2σ. For a complexity of 12, the mean intra-seasonal BC-dissimilarity is 0.198 $\pm$0.013, as seen in Fig. 4d.

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

Taking the minimum complexity of 12 as determined in Fig. 4a, the emergent global AEPs can be assessed. With knowledge of the geographical extent of the 12 AEPs, the overwhelming complexity in the 55D original data is revisited to gain insight into the emergent ecology. The expected region distinctions of biomass rich upwelling, picoplankton dominated oligotrophic gyres and diaton rich polar regions is apparent. Fig. 5 illustrates the ecological insights grouped by AEP a to L is geographical extent (Fig. 5c), functional group biomass composition (Fig. 5a) and nutrient supply (Fig. 5b).

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], features that stand out include Western Boundary Currents (WBCs), the Antarctic Circumpolar Current (ACC), known upwelling regions on the Eastern side of the ocean basins and the subtropical gyres. Taking the geographical extents of the provinces A to L, the ecological assemblages for each province are shown in the bar plot in Fig. 5a. The nutrient fluxes are similarly shown in Fig. 5b, and have been scaled by N in the Redfield ration.

Provinces are seen have very similar phytoplankton biomass, but very different communities (e.g. D, H and K). These three regions are very different, with H being present mainly in the equatorial Indian ocean and having a larger population of diazotrophs. Province 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. Province D has very few diazotrophs but more cocolithophores. Province K is found only in the high Arctic ocean, and is dominated by diatoms. The nutrient levels in each province are somewhat similar, with K having more Si, and H and D having very little. There is less N compared to P in province H. If only Chlorophyll were used to define provinces D, H and K they could not be distinguished. It is notable that the zooplancton biomass in the three regions are very different, with K having very little, but 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 provinces that have very different biomass are actually very similar in terms of their ecological community structure. This is seen in provinces D and E. These are close to each other, notably in the Pacific, where E is closer to the highly productive province J. There is again not a clear connection between phytoplankton biomass and zooplankton abundance. In this manner, monitoring Chlorophyll is 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. If this limitation were not present diatoms would dominate in all but the lowest nutrient supply regions. Diatoms are seen in the provinces A, J K, and L together with silicon supply. As expected, diatoms are found in the polar regions. Province L spans the southern Southern Ocean and province K is a region in the Arctic Ocean. The very productive region J has diatoms, and province A is often an extension of J outside of the Equatorial regions.

The diazotrophs are able to coexist with other phytoplankton where there is excess of Fe, P relative the the demands of the non-diazotrophs. It is notable that there is higher diazotroph biomass where the amount Fe,P supply are relatively larger. In this manner, the diazotroph biomass in province H is larger than in J, although the overall biomass in J is larger. It is worth noting that provinces J and H are very different, with H located in the equatorial Indian Ocean.

The patterns demonstrated in the lowest level of complexity would be very difficult to determine is the biomass data were not separated into provinces. The simultaneous comparison of a plethora of maps, compared to Fig. 5c, 5a and 5b is daunting. What the provinces also highlight effectively, is that chlorophyll is not a good proxy for productivity, and leads to questions about how accurate the use of chlorophyll is as a proxy for biomass.

Discussion and Conclusion

A framework for decoding the overwhelmingly complicated ecological data from DARWIN is presented. The determined provinces are global, and a framework for nesting them into AEPs is presented. The nesting can be adjusted between the full complexity of the original provinces and a minimum threshold set using the Longhurst provinces as a benchmark.

Efforts to determine true patterns and signals is a challenge even is simple systems. Emergent complexity can appear simply complicated until the underlying principles are determined that give rise to the observed patterns. The 55D model data is very difficult for an observer to process manually, and the presented framework demonstrates how patterns can emerge that are humanly-tractable, and the ecological implications evident.

The framework is global, and can span a range of complexity from >100 EBPs to 12. Th process of nesting was here applied to the lowest recommended complexity. However, regional studies can focus on a small subset of the global map, and find rich descriptions readily accessible. This is critical in interpretations of how general samples from one location may be. Long term observational datasets are, and will continue to be, invaluable. The presented framework presents a map of regions that should be samples, and a means of assessing how generally informative the region is, for example my illustrating how seasonally dominated the ecology is. The framework can be repeated for climatological data (not shown).

