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

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

[1]Massachusetts Institute of Technology, Department of Earth, Atmospheric and Planetary Sciences
Cambridge, MA 02139, USA
[2]Harvard University, Department of Earth and Planetary Sciences
Cambridge, MA 02138, USA

*To whom correspondence should be addressed; E-mail: maike_s@mit.edu.

An unsupervised learning method is presented for determining global marine ecological provinces (eco-provinces), from plankton community structure and nutrient flux data. The Systematic AGgregated Eco-province (SAGE) method identifies eco-provinces within a highly non-linear ecosystem model. To accommodate the non-Gaussian covariance of the data, SAGE employs t-stochastic neighbor embedding (t-SNE) to reduce dimensionality. Over a hundred eco-provinces are identified with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. Using a connectivity graph with ecological dissimilarity as the distance metric, robust aggregated eco-provinces (AEPs) are objectively defined by nesting the eco-provinces. Using the AEPs the control of nutrient supply rates on community structure is explored. Eco-provinces and AEPs are unique and aid model interpretation. They could facilitate model inter-comparison, and potentially improve under-

standing and monitoring of marine ecosystems.

Introduction

Provinces are regions in the ocean or on land where the complex biogeography has been organized into coherent and meaningful regions (*1*). Such provinces are important for comparing and contrasting locations, characterizing observations, monitoring, and conservation efforts. The intractably complicated and non-linear interactions that create these provinces make unsupervised machine learning (ML) methods well suited to objectively determine provinces, because the covariances within the data manifest as intricate and non-Gaussian. Here, an ML method is presented that systematically identifies unique marine ecological provinces (eco-provinces) from the Darwin global 3 dimensional physical/ecosystem model (*2*). The term "unique" is used to signify that the identified region is sufficiently different from other regions that they do not overlap. The method is called the Systematic AGgregated Eco-province method (SAGE). For useful classification, an algorithmic method needs to allow for both 1) global classification, and 2) a multi-scale analysis that can be both spatially and temporally nested/aggregated (*3*). In this study, the SAGE method is first presented, and the identified eco-provinces are discussed. The eco-provinces could facilitate understanding of factors controlling community structure, provide insight useful for monitoring strategies, and assist in the tracking of ecosystem changes.

Terrestrial provinces are often classified according to similarity in climate (precipitation, temperature), soil, vegetation, and fauna, and used to aid management, biodiversity studies, and disease control (*1, 4*). Ocean provinces are more difficult to define. The majority of organisms are microscopic, and the boundaries are fluid. Longhurst (*5*) provided one of the first global classifications of marine provinces based on environmental conditions. These "Longhurst"

provinces were defined using variables such as mixing rates, stratification and irradiance, along with Longhurst's extensive experience as a seagoing oceanographer of other key conditions important to the marine ecosystem. The Longhurst provinces have been widely used, for example to assess primary production, carbon fluxes, to aid fisheries and to plan in situ observational campaigns (*5–9*). Toward defining provinces more objectively, methods such as fuzzy logic and regional unsupervised clustering/statistics have been used (*9–14*). Such methods have the goal of identifying meaningful structures that can identify provinces in available observational data. For instance, dynamic seascape provinces (*12*) use self organizing maps to reduce noise, and hierarchical (tree based) clustering to identify provinces on the basis of regional satellite derived ocean color products (Chlorophyll-a, normalized fluorescence line height, colored dissolved organic material) and physical fields (sea surface temperature and salinity, absolute dynamic topography and sea ice).

Plankton community structure are of interest as their ecology has a large impact on higher trophic levels, and also on carbon uptake and hence climate. Despite this, identifying global eco-provinces based on plankton community structure remains a challenging and elusive goal. Ocean color satellites can potentially offer insight in terms of coarse grained size fractionation of phytoplankton, or suggest dominance of functional groups (*15*), but cannot currently provide details of community structure. Newer surveys (e.g. TARA ocean (*16*)) are providing unprecedented measurements of community structure, there are at present only sparse in situ observations at a global scale (*17*). Previous studies have largely determined "biogeochemical provinces" based on identifying biochemical similarities such as in primary production, Chl and available light (*12, 14, 18*). Here, numerical model output (Darwin (*2*)) is used, and eco-provinces are determined in terms of community structure and nutrient fluxes. The numerical model used in this study has global coverage and compares favorably to available in situ

data (*17*) and remotely sensed fields (note S1). The numerical model data used in this study has the advantage of global coverage. The model ecosystem consists of 35 phytoplankton and 16 zooplankton types (see materials and methods). The model plankton types interact non-linearly, with non-Gaussian covariance structure, such that simple diagnostics are not well suited to identifying unique and coherent patterns in the emergent community structure. The SAGE method presented here provides a novel method to examine the complex Darwin model output.

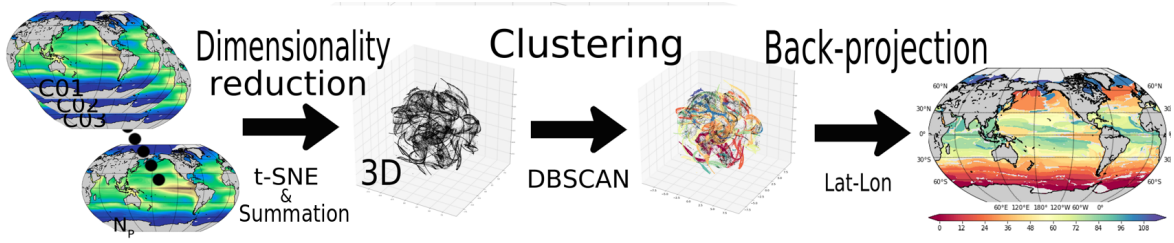
The transformative power of data science/ML techniques can allow overwhelmingly complicated model solutions to reveal complex, but robust, structures in the covariance of data. A robust method is defined as one that can faithfully reproduce results within a given error margin. Determining robust patterns and signals is a challenge even in simple systems. Emergent complexity can appear complicated/intractable until the underlying principles giving rise to the observed patterns are determined. Key processes setting ecosystem composition are inherently non-linear. The presence of non-linear interactions can confound robust classification, and methods that make strong assumptions about the underlying statistical distributions of the covariance of data need to be avoided. High-dimensional and non-linear data is common in oceanography, and likely have covariance structures with intricate, non-Gaussian, topology. While data with a non-Gaussian covariance structure can hamper robust classification, the SAGE method is novel as it was designed to allow identification of clusters with arbitrary topology.

The goal of the SAGE method is to objectively identify emergent patterns that could help further ecological understanding. Following a clustering based work-flow similar to that in (*19*), the ecological and nutrient flux variables are used to determine unique clusters within the data, referred to as eco-provinces. The SAGE method presented in this study (Fig. 1) first reduces

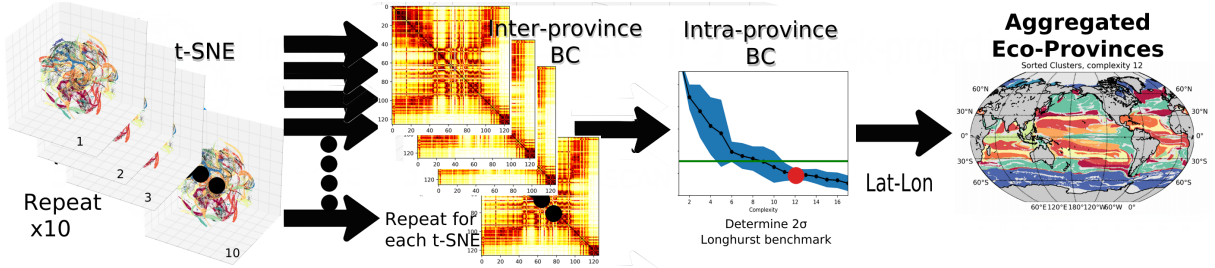
the dimensionality from 55 to 11 dimensions by summing over the a priori defined plankton functional groups (see materials and methods). The dimensionality is further reduced by a probabilistic projection onto a 3 dimensional space using the t-Stochastic Neighbourhood Embedding (t-SNE) method. Unsupervised clustering identifies regions of close ecological proximity (Density-based spatial Clustering of Applications with Noise, DBSCAN). Both t-SNE and DBSCAN are suitable for the inherently non-linear ecosystem numerical model data. The resulting eco-provinces are then back-projected onto the globe. Over a hundred unique eco-provinces are determined, suitable for regional studies. To consider global coherent ecosystem patterns, the eco-province utility is increased by aggregating eco-provinces (AEPs) down to an adjustable level of "complexity" defined as the level of aggregation. The minimum complexity number for robust AEPs is determined. The chosen focus is on the SAGE method, and on exploring the minimal complexity AEPs case to determine controls on the emergent community structures. Patterns can subsequently be analyzed, offering ecological insight. The approach presented here can also be used more widely, for example for model inter-comparison by assessing where similar eco-provinces are found in different models to highlight discrepancies and similarities.

Results: Identifying and aggregating eco-provinces

The SAGE method defines eco-provinces using output from a global 3 dimensional physical/ecosystem numerical model (Darwin (*2*), see materials and methods and note S1). The ecosystem component consists of 35 phytoplankton types and 16 zooplankton types, with 7 a priori defined functional groups: prokaryotes and eucaryotes adapted to low nutrient environments, coccolothophores with calcium carbonate coverings, nitrogen fixing diazotrophs (often a key missing nutrient), diatoms with silicous coverings, mixotrophic dinoflagellates that both



(a) Work flow repeated ten times



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

Figure 1: **The SAGE method workflow.** Upper panel is Fig. 1a is a sketch of the work-flow to determine the eco-provinces; The raw 55 dimensional data in reduced using summation within functional groups to 11 dimensional model output, it includes biomass of 7 functional/trophic groups of plankton and 4 nutrient supply rates. Negligible values and persistent ice cover areas are discarded. Data is normalized and standardized. The 11 dimensional 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 latude/longitude projection. Note this process is repeated 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 biomass of the 51 phytoplankton types. The BC-dissimilarity within the aggregated provinces is determined going from a complexity of 1 AEP to full complexity of 115. The BC-benchmark is set by Longhurst provinces.

photosynthesize and graze other plankton, and zooplankton grazers. Sizes span 0.6μm to 2500 μm equivalent spherical diameter. The model distribution of size and functional grouping of phytoplankton capture gross features seen in satellite and in situ observations (see Fig. S1-3). The similarity between the numerical model and the observed ocean suggests that provinces defined from the model may have application to the in situ ocean. Note the caveats that the model only captures some of the diversity of phytoplankton, and only some of the range of physical and chemical forcings of the in situ ocean. The SAGE method could lead to a better understanding of the highly regional controlling mechanisms of the model community structure.

