Decoding Ecological Complexity: Unsupervised Learning determines marine eco-provinces

Maike Sonnewald, with Steph Dutkiewicz, Chris Hill and Gael Forget

Princeton University & NOAA/Geophysical Fluid Dynamics Laboratory
visitor University of Washington

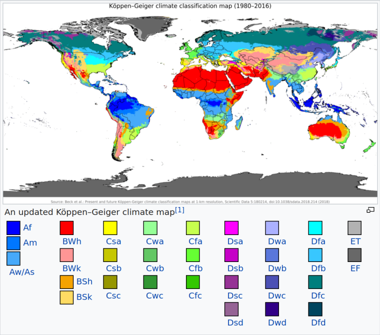
NOAA Geophysical Fluid Dynamics Laboratory

Utility of ocean ‘Provinces’

Question: How well can we see ocean biogeographic regions?

- Obs. data are sparse/complicated

- Ocean has liquid boundaries



- Compare locations
- Conservation/monitoring
- Assess base of food chain

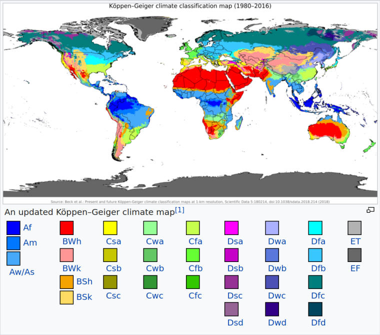
Develop objective method to classify ocean ecology for both regional and global work

Longhurst et al. 1995, Zimmermann et al. 2018

Utility of ocean 'Provinces'

Question: How well can we see ocean biogeographic regions?

- Obs. data are sparse/complicated
- Ocean has liquid boundaries



An updated Köppen–Geiger climate map[1]

- Compare locations
- Conservation/monitoring
- Assess base of food chain

Develop objective method to classify ocean ecology for both regional and global work