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 the Poincarre section are used to reduce the dimensionality and the volume occupied demine "regular" or "chaotic" areas, we organize the 55D 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 eco-provinces are all unique by construction. They conform to firmly established biological dynamics prescribed by the DARWIN model. The provinces isolate regions that are highly seasonal, dominated by upwelling, light limited Polar regions, high nutrient low chlorophyll regions, as well as highly productive regions and oligotrophic regions with very low biomass.

The full complexity features over 100 provinces. The aggregation method featuring the BC-dissimilarity allows nesting of the provinces down to a complexity of 12 AEPs (Fig. 5). This nesting provides allows the study of broader swaths of similar ecological composition to be studied as a whole to determine for example what the controls on the community structure are. Assessing the 12 AEPs, the presence of similar biomass but significantly different ecological composition highlights regions

where using chlorophyll alone is not a sufficient indication of the ecology (e.g. D and E). AEPs such as D and K have very different biomass but similar ecological composition. Biomass is seen to be a poor predictor of of zooplankton (e.g. D+E or K+L). This has implications for using chlorophyll to assess the base of the foodchain for fisheries, and the potential for trophic cascades in the biomass were to change. The presented framework allows identification of key regions where relying on chlorophyll to assess grazer biomass and ecological structure is misguided. Using the provinces and AEPs, these studies can be ameliorated to be more accurate and increase utility.

The utility of the provinces and AEPs includes being able to assess the controls on the community structure. The relative amount of diatoms is set by the imbalance in the Si to N,P, Fe supplies. With balanced supplies 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 provinces and AEPs can be used to provide a context for the sparse in-situ sampling of observational studies. Some areas can be seen as representative of larger areas of the ocean, and samples can be readily extrapolated, while others are variable and need high spatial resolution. For example, the timeseries ALOHA is located quite close the a boundary to a very different AEP, suggesting that it should not be considered representative of the entire AEP it belongs to. The PAP station 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 how representative an observation likely is.

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, arranged as "streaks" in the 3D space, or associated with very few datapoints. Both likely failing the "minimum" criterion of DBSCAN.

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

Acknowledgements

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

Author contributions statement

M. Sonnewald developed the method and ran the analysis, she wrote the main body of the text.
Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

References

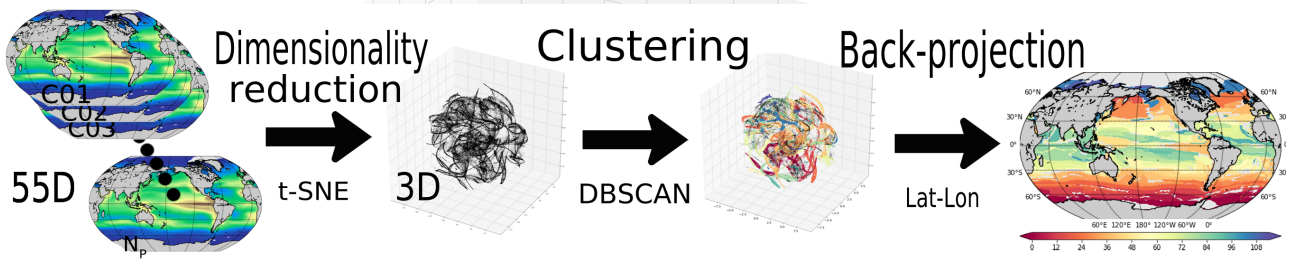
Note, in progress. : This list of references is not complete.

Adcroft *et al.*, 2004. 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.

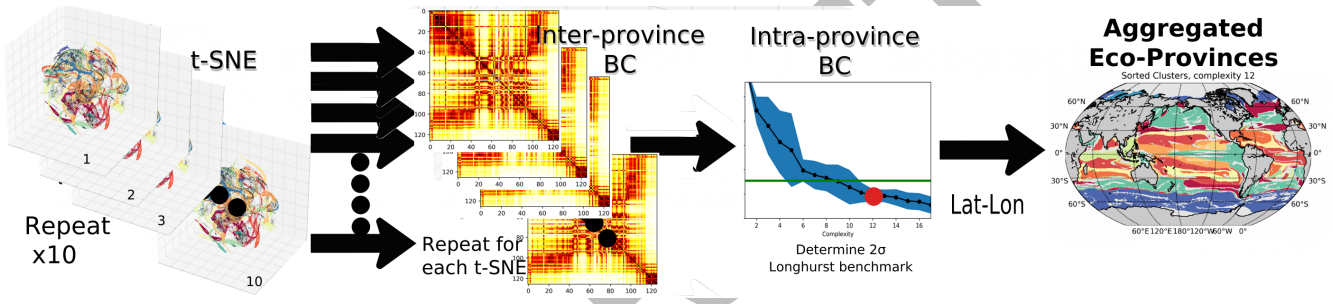
Bray and Curtis, 1957. Bray, J. R. and J. T. Curtis. 1957. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27:325-349.