The dimensionality of the data is initially reduced by including only the surface, 20 year time-mean sum of biomass, within each plankton functional group. Surface source terms for the flux of nutrients (Nitrogen, Iron, Phosphate and silicic acid supply) are also included, following earlier studies showing their key roles in setting community structure (e.g. (*20, 21*)). Summation over functional groups reduces the problem from 55 (51 plankton and 4 nutrient fluxes) to 11 dimensions. In this initial study, depth and temporal variability are not considered due to computational limitations imposed by the algorithms.

Dimensionality reduction with t-SNE

The SAGE method is able to identify important relationships between the non-linear processes and interacting key features in the biomass of functional groups and nutrient fluxes. Obtaining robust, reproducible, provinces is not possible with the 11 dimensional data using learning methods based on Euclidean distances such as K-means (*19, 22*). This is because the underlying distribution of the covariance of key features that define the eco-provinces *are not* seen to inhabit shapes that are Gaussian. K-means, using Voronoi cells (straight lines), is not able to preserve non-Gaussian underlying distribution.

The 7 plankton functional group biomasses and the 4 nutrient fluxes form an 11 dimensional vector, $\mathbf{x}$. Thus, $\mathbf{x}$ is a vector field on the model grid, where each element $\mathbf{x}_i$ represents the 11 dimensional vector defined 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 log of the biomass data is used, and is discarded if a model grid cell has a biomass less than $1.2 \times 10^{-3} \mathrm{mgChl/m^3}$ or ice cover is over 70%. The data is normalized and standardized such that all data exist on the range [0 to 1], with the mean removed and scaled to unit variance. This is done so the features (biomass and nutrient fluxes) do not become conditioned by contrasts in the ranges of possible values. The clustering should capture the variational relationships from the key probabilistic distances between features rather the geographic distances. Quantifying these distances, important features emerge while unnecessary detail is discarded. In ecological terms, this is necessary because some phytoplankton types that have little biomass can have large biogeochemical impact, e.g. diazotrophs fixing nitrogen. The covariability of these types is highlighted when the data is standardized and normalized.

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 (*23*). This is a necessary step towards identifying robust clusters in the feature data, while avoiding non-convergent solutions (note S2). Using a Gaussian kernel, t-SNE preserves the statistical properties of the data by mapping each high-dimensional object onto a point in three-dimensional phase space in a way that ensures a high probability of similar objects being close in both the high and low-dimensional space (*24*). Given a set of N high-

dimensional objects x_1 ,..., x_N, the t-SNE algorithm performs a reduction by minimizing the Kullback-Leibler (KL) divergence (*25*). The KL divergence is a measure of how different one probability distribution is from a second reference probability distribution. Effectively, assessing the likelihood of association between a low dimensional rendition of 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, t-SNE defines a probability of similarity, p:

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

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 11 dimensional combined biomass and nutrient flux vector set to 3D. 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 dimensionality. The t-SNE method was found to be superior to PCA in delivering robust and reproducible results for the eco-provinces (see note S2). This is likely because the orthogonality assumption that underlies PCA is not appropriate for identifying key interactions between highly non-linearly interacting features, because PCA focuses on linear covariance structure (*26*). Using remotely sensed data, Lunga et. al (*27*) illustrates how complex and non-linear spectral features that depart from Gaussian distributions can be highlighted employing SNE methods.

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. The colors in Fig. 2a are the clusters found using the DBSCAN method (*28*). Looking for densely packed observations, the DBSCAN algorithm uses the distance in the 3 dimensional representation between points (ϵ=0.39, see materials and methods for a discussion of this choice), and the number of similar points needed to define a cluster (here 100 points, see above). The DBSCAN method makes no assumptions about the shapes or numbers of clusters in the data, as follows:

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 color. DBSCAN is a fast and scalable algorithm, with a worst-case performance of O(n^2), and is effectively not stochastic for the present analysis. Setting the minimum number of points was determined using expert assessment, with results not being robust within $\approx \pm 10$ after adjustment of the distance ϵ. This distance was set using the degree of connectiveness (Fig. 6a) and the percentage of ocean covered (Fig. 6b). The connectiveness

is defined as the resultant number of clusters, and is sensitive to the ϵ parameter. A low connectiveness indicates under-fitting, artificially grouping areas together. A high connectiveness indicates over-fitting. A higher minimum number could conceivably be used, but arriving at a robust solution would be unlikely if the minimum exceeds ca. 135 (see materials and methods for further details).

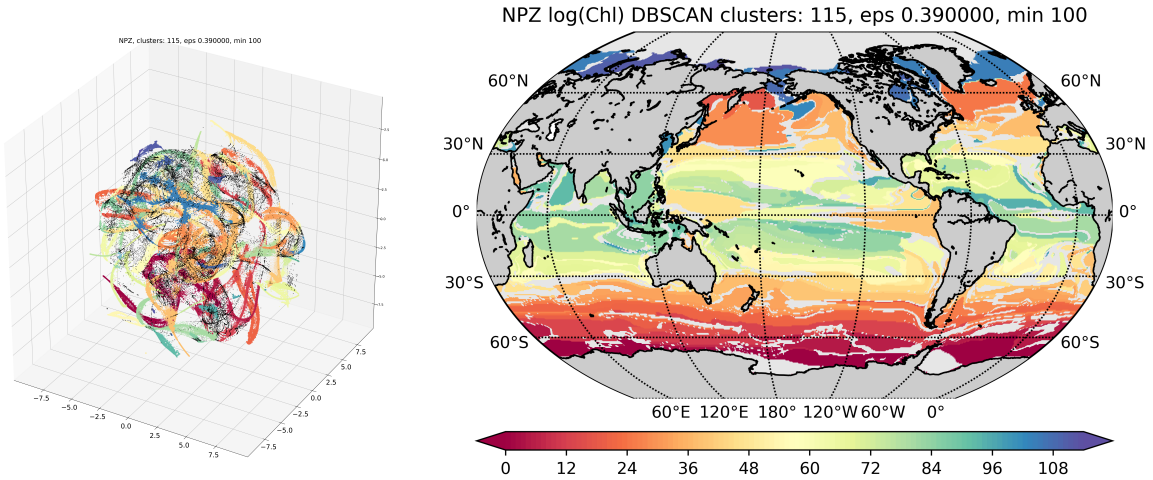
Back-projecting onto the globe

The 115 clusters identified in Fig. 2a are presented projected back onto the globe in Fig. 2b. Each color corresponds to a coherent combination of biogeochemical and ecological factors identified by DBSCAN. 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 colors of clusters the same as in Fig. 2a. Similar colors should not be interpreted as ecological similarity, as they are assigned by the order in which the algorithm discovers clusters.

Regions in Fig. 2b can be seen as qualitatively similar to established regions in the physics and/or biogeochemistry of the ocean. For example, the clusters in the Southern Ocean are zonally symmetric, oligotrophic gyres emerge, sharp transitions suggest the influence of trade winds, and distinct regions associated with upwelling are seen e.g. in the equatorial Pacific.

Ecological similarity: BC-dissimilarity

To understand the ecological context of the eco-provinces, the intra-cluster ecology is assessed using a variant on the Bray-Curtis Dissimilarity metric (BC, (*29*)). The BC metric is a statistic



(a) t-SNE projection with provinces in colour (b) Spatial representation provinces in Fig. 2a BC-dissimilarity

Figure 2: **Eco-provinces geographical and in t-SNE space.** The left Fig. 2a showing modelled nutrient supply rates, phytoplankton and zooplankton functional group biomass 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 majority of points captured as is demonstrated in the Fig. 6b. Axes refer to the "t-SNE" dimensions 1, 2 and 3. To the right Fig. 2b shows the geographical projection of the provinces discovered by DBSCAN onto the origin latitude longitude grid. Colours should be considered arbitrary but correspond to Fig. 2a.

used to quantify the community structure dissimilarity between two different sites. The BC metric is applied to the biomass of the 51 types of phyto- and zooplankton:

$$BC_{n_i n_j} = 1 - \frac{2C_{n_i n_j}}{S_{n_i} + S_{n_j}},$$

$BC_{n_i n_j}$ 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 types 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, but operates in a non-euclidean space which is likely better suited to ecological data and its 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 and upwelling regions are markedly different.

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 SAGE method is able to separate the 51 types of the model data according to ecological

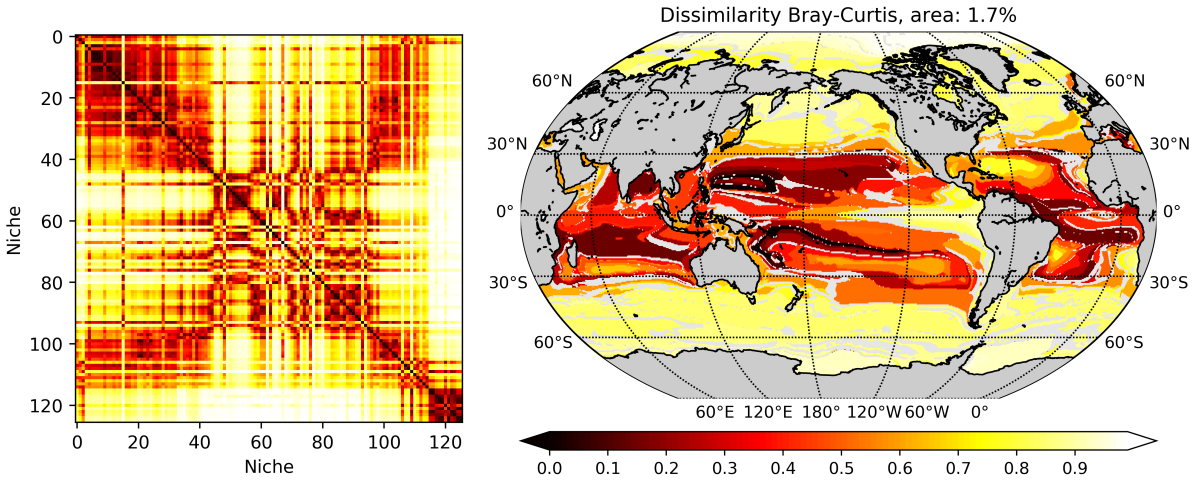
similarity. The global mean intra-cluster BC-dissimilarity for all 51 types is 0.102 ± 0.0049.