Longhurst et al. 1995, Zimmermann et al. 2018

Unsupervised Learning: Find structure

Supervised

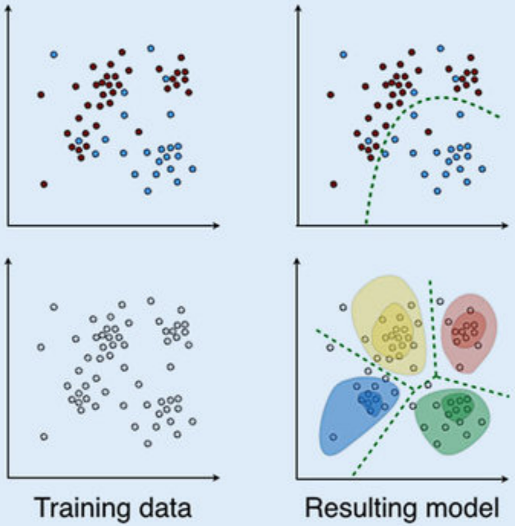
-Labeled data

-Decision boundary

Unsupervised

-No labels

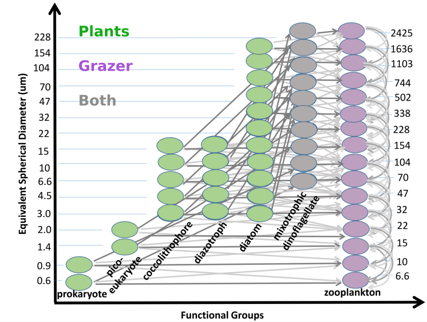
-Identify structures

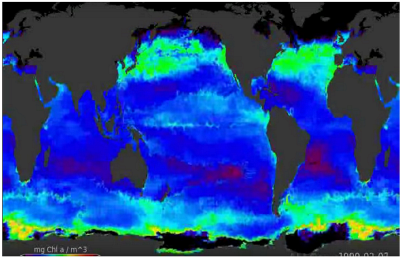


Langs et al. 2018

Darwin: Global biogeochemical ecosystem model

Physical model (ECCO)+biogeochemistry+trait based ecology

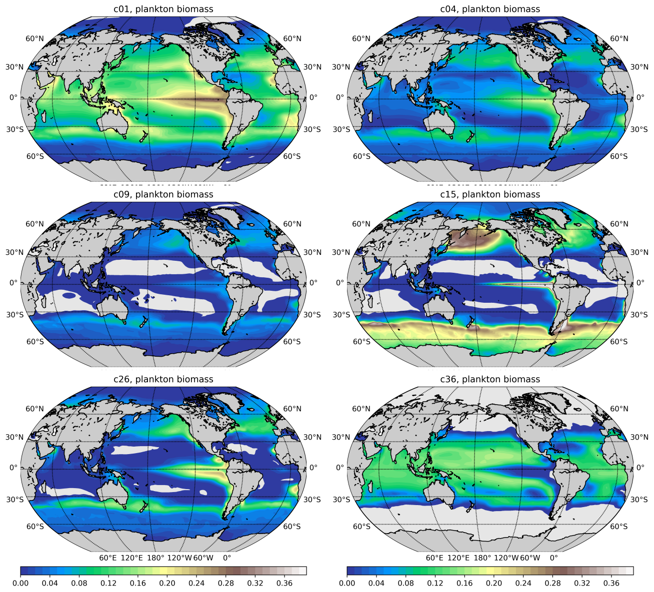




mg Chlorophyll a/m^3

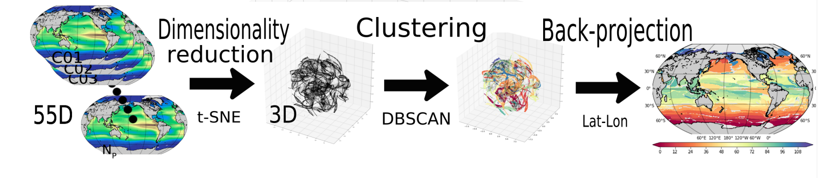
Dutkiewicz et al., 2015

51 species (biomass) and 4 nutrients (concentration)

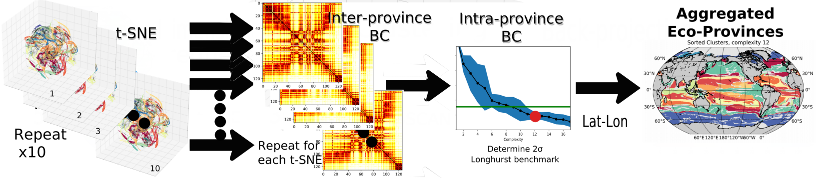


SAGE: Systematic AGgregated Eco-province method

Eco-Provinces



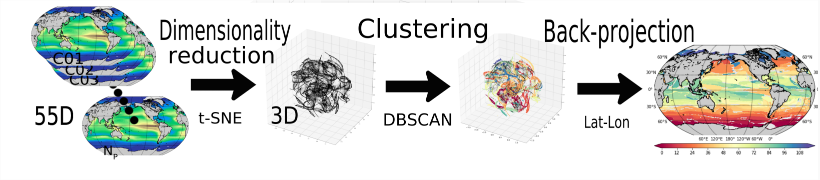
Aggregated Eco-Provinces (AEPs)



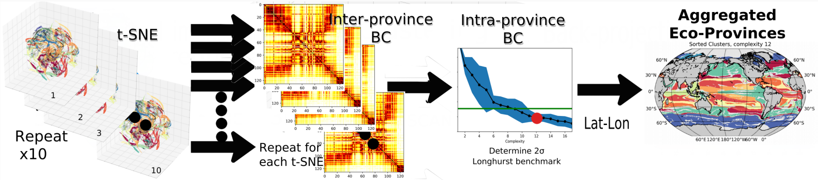
Sonnewald et al., Science Advances, 2020

SAGE: Systematic AGgregated Eco-province method

Eco-Provinces



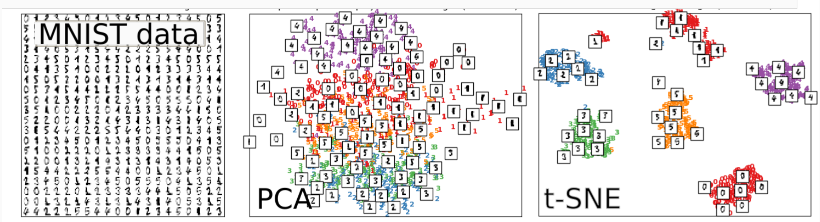
Aggregated Eco-Provinces (AEPs)



Sonnewald et al., Science Advances, 2020

Accounting for nonlinearities matters

ML benchmark: Classify 70k hand written digits (MNIST data)



- PCA assumes underlying normal distribution
- t-Statistic Neighborhood Embedding makes no assumptions
  - $\rightarrow$ more appropriate for geoscience data

scikit-learn.org

t-Statistic Neighbourhood Embedding

t-Statistic Neighbourhood Embedding helps 'flatten' the data.

We minimize 'distance' between lat+lon points in 11D and a low dimensional projection using the Kullbach-Leibner (KL) divergence.