Costanzo *et al.*, 2016. 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.

ECCO Consortium, 2017a. 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.

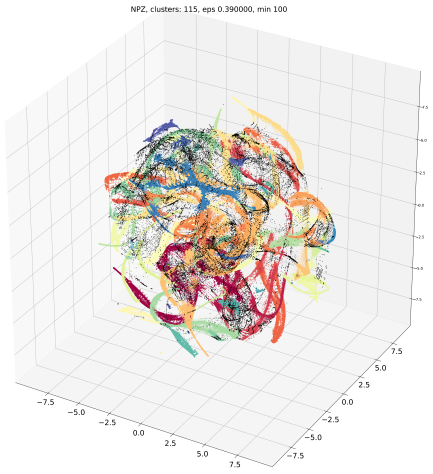


(a) Work flow repeated ten times

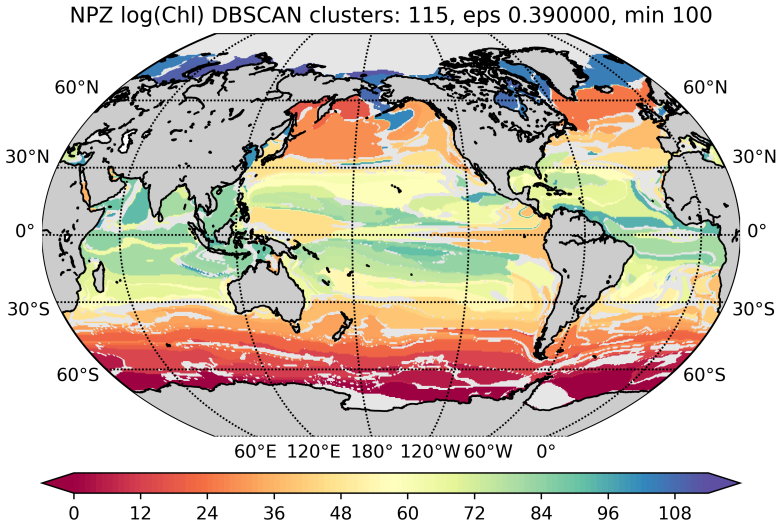


(b) Determining the Aggregated Eco-Provinces (AEPs)

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.