The equivalent Longhurst intra-province BC-dissimilarity is presented in Fig. 4b using the biomass of the 51 plankton types, with a global mean across provinces of 0.227, and a standard deviation across grid-points referenced to the province BC dissimilarity of 0.046. This is larger than for the clusters identified in Fig. 1b. Using the sum of the 7 functional groups instead, the mean intra-seasonal BC-dissimilarity of the Longhurst provinces increases to 0.232.

The maps of the global eco-provinces offer intricate detail of ecological interactions that are unique, and offer a refinement in terms of ecosystem structure over using Longhurst provinces. The eco-provinces are anticipated to provide insight into the processes controlling the numerical model ecosystem, and such insights could assist exploration of in situ efforts. For the purpose of this study, the over hundred provinces cannot be adequately showcased. The following section presents the SAGE method to aggregate provinces.

Defining Aggregated Eco-Provinces (AEPs)

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. 1b illustrates the nesting of ecologically similar provinces. Eco-provinces are grouped together in terms of their ecological similarity, and these grouping of provinces are called "Aggregated Eco-Provinces" (AEPs). An adjustable level of "complexity" is set in terms of the number of aggregated provinces that will be considered. The term "complexity" is used, as it allows the level of the emergent properties to be adjusted. For defining meaningful aggregation, the mean intra-province BC-dissimilarity from the Longhurst provinces of 0.227 is used as a benchmark below which the aggregated



(a) Symmetric BC matrix

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

Figure 3: **The eco-province Bray-Curtis Dissimilarity.** The left Fig. 3a Bray-Curtis Dissimilarity metric evaluated for **every province compared to every other** for the global surface 20 year mean of the 51 plankton biomasses. 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 a province in the oligotrophic gyre compared to every other for the global surface 20 year mean. Black (Bray-Curtis = 0) denotes an identical region, while white (Bray-Curtis = 1) denotes no similarity.

provinces are no longer considered useful.

The eco-provinces are coherent across the globe as Fig. 3b demonstrates. Some configurations are very "common", as seen using the inter-province BC-dissimilarity. Inspired by methods from genetics and graph theory, "connectivity graphs" are used to sort the > 100 provinces according to which province they are most similar to. The metric of "connectivity" here is determined using the inter-province BC-dissimilarity (*30*). The number of spatially larger provinces that the > 100 provinces can be sorted into is here referred to as the "complexity". The aggregated eco-provinces (AEPs) are the product of sorting the full > 100 provinces into this subset of the most dominant/highly connected eco-provinces; each eco-province is assigned to the 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 complexity can be anything from 1 to the full complexity from Fig. 2a. At low complexities the AEPs can become degenerate, due to the probabilistic dimensionality-reduction step (t-SNE). Degeneracy implies that the eco-provinces could be assigned to different AEPs between iterations, changing the geographical area covered. 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-province BC-dissimilarity assessed using the 51 plankton types is 0.198 $\pm$0.013, as seen in Fig. 4d. Using the sum of the 7 plankton functional groups, the mean intra-province BC-dissimilarity 2σ is instead 0.198

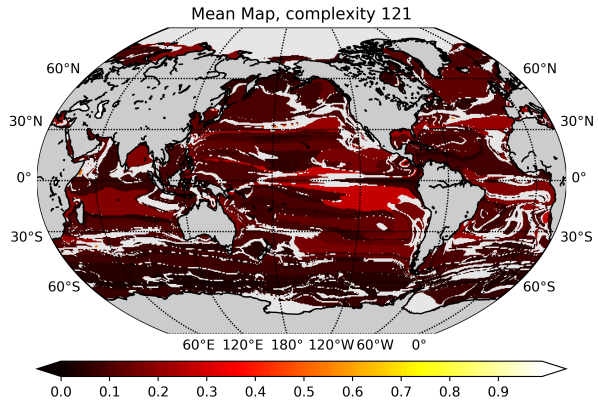
± 0.004. The comparison between the BC-dissimilarity computed with either the 7 functional group summed biomass or the full 51 plankton types' biomasses suggests the SAGE method is appropriate for the 51 dimensional case although it was trained on the biomass sum of the 7 functional groups.

Depending on the purpose of any study, a different level of complexity could be considered. A regional study might want the full complexity (i.e. all 115 provinces). As an example and for clarity, the lowest recommended complexity 12 is considered.

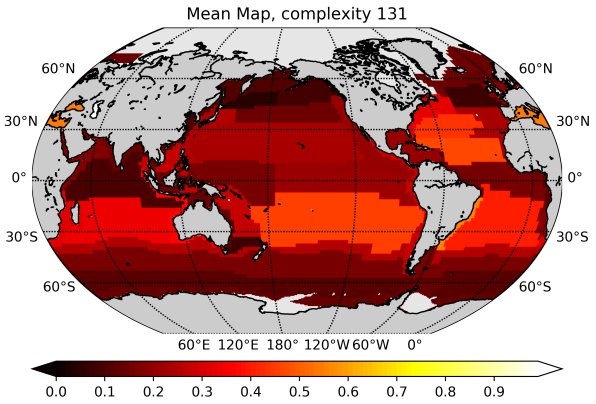
Utility of Aggregated Eco-Provinces: Community structure and their controls

As an example of the utility of the SAGE method, here the minimum complexity 12 AEPs are used to explore the controls on the emergent community structure. Fig. 5 illustrates the ecological insights grouped by AEPs (named 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, 1:1:16:16×10^3) are shown. For this latter panel, P is multiplied by 16 and Fe by 16×10^3 so the bars are comparable to the phytoplankton nutrient requirements.

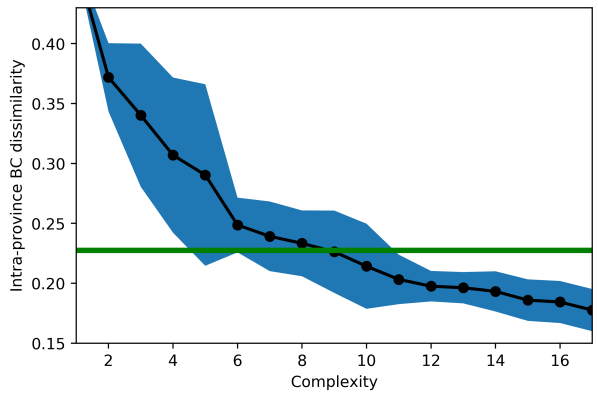
The identified AEPs are unique. There is some symmetry around the equator in the Atlantic and Pacific ocean, and similar, but augmented regions exist in the Indian ocean. Some AEPs hug the western sides of continents associated with upwelling regions. The Antarctic Circumpolar Current (ACC) is seen as a large zonal feature. The subtropical gyres stand out as complex series of oligotrophic AEPs. The familiar patterns of differences in biomass between



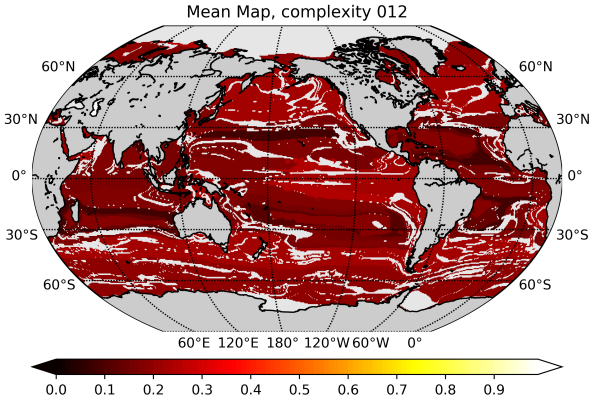
(a) Full complexity intra-province BC-dissimilarity



(b) Longhurst intra-province BC-dissimilarity



(c) Longhurst criterion



(d) Complexity 12 intra-province BC-dissimilarity

Figure 4: **Heuristic Processes to determine a minimum level of biogeochemical complexity.** For Fig. 4a, b and d 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 ± 0.117. This is the benchmark for the ecologically motivated sorting presented in this work (green line in Fig. 4c). The bottom left Fig. 4c shows the averaged intra-province BC-dissimilarity: 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 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 results in an intra-province BC-dissimilarity of 0.200 is reached as demonstrated in Fig. 4d.

picoplankton dominated oligotrophic gyres and diatom rich polar regions is apparent in these provinces.

AEPs with very similar total phytoplankton biomass can have very different community structure, and cover very different geographical areas, such as D,H and K which have similar total phytoplankton biomass. 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 has more diatoms, and fewer picoplankton than the other two provinces. It is notable that zooplankton biomass in the three regions is also very different, with AEP K having relatively low zooplankton abundance, but AEP D and H having relatively similar, higher, levels. Thus though their biomass (and hence also Chl-a) are similar, these provinces are very different: Chlorophyll based province detection would likely not capture these differences.

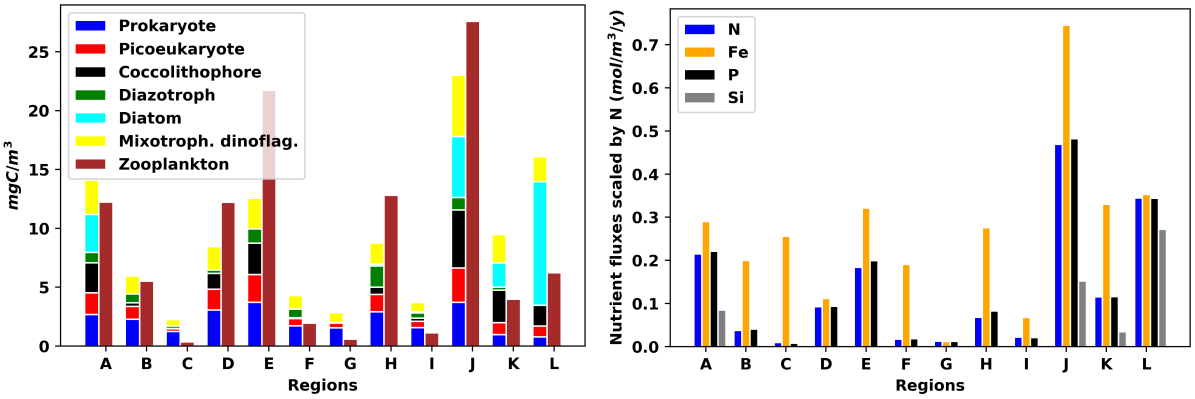
It is also apparent that some AEPs that have very different biomass can be similar in terms of their phytoplankton community structure. This is seen in AEP D and E for example. These are close to each other, notably in the Pacific, where AEP E is close to the highly productive AEP J. Again, there is not a clear connection between phytoplankton biomass and zooplankton abundance.