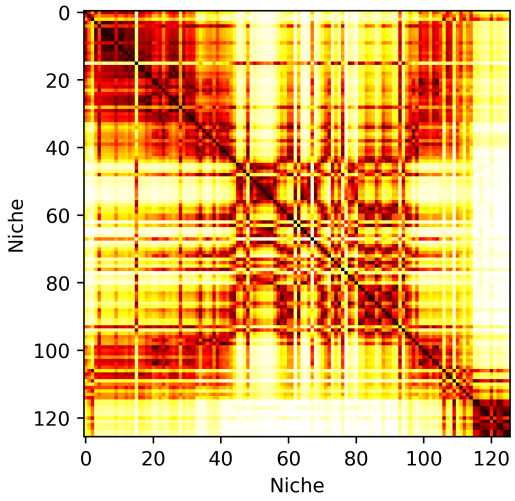


(a) t-SNE projection with provinces in colour

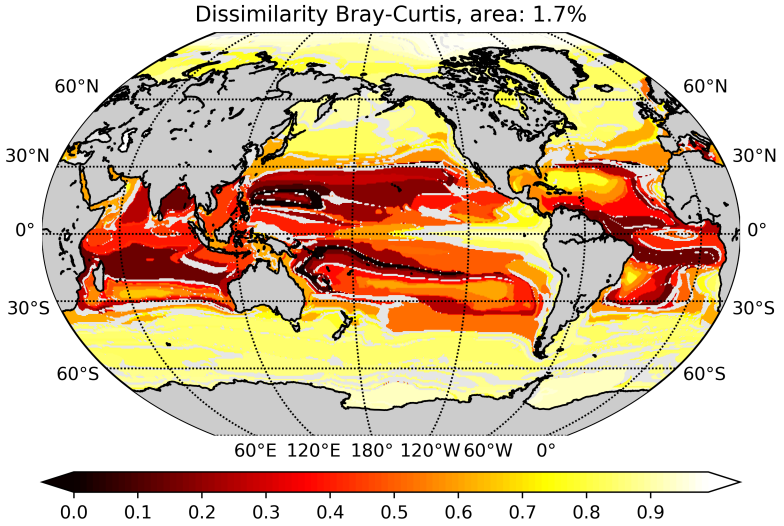


(b) Spatial representation provinces in Fig. 2a BC-dissimilarity

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.