The AEPs can be understood in terms of the nutrient supplies to them (Fig. 5b). Diatoms only exist where there is sufficient silicic acid supply; generally the higher the silicic acid supply the higher the diatom biomass. Diatoms are seen in the AEPs A, J, K, and L. The proportion

of diatom biomass relative to other phytoplankton is dictated by how much N, P and Fe are supplied, relative to the diatoms demands. For instance AEP L is dominated by diatoms, and has the highest supply of Si relative to the other nutrients. In contrast, though more productive, AEP J has fewer diatoms and less Si supply (both total and relative to the other nutrients).

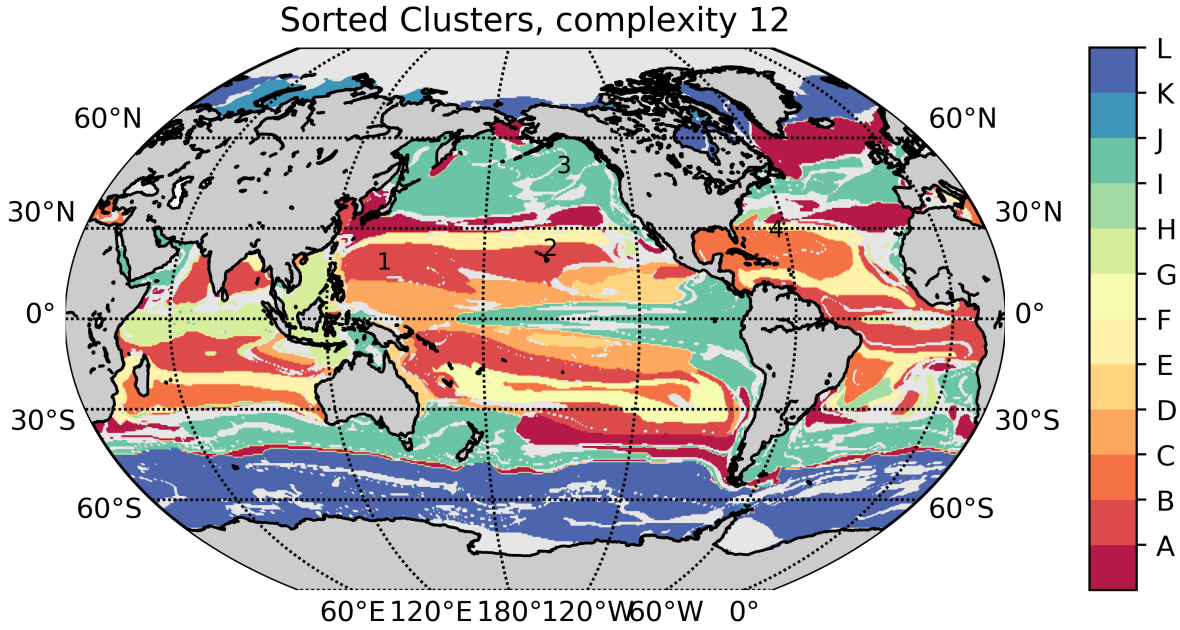
Diazotrophs have the ability to fix N, but also grow slowly (*31*). They coexist with other phytoplankton where there is an excess of Fe and P relative the the demands of the non-diazotrophs (*20, 21*). It is notable that there is higher diazotroph biomass where the amount of Fe and P supply are relatively large relative to the N supply. In this manner, the diazotroph biomass is larger in AEP H than in J, although the overall biomass in AEP J is higher. It is worth noting that AEP J and H are very different geographically, 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 unique ecosystem structure were not separated into provinces. SAGE generated AEPs facilitate the coherent and simultaneous comparison of the complicated and high-dimensional information from the ecosystem model. The AEPs effectively highlight why and where chlorophyll is not a good proxy for determining community structure, or abundance of zooplankton in higher trophic levels. A detailed analysis of the topic of an ongoing study beyond scope of this paper. The SAGE method provides a way to explore other mechanisms in the model in a more tractable way than looking from point to point.



(a) Ecological ensemble

(b) Nutrient fluxes



(c) Global provinces

Figure 5: **AEP interpretation for complexity 12.** Sorting the provinces into the 12 aggregated eco-provinces (AEPs) A to L. Top left Fig. 5a showing biomass (mgC/m^3) of the ecological ensemble in the 12 provinces. Top right Fig. 5b the nutrient fluxes rates (mmol/m^3/y) for dissolved inorganic nitrogen (N), Iron (Fe), phosphate (P) and silicic acid (Si). Fe and P are multiplied by 16 and 16×10^3 respectively, so that the bars are normalized to the phytoplankton stoichometric requirements. 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: SEATS, 2: ALOHA, 3: Station P, and 3: BATS.

Discussion and Conclusion

The SAGE method is presented, designed to help elucidate the overwhelmingly complicated ecological data from a global physical/biogeochemical/ecosystem numerical model. Eco-provinces are determined by summation of biomass across plankton functional groups, application of the t-SNE probabilistic dimensionality reduction algorithm, and clustering using the unsupervised ML method DBSCAN. An inter-province BC-dissimilarity/graph theory method for nesting is applied to arrive at robust AEPs, useful for global interpretation. Both the eco-provinces and AEPs are unique by construction. The AEP nesting can be adjusted between the full complexity of the original eco-provinces and the minimum recommended threshold of 12 AEPs. The nesting and determination of a minimal complexity for AEPs is seen as a crucial step, as the probabilistic t-SNE makes the $<$12 complexity AEPs degenerate. The SAGE method is global, and spans a complexity range from $>$100 AEPs to 12. For simplicity, the present focus is on the complexity 12 global AEPs. Future studies, particularly regional ones, could find a smaller spatial subset within the global eco-provinces useful, and potentially perform the aggregation within such a smaller region to leverage the same ecological insight as is discussed here. Suggestions are offered regarding how these eco-provinces, and insight gained from them, could be used to further ecological understanding, facilitate model inter-comparison, and potentially improve monitoring of marine ecosystems.

The eco-province and AEPs that the SAGE method identified are based on data from a numerical model. Numerical models are by definition simplified constructs that attempt to capture the essence of a target system, and different models can vary in their plankton distributions. The numerical model used in this study does not fully capture some of the patterns observed (e.g. in Chl estimates of the equatorial regions and Southern ocean). Capturing only a fraction

of the diversity in the real ocean, and not resolving the meso- and sub-mesoscale, likely impact nutrient fluxes and smaller scale community structures. Despite these shortcomings, AEPs are shown to be useful in helping to understand the complex model. The AEPs offer a potential numerical model inter-comparison tool, by assessing where similar ecological provinces are found. The present numerical model captures gross patterns of remotely sensed phytoplankton Chl-a concentrations, and distributions of plankton size and functional groups (note S1 and Fig. S1 (*2, 32*)).

The AEPs fit into oligotrophic versus mesotrophic regions as indicated by the 0.1mg Chl-a/m^{-3} contour (Fig. S1b): AEPs B, C, D, E, F, G are oligotrophic, and the remainder are in regions of higher Chl-a. The AEPs show some correspondence to the Longhurst provinces (Fig. S3a), for example the Southern Ocean and equatorial Pacific. In some regions the AEPs cover several Longhurst regions, and visa versa. Since the intent of the delineation of provinces here and in Longhurst are not the same, differences are anticipated. Multiple AEPs within a single Longhurst provinces suggest that some regions with similar biogeochemistry may have very different ecosystem structure. The AEPs show some correspondence to physical regimes as revealed using unsupervised learning (*19*), such as in high upwelling regimes (e.g. Southern Ocean and the Equatorial Pacific, Fig. S3c,d). Such correspondences suggest where plankton community structure is strongly influenced by ocean dynamics. In regions such as the North Atlantic, AEPs cross through physical provinces. Mechanisms causing these discrepancies could include processes such as dust delivery leading to very different nutrient regimes even within a similar physical regime.

The eco-provinces and AEPs suggest that using chlorophyll alone is not able to identify ecological composition, as is already appreciated by the marine ecological community. This is

seen in AEPs with similar biomass but markedly different ecological composition in (e.g. D and E). In contrast, AEPs such as D and K have very different biomass but similar ecological composition. The AEPs emphasize that the relationship between biomass, ecological composition and zooplankton abundance is complex. For example, while AEP J stands out in terms of both high phytoplankton and zooplankton biomass, AEP's A and L have similar phytoplankton biomass but A has much higher zooplankton abundance. The AEPs highlight that phytoplankton biomass (or Chl) cannot be used to predict zooplankton biomass. Zooplankton are the base of the food-chain for fisheries, and more accurate estimates could lead to better resource management. Future ocean colour satellites (e.g. PACE) might be better positioned to help estimate phytoplankton community structure. Using AEP predictions, estimates of zooplankton from space could potentially be facilitated. Methods like SAGE, together with new technology, as well as the increasing in situ data available (e.g. TARA and follow on studies) for ground-truthing, could together provide a step toward satellite based monitoring of the health of an ecosystem.

The SAGE method provides a convenient way to assess some of the mechanisms that control the features in the provinces e.g. biomass/chlorophyll, net primary production and community structure. For example, the relative amount of diatoms is set by the imbalance in the Si to N,P and Fe supplies relative to the phytoplankton stoichiometric requirements. With balanced supply rates, communities are diatom dominated (L) and where supply rates are less balanced (i.e. with lower Si supply relative to the diatoms nutrient demands) diatoms comprise only a smaller fraction (K). Diazotrophs thrive where the Fe and P supplies are in excess of the N supplies (e.g. E and H). Explorations of controlling mechanisms are made significantly more useful through the context provided by AEPs.

The eco-provinces and AEPs are regions of similar community structures. A time-series from a location within one eco-province or AEP could be seen as a point of reference, and representative of the area covered by the eco-province or AEP. Long term in situ monitoring stations offer such time-series. Long term in situ data-sets will continue to be invaluable, and the SAGE method could be seen as a method to help determine locations where new sites would be most useful from the perspective of monitoring community structure. For example, the time-series from ALOHA is in AEP B (Fig. 5c, label 2), in an oligotrophic region. Because ALOHA is close to the boundary to another AEP, the time-series may not be representative of the whole region, as previously suggested (*33*). Within the same AEP B, the time-series SEATS is south-west of Taiwan (*34*), further from the boundaries of other AEPs (Fig. 5c, label 1), and could serve as a better location within which to monitor AEP B. The BATS time-series in AEP C (Fig. 5c, label 4) is very close to the border of AEPs C and F, suggesting that monitoring AEP C using the BATS time series directly could be problematic. The P Station (Fig. 5c, label 3) in AEP J is quite far from an AEP boundary, and could therefore be more representative. The eco-provinces and AEPs could help establish a monitoring framework suitable for assessing global change, as the provinces allow assessment of where in situ sampling could offer key insight. The SAGE method can be developed further for application to climatological data to assess temporal province variability.

The success of the SAGE method is achieved through careful application of data science/ML methods, together with domain specific knowledge. Specifically, dimensionality reduction is performed using t-SNE, retaining the covariance structure of the high dimensional data, and facilitating visualising the covariance topology. The data is arranged in streaks and sheets of covariance (Fig. 2a), clearly indicating that purely distance based metrics such as k-means are inappropriate as they often assume a Gaussian (round) underlying distribution (discussed in note

S2). The DBSCAN method is appropriate for arbitrary covariance topologies, offering robust identification provided careful attention is given to setting parameters. The t-SNE algorithm is computationally costly, limiting its present application to larger data-volumes, meaning that application to depth or time varying fields is difficult. Work on the scalability of t-SNE is ongoing. The t-SNE algorithm has the potential to scale well in the future, as the Kullback-Leibler distance is readily parallelisable (*35*). Alternative promising methods of dimensionality reduction that to date scale better include the Uniform Manifold Approximation and Projection (UMAP) technique, but evaluation in the context of oceanographic data is necessary. Implications of better scalability would be classification e.g. over the mixed layer, for global climatologies or models with varying complexity. The regions that fail to be classified within any province by SAGE can be seen as the remaining black dots in Fig. 2a. Geographically, these regions are largely in highly seasonal areas, suggesting that capturing the time evolving eco-provinces would provide better coverage.

To construct the SAGE method, ideas from complex systems/data science have been leveraged. Exploiting the ability to determine clusters of functional groups (high probability of close proximity in an 11 dimensional space), and determine provinces. These provinces delineate a specific volume in our 3 dimensional t-SNE phase space. Similarly, Poincaré sections can be used to assess the ”volume” of state space occupied by a trajectory, in order to determine ”regular” or ”chaotic” behaviour (*36*). For the static 11 dimensional model output, the volume occupied after data is cast into a 3 dimensional phase space could be interpreted similarly. The relation between geographical area and the area in 3 dimensional phase space is not simple, but can be interpreted in terms of ecological similarity. The more conventional BC-dissimilarity metric was preferred for this reason.

Future work will repeat the SAGE method for seasonally varying data, to assess the spatial variability in the identified provinces and AEPs. A future goal is to leverage this method to help determine which provinces could be determined by satellite measurements such as Chl-a, remotely sensed reflectance, sea-surface temperature etc. This would allow remote sensing assessments of ecological composition and highly agile monitoring of the eco-provinces and their variability.

Materials and methods

The purpose of this study is to present the SAGE method for defining eco-provinces by their distinct plankton community structure. Here more detail is provided on the physical/biogeochemical/ecosystem model, as well on parameter selection for t-SNE and DBSCAN algorithms.

Model framework

The physical component of the model comes from the Estimating the Circulation and Climate of the Ocean (ECCOv4, (*37*)) global state estimate described by (*38*). 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, (*39*)) that is optimized to track observations.

The biogeochemical/ecosystem is described more fully (i.e. equations and parameter values) in (*2*). The model captures the cycling of C, N, P, Si and Fe through inorganic and organic pools. The version used here includes 35 phytoplankton: 2 Pico-prokaryotes and 2 Pico-eukaryotes

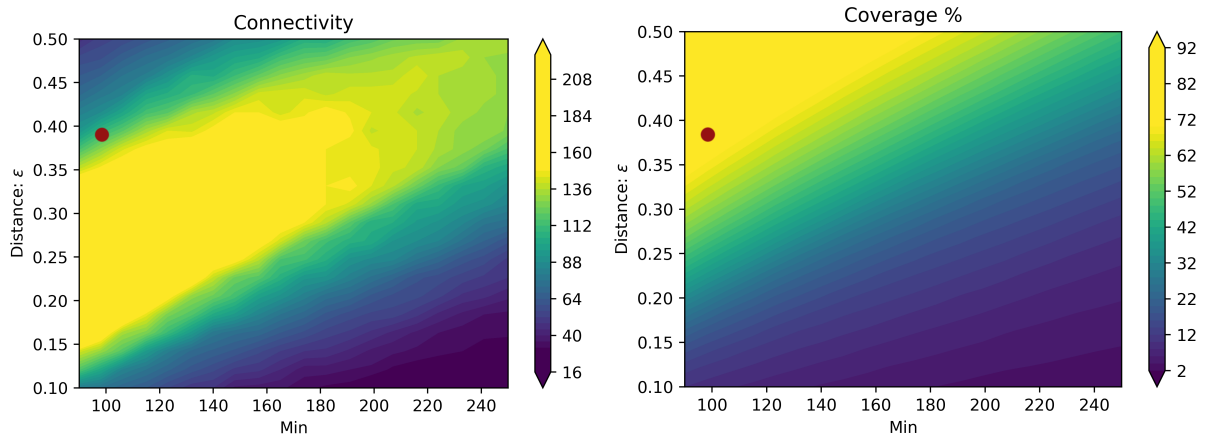
(that are adapted to low nutrient environments), 5 coccolothophores (that have calcium carbonate coverings), 5 diazotrophs (that fix nitrogen gas, and thus are not limited by availability of dissolved inorganic nitrogen), 11 diatoms (that form silicous coverings), 10 mixotrophic dinoflagellates (that can both photosynthesize and graze other plankton), and 16 zooplankton (which graze on other plankton). These are referred to as "biogeochemical functional groups" as they each impact the biogeochemistry of the ocean differently (*40, 41*) and are frequently used in observational and modelling studies. In this model each functional group is made up of several plankton of different sizes spanning 0.6μm to 2500 μm equivalent spherical diameter.

Parameters influencing phytoplankton growth, grazing, and sinking are related to size, with specific differences between the 6 phytoplankton functional groups (*32*). Results from this 51 plankton component of the model has been used in several recent studies (*42–44*), 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, the 20 year mean of these outputs was used as the input for the eco-provinces. Distribution of Chl, plankton biomass, nutrient concentrations, as well as distributions of functional groups compare well with satellite and in situ observations (see (*2, 44*), and note S1, and Fig. S1-3).

Parameter selection for t-SNE and DBSCAN

For the SAGE method, the main source of stochasticity comes from the t-SNE step. Stochasticity can hinder reproducibility, meaning that results are not robust. The SAGE method uses a



(a) Connectiveness (b) Coverage

Figure 6: **Setting the DBSCAN parameters.** Setting the parameters for t-SNE the resultant number of found clusters is used as a measure of the connectiveness (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.

stringent test of robustness, by identifying one set of parameters for t-SNE and DBSCAN that consistently identify clusters when repeated. Determining t-SNE parameter "perplexity" can be understood as determining the degree to which the mapping from high to low dimensionality should respect local or more global features of the data. A perplexity of 400 and 300 iterations was arrived at.

For the clustering algorithm DBSCAN, the minimum size of the data points within a cluster, and the distance metric ϵ need to be determined. The minimum number is set using expert guidance with knowledge of what is appropriate for the present numerical modeling framework and resolution, a minimum number of 100 was set. A higher minimum number (ca. < 135, before the upper band of green widens) could conceivably be used, but would be not be able to act as a substitute for the aggregation method based on the BC-dissimilarity. The degree

of connectiveness (Fig. 6a) is used to set the ϵ parameter, favouring a higher coverage (Fig. 6b). The connectiveness is defined as the resultant number of clusters, and is sensitive to the ϵ parameter. A low connectiveness indicates under-fitting, artificially grouping areas together. A high connectiveness indicates over-fitting. Over fitting is problematic also because it indicates that the initial stochastic guess can lead to results that are not reproducible. Between these two extremes there is a drastic increase (often referred to as an "elbow"), indicating the optimal ϵ. 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 ($\epsilon < 0.42$), or most of the ocean is not classified and is deemed as noise ($\epsilon > 0.99$). 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". This is the optimal region, where robust clusters can be identified, as determined using the intra-province BC-dissimilarity, despite the probabilistic t-SNE. Using Fig. 6a and 6b ϵ was set to 0.39. With a larger minimum number, arriving at an ϵ that allows robust classification would be unlikely, with values > 135 seen to have a wider green region. The widening of this region suggests that the "elbow" will be more difficult to find, or absent.

References

1. R. Bailey, *Ecoregions: The Ecosystem Geography of the Oceans and Continents* (Springer New York, 2014).

2. S. Dutkiewicz, A. E. Hickman, O. Jahn, W. W. Gregg, C. B. Mouw, M. J. Follows, Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model. *Biogeosciences* **12**, 4447–4481 (2015).

3. M. D. Spalding, H. E. Fox, G. R. Allen, N. Davidson, Z. A. Ferdaña, M. Finlayson, B. S. Halpern, M. A. Jorge, A. Lombana, S. A. Lourie, K. D. Martin, E. McManus, J. Molnar, C. A. Recchia, J. Robertson, Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. *BioScience* **57**, 573-583 (2007).

4. J. M. Omernik, G. E. Griffith, Ecoregions of the conterminous united states: Evolution of a hierarchical spatial framework. *Environmental Management* **54**, 1249–1266 (2014).

5. A. Longhurst, S. Sathyendranath, T. Platt, C. Caverhill, An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research* **17**, 1245-1271 (1995).

6. L. Gloege, G. A. McKinley, C. B. Mouw, A. B. Ciochetto, Global evaluation of particulate organic carbon flux parameterizations and implications for atmospheric pco2. *Global Biogeochemical Cycles* **31**, 1192-1215 (2017).