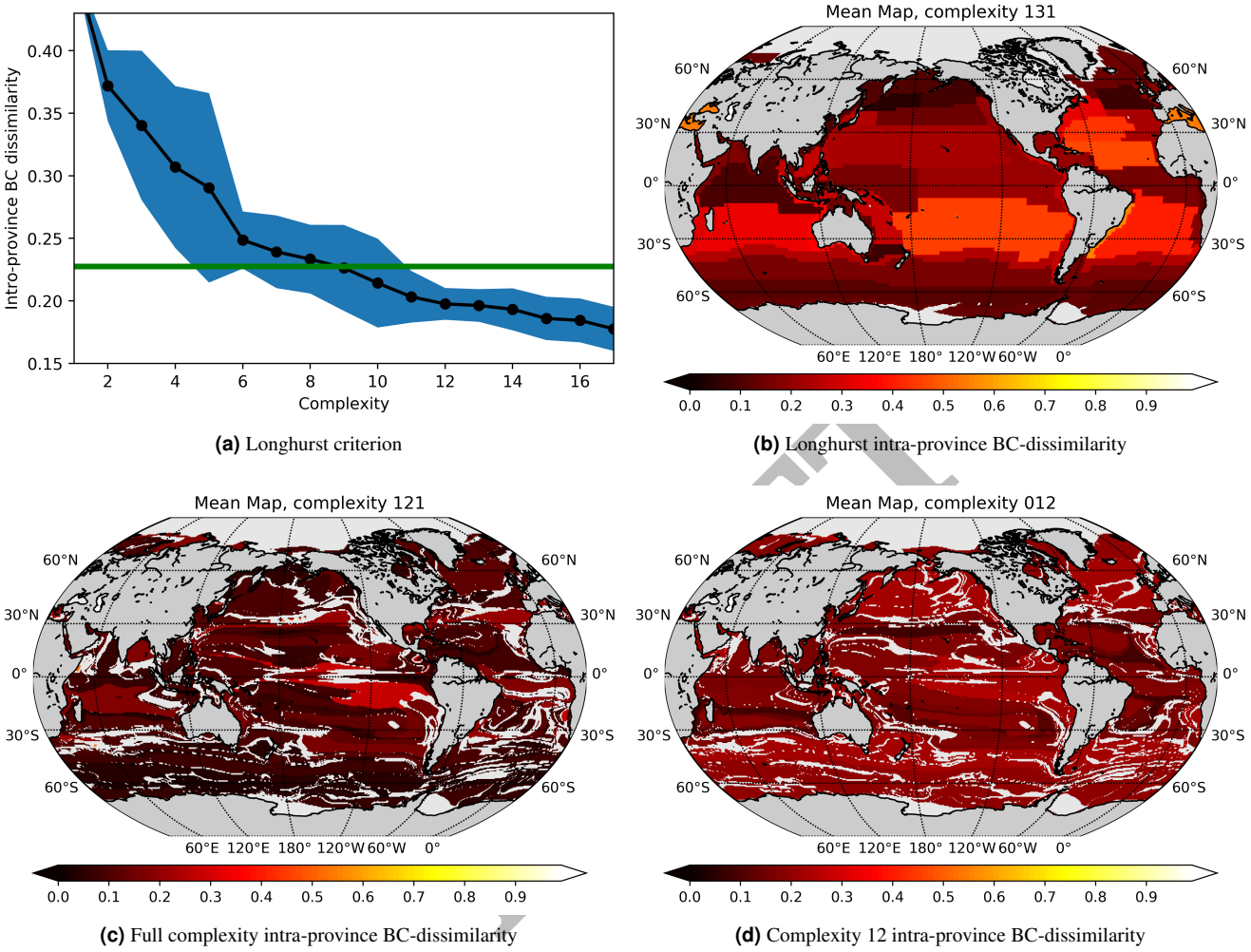


(a) Symmetric BC matrix



(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.