7. J. Roff, *Marine Conservation Ecology* (Taylor & Francis, 2013).

8. R. Watson, D. Pauly, V. Christensen, R. Froese, A. Longhurst, T. Platt, S. Sathyendranath, K. Sherman, P. Celone, *In: Trends in Exploitation, Protection, and Research (eds* (2003), pp. 375–395.

9. IOCCG, *IOCCG report 9: Partition of the Ocean into Ecological Provinces: Role of Ocean-Colour Radiometry* (2009), pp. Dowell, M. and T. Platt (eds.).

10. T. Hattab, F. Lasram, C. Albouy, C. Sammari, R. Ms, P. Cury, F. Leprieur, F. Le Loc'h, The use of a predictive habitat model and a fuzzy logic approach for marine management and planning. *PloS one* **8**, e76430 (2013).

11. M. J. Costello, P. Tsai, P. S. P. Wong, A. K. L. Cheung, Z. Basher, C. Chaudhary, *Nature Communications* (2017).

12. M. T. Kavanaugh, B. Hales, M. Saraceno, Y. H. Spitz, A. E. White, R. M. Letelier, Hierarchical and dynamic seascapes: A quantitative framework for scaling pelagic biogeochemistry and ecology. *Progress in Oceanography* **120**, 291 - 304 (2014).

13. M. J. Oliver, A. J. Irwin, Objective global ocean biogeographic provinces. *Geophysical Research Letters* **35** (2008).

14. G. Reygondeau, A. Longhurst, E. Martinez, G. Beaugrand, D. Antoine, O. Maury, Dynamic biogeochemical provinces in the global ocean. *Global Biogeochemical Cycles* **27**, 1046-1058 (2013).

15. C. Mouw, N. Hardman-Mountford, S. Alvain, A. Bracher, B. Robert, A. Bricaud, A. Ciotti, E. Devred, F. Amane, T. Hirata, H. Toru, T. Kostadinov, S. Roy, J. Uitz, A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global ocean. *Frontiers in Marine Science* **4**, 41 (2017).

16. G. Lima-Mendez, *et al.*, Determinants of community structure in the global plankton interactome. *Science* **348** (2015).

17. E. Buitenhuis, M. Vogt, R. Moriarty, N. Bednaršek, S. Doney, K. Leblanc, C. Le Quéré, Y.-W. Luo, C. O'Brien, T. O'Brien, J. Peloquin, R. Schiebel, C. Swan, Maredat: towards a world atlas of marine ecosystem data. *Earth System Science Data* **5**, 227-239 (2013).

18. A. Longhurst, *Ecological Geography of the Sea*, Agricultural and Biological Sciences (Academic Press, 1998).

19. M. Sonnewald, C. Wunsch, P. Heimbach, Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science* **6**, 784-794.

20. B. A. Ward, S. Dutkiewicz, C. M. Moore, M. J. Follows, Iron, phosphorus, and nitrogen supply ratios define the biogeography of nitrogen fixation. *Limnology and Oceanography* **58**, 2059-2075.

21. S. Dutkiewicz, B. A. Ward, F. Monteiro, M. J. Follows, Interconnection of nitrogen fixers and iron in the pacific ocean: Theory and numerical simulations. *Global Biogeochemical Cycles* **26**.

22. G. Maze, H. Mercier, R. Fablet, P. Tandeo, M. L. Radcenco, P. Lenca, C. Feucher, C. L. Goff, Coherent heat patterns revealed by unsupervised classification of argo temperature profiles in the north atlantic ocean. *Progress in Oceanography* **151**, 275 - 292 (2017).

23. D. Marmanis, M. Datcu, T. Esch, U. Stilla, Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters* **13**, 105-109 (2016).

24. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

25. S. Ghosh, K. P. Burnham, N. F. Laubscher, G. E. Dallal, L. Wilkinson, D. F. Morrison, M. W. Loyer, B. Eisenberg, S. Kullback, I. T. Jolliffe, J. S. Simonoff, Letters to the editor. *The American Statistician* **41**, 338–341 (1987).

26. J. M. Lewis, P. M. Hull, K. Q. Weinberger, L. K. Saul, Mapping uncharted waters: Exploratory analysis, visualization, and clustering of oceanographic data.

27. D. Lunga, S. Prasad, M. M. Crawford, O. Ersoy, Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine* **31**, 55-66 (2014).

28. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96 (AAAI Press, 1996), pp. 226–231.

29. J. Bray, J. Curtis, An ordination of upland forest communities of southern wisconsin.-ecological monographs. *Ecological Monographs* **27**, 325–349 (1957).

30. M. Costanzo, B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons, G. Tan, W. Wang, M. Usaj, J. Hanchard, S. D. Lee, *et al.*, A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).

31. I. Berman-Frank, A. Quigg, Z. V. Finkel, A. J. Irwin, L. Haramaty, Nitrogen-fixation strategies and fe requirements in cyanobacteria. *Limnology and Oceanography* **52**, 2260-2269 (2007).

32. S. Dutkiewicz, P. Cermeno, O. Jahn, M. J. Follows, A. E. Hickman, D. A. A. Taniguchi, B. A. Ward, Dimensions of marine phytoplankton diversity. *Biogeosciences Discussions* **2019**, 1–46 (2019).

33. M. T. Kavanaugh, M. J. Church, C. O. Davis, D. M. Karl, R. M. Letelier, S. C. Doney, 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 (2018).

34. D. Karl, N. Bates, S. Emerson, P. Harrison, C. Jeandel, O. Llinás, K.-K. Liu, J.-C. Marty, A. Michaels, J. Miquel, S. Neuer, Y. Nojiri, Temporal studies of biogeochemical processes determined from ocean time-series observations during the jgofs era. *Ocean Biogeochemistry* (2003).

35. L. Van Der Maaten, Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).

36. M. Henon, C. Heiles, The Applicability of the Third Integral of Motion: Some Numerical Experiments. *Astron. J.* **69**, 73-79 (1964).

37. G. Forget, J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, C. Wunsch, Ecco version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development* **8**, 3071–3104 (2015).

38. C. Wunsch, P. Heimbach, *Ocean Circulation and Climate*, G. Siedler, S. M. Griffies, J. Gould, J. A. Church, eds. (Academic Press, 2013), vol. 103 of *International Geophysics*, pp. 553 – 579.

39. A. Adcroft, C. Hill, J.-M. Campin, J. Marshall, P. Heimbach, *Proceedings of the ECMWF seminar series on Numerical Methods, Recent developments in numerical methods for atmosphere and ocean modelling* (2004), pp. 139–149.

40. E. T. Buitenhuis, T. Hashioka, C. L. Quéré, Combined constraints on global ocean primary production using observations and models. *Global Biogeochemical Cycles* **27**, 847-858 (2013).

41. R. R. Hood, E. A. Laws, R. A. Armstrong, N. R. Bates, C. W. Brown, C. A. Carlson, F. Chai, S. C. Doney, P. G. Falkowski, R. A. Feely, M. A. M. Friedrichs, M. R. Landry, J. Keith Moore, D. M. Nelson, T. L. Richardson, B. Salihoglu, M. Schartau, D. A. Toole, J. D. Wiggert, Pelagic functional group modeling: Progress, challenges and prospects. *Deep Sea Research Part II: Topical Studies in Oceanography* **53**, 459-512 (2006).

42. P. Tréguer, C. Bowler, B. Moriceau, S. Dutkiewicz, M. Gehlen, O. Aumont, L. Bittner, R. Dugdale, Z. Finkel, D. Iudicone, O. Jahn, L. Guidi, M. Lasbleiz, K. Leblanc, M. Levy, P. Pondaven, Influence of diatom diversity on the ocean biological carbon pump. *Nature Geoscience* (2017).

43. E. L. McParland, N. M. Levine, The role of differential dmsp production and community composition in predicting variability of global surface dmsp concentrations. *Limnology and Oceanography* **64**, 757-773.

44. A. Kuhn, S. Dutkiewicz, O. Jahn, S. Clayton, T. Rynearson, M. Mazloff, A. Barton, Temporal and spatial scales of correlation in marine phytoplankton communities. *Journal of Geophysical Research: Oceans* **n/a**.

45. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

46. T. Moore, J. Campbell, M. Dowell, A class based approach for characterizing the uncertainty of the modis chlorophyll product. *Remote Sensing of Environment - REMOTE SENS ENVIRON* **113**, 2424-2430 (2009).

47. S. Clayton, S. Dutkiewicz, O. Jahn, C. Hill, P. Heimbach, M. Follows, Biogeochemical versus ecological consequences of modeled ocean physics. *Biogeosciences* **14**, 2877-2889 (2017).

48. M. Szeto, J. Werdell, T. Moore, J. Campbell, Are the world's oceans optically different? *Journal of Geophysical Research* **116** (2011).

49. R. Johnson, P. Strutton, S. Wright, A. McMinn, K. Meiners, Three improved satellite chlorophyll algorithms for the southern ocean. *Journal of Geophysical Research: Oceans* **118** (2013).

50. C. O'Brien, J. Peloquin, M. Vogt, M. Heinle, N. Gruber, P. Ajani, H. Andruleit, J. Aristegui, L. Beaufort, M. Estrada, D. Karentz, E. Kopczyńska, R. Lee, A. Poulton, T. Pritchard, C. Widdicombe, Global marine plankton functional type biomass distributions: Coccolithophores. *Earth System Science Data* **5**, 259-276 (2013).

51. I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society of London Series A* **374**, 20150202 (2016).

52. A. Hannachi, I. Jolliffe, D. Stephenson, Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology* **27** (2007).

53. H. Akaike, Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (1971).

54. A.-K. Seghouane, S.-i. Amari, The aic criterion and symmetrizing the kullback–leibler divergence. *Neural Networks, IEEE Transactions on* **18**, 97 - 106 (2007).

55. J. Dziak, D. Coffman, S. Lanza, R. Li, L. Jermiin, Sensitivity and specificity of information criteria (2018).

56. C. Ding, X. He, K-means clustering via principal component analysis. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004* **1** (2004).

Acknowledgements

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. All authors declare that they have no competing interests.

The AEP data is available with DOI: 10.5281/zenodo.3579371

The nutrient flux and plankton data used in this study is available with DOI: 10.5281/zenodo.3579369

The grid file is available with DOI: 10.5281/zenodo.3697102

A visualization example of varying AEP complexity is available: `github.com/maikejulie/plottingAEPs/`

This work was supported by grant NASA-IDS (80NSSC17K0561), ECCO Consortium funding via the Jet Propulsion Laboratory. SD and CH were also supported by the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems (CBIOMES) (grant no. 549931). The code fort the scaling, t-SNE and DBSCAN are from the python library scikit-learn (*45*), version 0.22.1. The physical fields are available online (`https://ecco.jpl.nasa.gov`) and documentation is available: `http://doi.org/10.5281/zenodo.2533351`

Supplementary Materials

Note S1: Model Evaluation

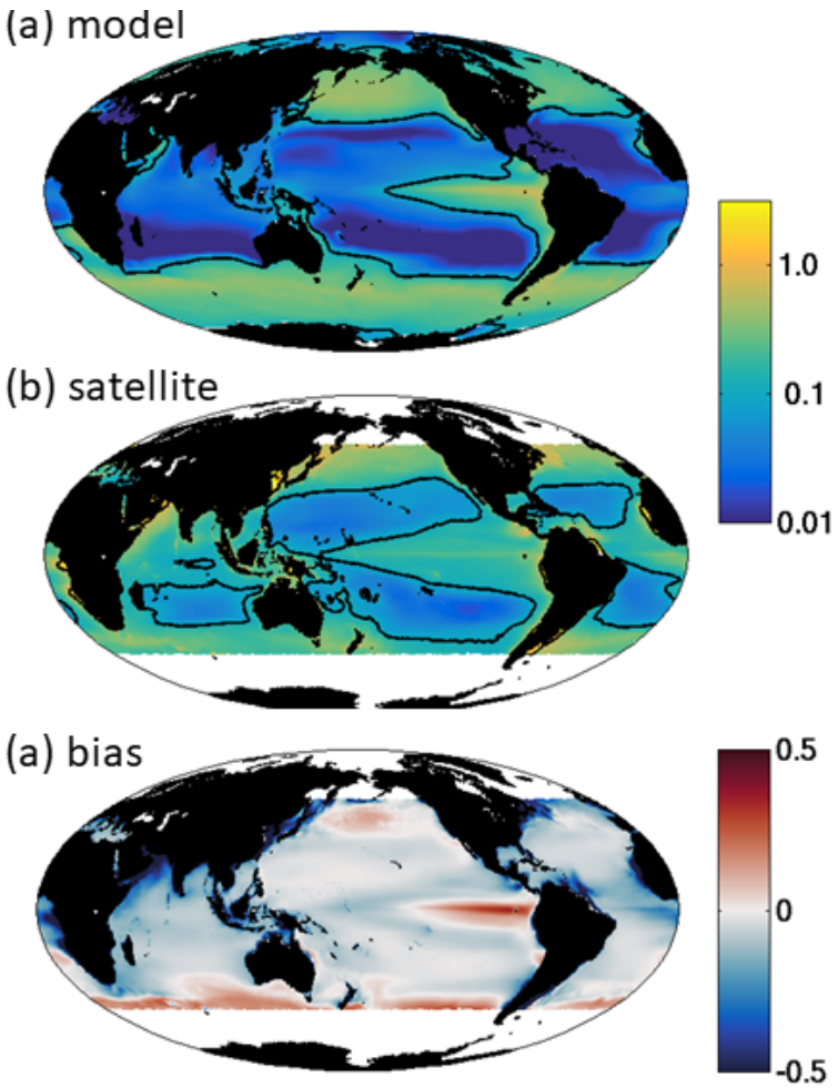
The ecosystem model has been used in various ecological configuration and within different physical frameworks (*2, 32, 42, 43*). These studies have provided extensive evaluation against satellite and available in situ data. An evaluation is presented of the model output used in this study against satellite Chl-a, and against the in situ compilation of phytoplankton functional group biomass from the MAREDAT data-set (*40*).

Comparing the model annual climatological surface (0-10m) Chl-a, to Ocean Colour Climate Change Initiative project (OC-CCI, 1998-2015) estimates of Chl-a (Fig. S1), the model is seen to capture the patterns of high Chl-a in both subpolar and equatorial upwelling regions, and captures low Chl-a in subtropical gyres. Only data for regions with full annual coverage are shown (optical satellite sensors do not capture a signal in the polar winters). Note that the satellite estimates have non-negligible uncertainties associated with them (e.g. estimates have more than 35% errors (*46*)). The spatial resolution of the Darwin model does not capture important physical processes near coastlines, and lack of sedimentary and terrestrial supplies of nutrients and organic matter lead to Chl-a being too low in these regions. Chl-a is under-estimated by the model in the subtropical gyres, likely due to lack of mesoscale processes in the model that would supply additional nutrients in these regions (see e.g. (*47*)). The model Chl-a is higher than the satellite estimates in the Southern Ocean. There are likely regional biases in the satellite algorithms, these are potentially enhanced in Southern Ocean signals e.g. (*48, 49*). The model is also higher in the equatorial Pacific, likely due to insufficient iron limitation in this region. The model underestimates the Chl-a in the Atlantic Equatorial region.

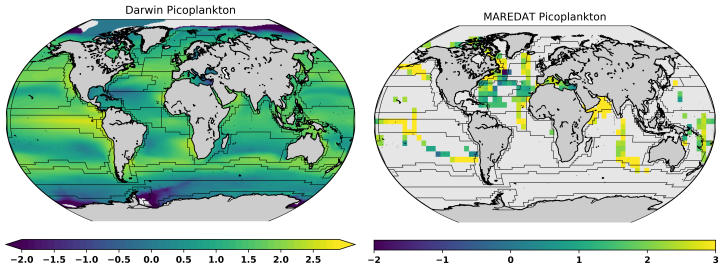
The numerical model functional group distribution is compared to the latest compilation of observations (Fig. S2, (*40*), and references therein). Observations are sparse both temporally and spatially, and were averaged into 5 degree bins to facilitate visual comparison. Visual evaluation suggests that even with the spatial and temporal data taken into account overall features are captured: the ubiquitous nature of the pico-phytoplankton, the limited domain of the diazotrophs (including observed lack of diazotrophs in the South Pacific gyre), the pattern of enhanced diatom biomass in high latitudes, and low biomass in subtropical gyres. The model underestimates diazotrophy in the western equatorial Atlantic, possibly due to lack of riverine influx of nutrients/organic matter in this region. Coccolithophore biomass is overestimated relative to MAREDAT in many regions, but note that the conversion from cells to biomass in that compilation was estimated to have uncertainties of several 100% (*50*).

Note S2: PCA and k-means methods in the presence of non-Gaussian covariance structures

The interactions between the types and nutrient fluxes in the feature vector in this study are highly non-linear, implying an underlying distribution of the covariance structures are not Gaussian. The ultimate goal of clustering algorithms is to arrive at a statistical model that approximates the ”true” model from which the available data has been drawn. The underlying distribution of the data's covariance structure has implications for which clustering algorithm is appropriate, because these are designed to ”identify” different types of underlying distributions. A Gaussian covariance would manifest as a ”round” shape. A more complex distribution will have a correspondingly complex ”shape”. Most clustering algorithms are well suited for data with an underlying Gaussian covariance distribution, but a highly non-Gaussian distribution of-

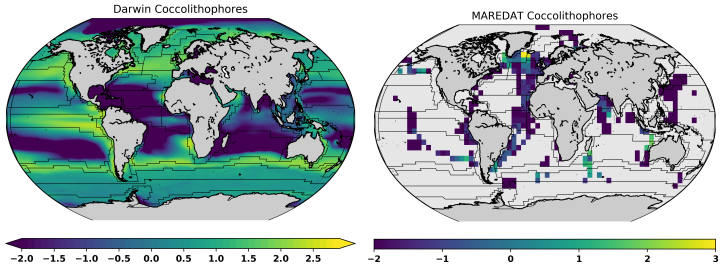


S1: Model surface Chl-a comparison to satellite data. The model annual climatological surface (0-10m) Chl-a (top), and Ocean Colour Climate Change Initiative project (OC-CCI, 1998-2015) (middle), and the bias (bottom). The model captures the patterns of high Chl-a in the subpolar regions and along the equatorial upwelling and low Chl-a in the subtropical gyres. Contours indicates 0.1 mg Chl/m^{-3}