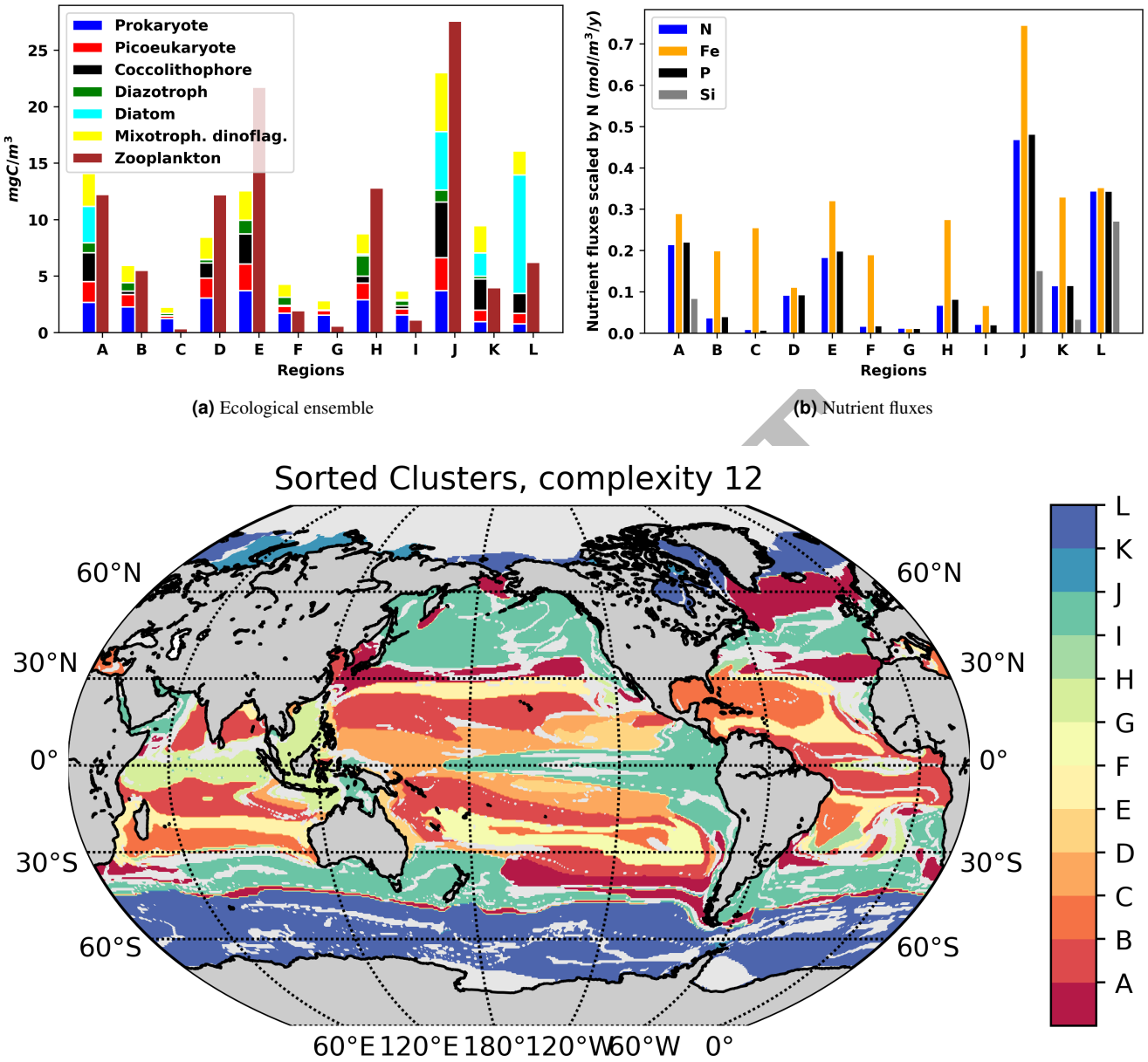
(a) Longhurst criterion

(b) Longhurst intra-province BC-dissimilarity

(c) Full complexity intra-province BC-dissimilarity

(d) Complexity 12 intra-province BC-dissimilarity

Figure 4. How complex should the biogeochemistry be? The top left Fig. 4a 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 province recognition process. For Fig. 4b, 4d and 4c 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. 4a). For the full complexity in the provinces discovered by DBSCAN, Fig. 4c illustrates that an intra-province BC-dissimilarity of 0.099 is reached, while sorting into a complexity of 12 as suggested by Fig. 4a 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.

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

Ester *et al.*, 1996. 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. Flanders Marine Institute (2009). Longhurst Provinces. Available online at http://www.marineregions.org/. Consulted on 2019-03-15.

Forget *et al.*, 2015. 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

Fukumori *et al.*, 2018. 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

Henon and Heiles, 1964. 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.

Kullback, 1987. 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, 2007. Longhurst, Ecological Geography of the Sea (ISBN 0124555217), Academic Press, 2007, San Diego, 560p.

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

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

Lunga *et al.*, 2014. 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

Marmanis *et al.*, 2016. 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

Reynolds *et al.*, 2013. 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,

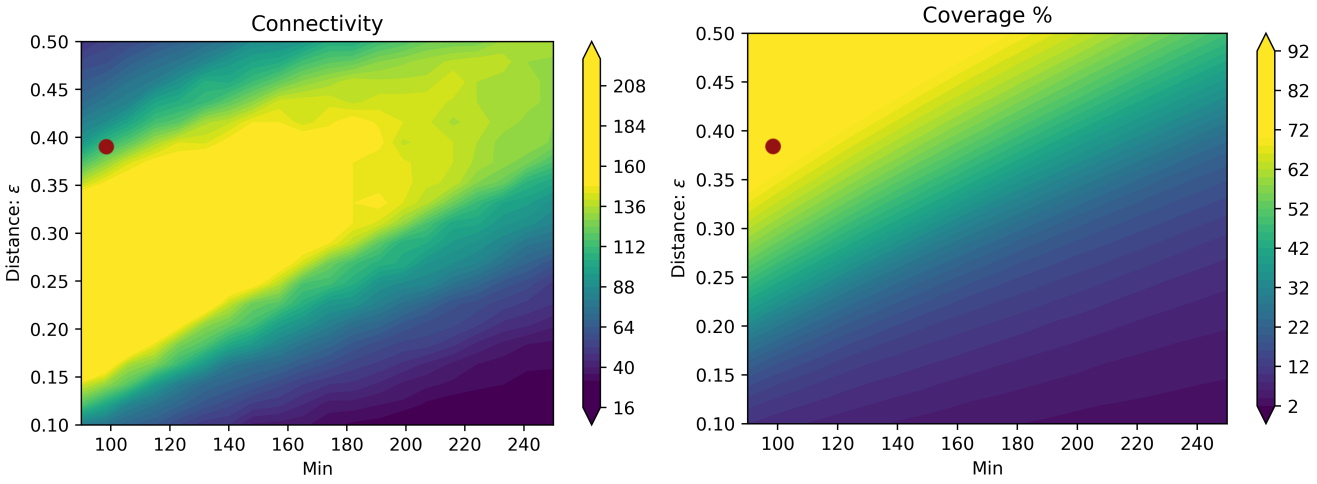
Sonnewald *et al.*, 2019. Sonnewald, M., Wunsch, C., Heimbach, P. (2019). Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions. Earth and Space Science, 6. https://doi.org/10.1029/2018EA000519

van der Maaten & Hinton, 2008. 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.

Villar *et al.*, 2018. 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 and Heimbach, 2007. Wunsch, C., and P. Heimbach, 2007: Practical global oceanic state estimation. Physica D, 230, 197-208.

Wunsch and Heimbach, 2013. 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.



(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.