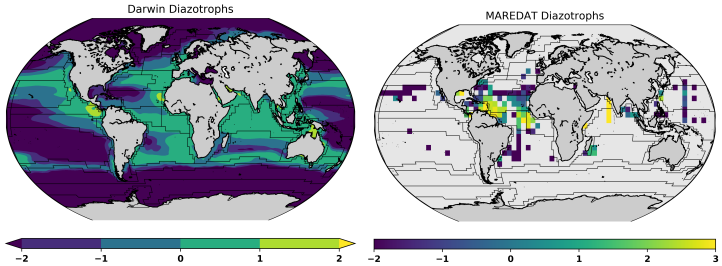
(a) Pico

(b) Pico MAREDAT



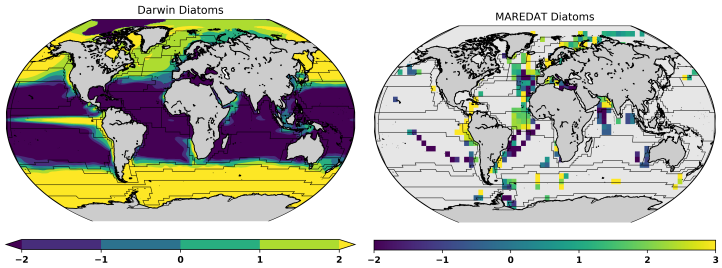
(c) Cocco

(d) Cocco MAREDAT



(e) Diaz

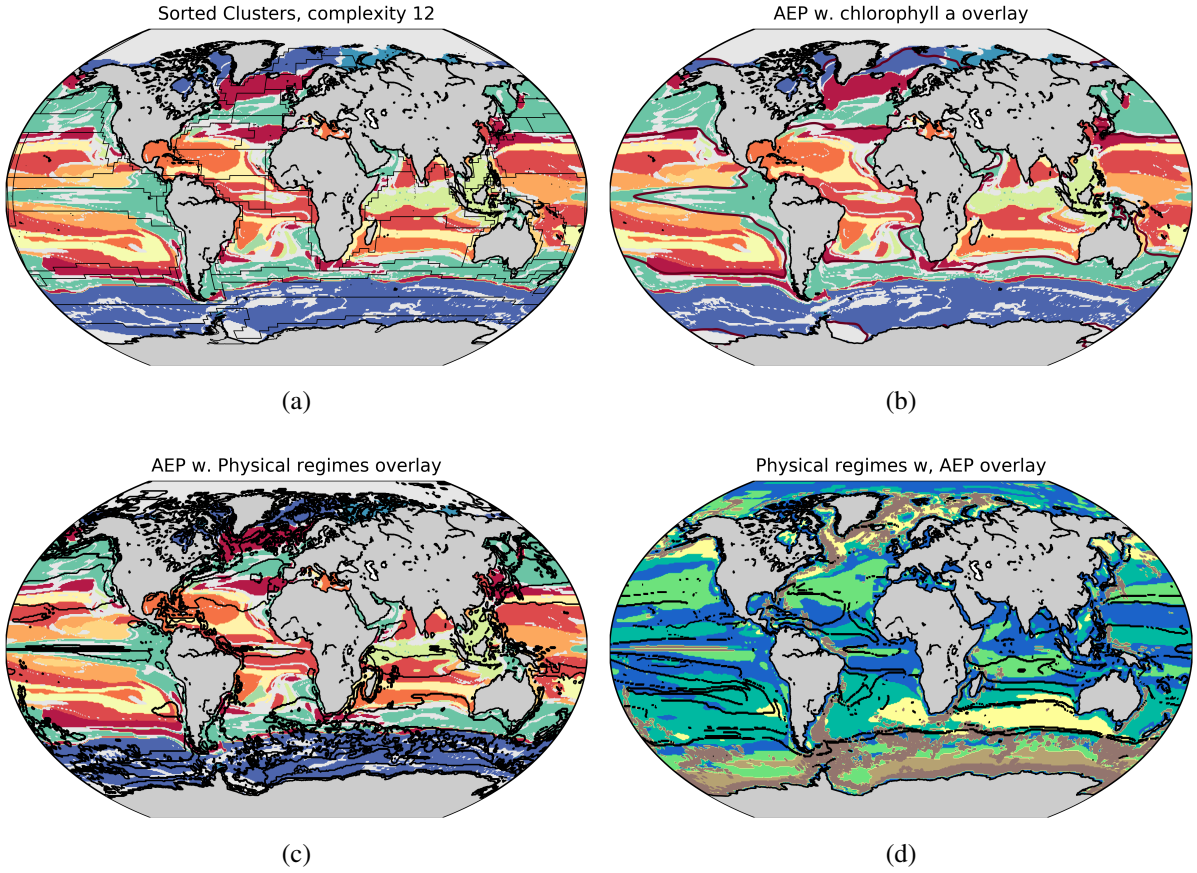
(f) Diaz MAREDAT



(g) Diatom

(h) Diatom MAREDAT

S2: Comparison of model and observational phytoplankton functional group biomass. Phytoplankton functional group biomass (mg C/m^{-3}) from the numerical model (a, c, e, g) and MAREDAT (b, d, f, h) (*40*), biomass on log scale.



S3: Further context for the AEP of complexity 12. Comparison of the AEP complexity 12 to Longhurst (a), to the Chl-a 0.1 contour from the numerical model (b, see Fig S1), to the physical regimes in (*19*) (c), and a select number of physical regimes are overlaid (black contours) onto the AEP complexity 12 (d). Contours in c and d are overlaid as visual aids, and not all regions are shown.

ten requiring a more specialized approach. The choice of algorithm needs to be tailored to the data (e.g. DBSCAN in this study), and the results verified and validated to the extent that this is possible to avoid false positives. Starting to explore a new dataset, there is no a-priori reason to assume that the covariance ot the data is not Gaussian as a first guess, and moving to a more complicated, and potentially computationally expensive, method is first merited when simpler approaches fail. In this study, the initial analysis of the feature vector was done using methods assuming an underlying Gaussian covariance distribution; Principal component analysis (PCA) for the initial dimensionality reduction and k-means to identify clusters.

Having widespread use for dimensionality reduction (*51, 52*), PCA increases interpretability and minimizes information loss. Solving an eigenvector/eigenvalue problem, PCA imposes a geometric constraint as the covariance matrix of the remaining subset of features is always diagonal, and that the data can be represented by a linear combination of the identified eigenvectors. The new uncorrelated features (variables) successively maximize variance, but the main use of PCA should be descriptive rather than inferential. For example, for spatial patterns the orthogonality constraint can give rise to spurious global structures with large amplitude even when the true pattern is known to be local.

The implication of assuming that linear combinations of the input features can capture the dominant underlying patterns, is that PCA is the optimal method for dimensionality reduction if the underlying covariance structure is Gaussian. If the input features interact non-linearly, the underlying covariance structure is likely to be non-Gaussian. These features will not be detected by the PCA. It follows that feature normalization and standardization are recommended steps.

The clustering algorithm k-means minimizes the average squared Euclidean distance from

one data point to a cluster centroid (k), where each data point belongs to the cluster with the closest mean. In our application, the types and nutrient flux data would be the dimensions/feature space that the k-means algorithm operates in. It effectively partitions the parameter space using Voronoi cells (straight lines). Relying on Euclidean distances, k-means assumes that the underlying covariance distribution associated with the clusters is Gaussian, looking for ”round” shapes. As with PCA, this assumption can mean that the algorithm fails in the presence of non-Gaussian distributions.

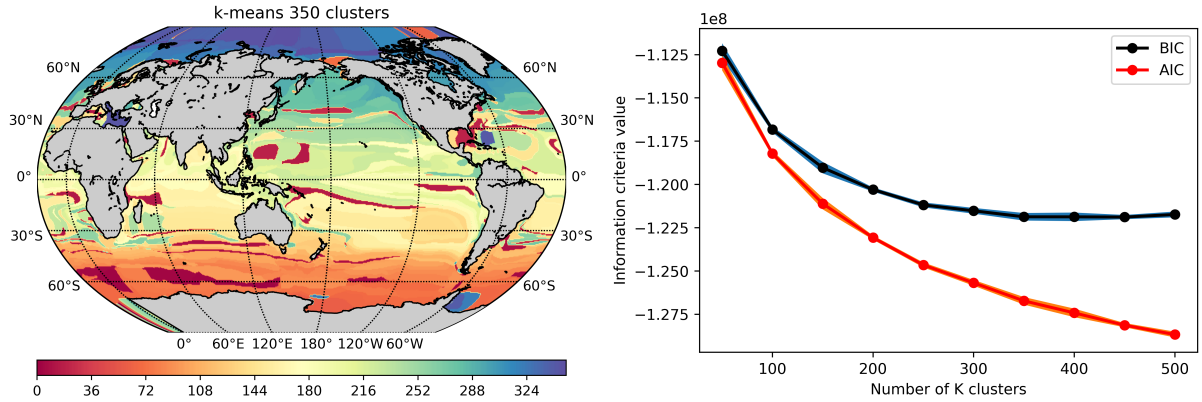
The clustering algorithm is used to arrive at a statistical model that can represent the process that generated the data. To assess if robust clusters have been identified in the types and nutrient flux data, the goodness-of-fit of the k-means algorithm as compared to the ”true” model that generated the data should be assessed. If successful, different numbers of parameters (k clusters) approximate the data, but the models can over, and under, fit. The optimal model in the context of k-means is the one arrived at with a number of k that closest approximates the ”true” model. If the ”true” model is known, the Kullback-Leibler divergence (cross entropy+entropy) can be used. In most cases the ”true” model is unknown, and it is common practice to assess the goodness-of-fit using information criteria. The advantage of using t-SNE comes from that the original high dimensional data is used as the ”true” model and the Kullback-Leibler divergence can in this manner probabilistically compress the high dimensional data onto lower dimensions. In this manner the topology of the data is conserved in the lower dimensional rendition.

Akaike (1971) formalised the intuition that some information is lost using a model to represent the process that generated the original data (*53*). The Akaike Information Criteria (AIC) estimates the relative amount of information lost by a given model. The AIC approximates the

”true” model by penalising an increase in candidate model complexity (for k-means increasing the number of k) using the likelyhood-function. For large volumes of data, the AIC is an asymptotically unbiased estimate of the cross-entropy risk, meaning that the model with the minimum AIC score will possess the smallest Kullback-Leibler divergence. Using this number of k gives us the statistical model that best approximates the underlying ”true” model that generated the data (*54*). The Bayesian Information Criteria (BIC) is also based on the likelihood-function, estimating a function of the posterior probability of a model being true. The BIC has a higher penalty for increasing candidate model complexity. Both the AIC and BIC are based on assumptions and asymptotic approximates, implying that they should only disagree if the AIC chooses a larger number of parameters than the BIC. For practical applications as in (*19, 22*) where k-means is used, the combination of the AIC and BIC is recommended to assess goodness-of-fit (*55*). As an example, if the BIC reaches a minimum and starts increasing, while the AIC asymptotes shortly thereafter, a parameter number between the BIC minimum and the point where the AIC asymptotes is optimal.

Both PCA and k-means minimize the mean-squared reconstruction error, and PCA is a super-sparse k-means (*56*). Applying PCA reduces the number of ”features” while preserving the variance. K-means reduces the number of data points, assigning them to the clusters, but it does **not** preserve the underlying covariance distribution of the data. This implies that k-means and PCA will agree only when the cluster centroids is sufficiently close to the PC. Note that other PCA methods exist e.g. kernel based, but are recommended only in situations where known non-linear relationships and correlations exist.

A practical example, and cautionary note, is given using the data from the Darwin model. For this data the results that the k-means algorithm produce can look reasonable (Figure S4a).



(a) Spatial projection with k=350

(b) AIC and BIC

S4: Illustration of k-means applied to the Darwin data.

Although the results visually look reasonable, the AIC and BIC test both failed, with and without PCA as a dimensionality reduction method (Figure S4b). The failure of the AIC and BIC test suggests that the underlying covariance distribution of the data is highly non-Gaussian, such that the statistical models like k-means fail to converge. This implies that the models are not able represent the underlying ”true” model, because the centroids are not able to partition the feature space so that e.g. consistent regions can be found. The suggested importance of the highly non-linear interactions in the Darwin data means that a different algorithm could be more appropriate. With the application of t-SNE, a clustering algorithm, in the Darwin model case DBSCAN, can be chosen such that emergent features of the non-linear, and non-Gaussian covariance, of the data can be captured.