If x_i is the i-th object in 11D, and y_j is the j-th object low-dim: space

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

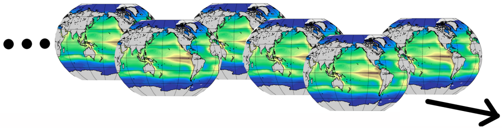
and the same for a reduced dimensional set:

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

This is done as:

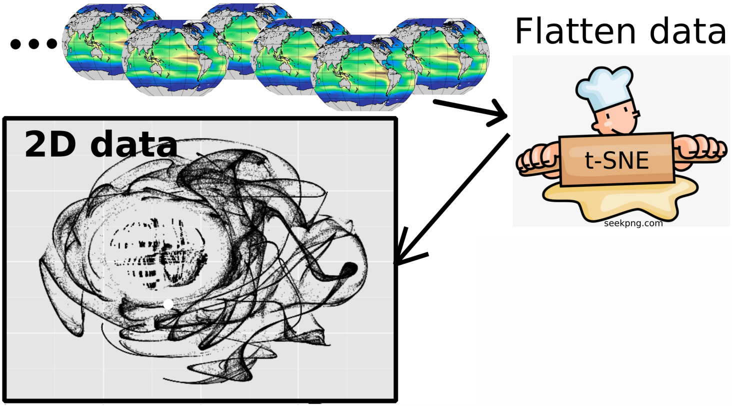
$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Learning: Understanding the data

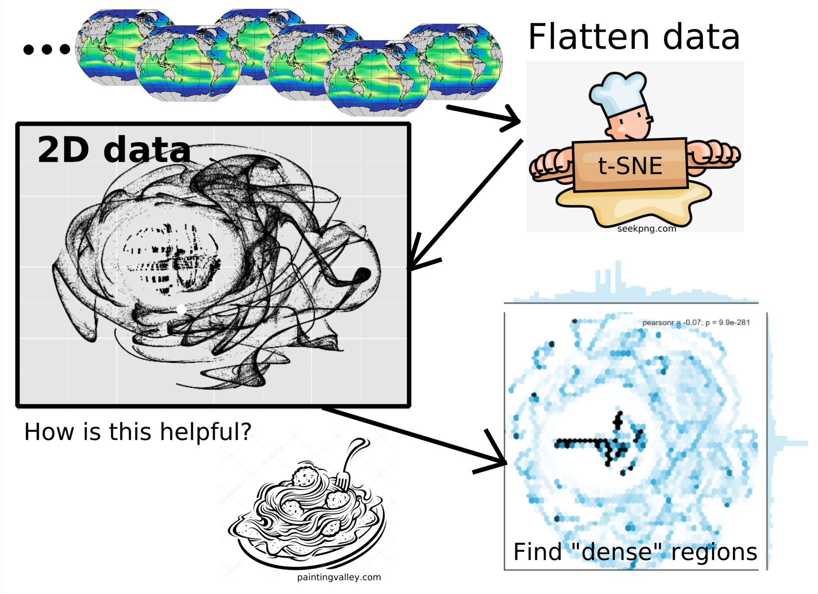


Flatten data

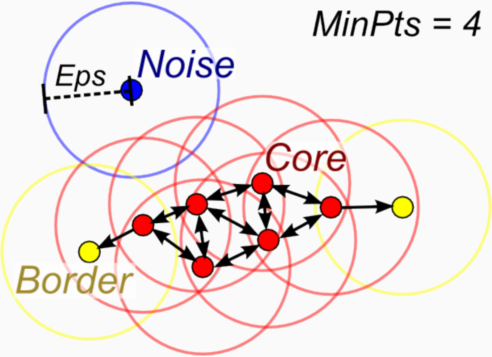
Learning: Understanding the data



Learning: Understanding the data



Density-based spatial clustering of apps with noise

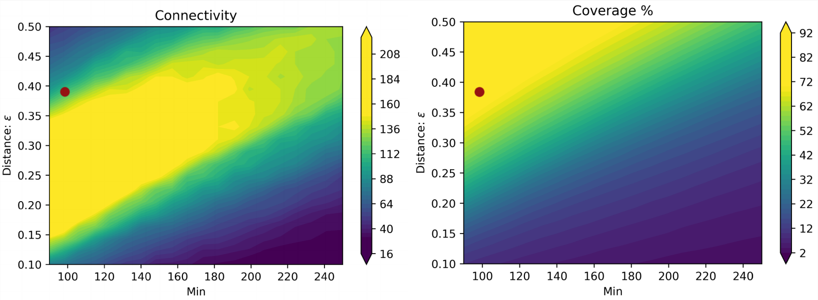


stackexchange.org

- Parameters: distance ‘Eps’ and minimum points ‘MinPts’

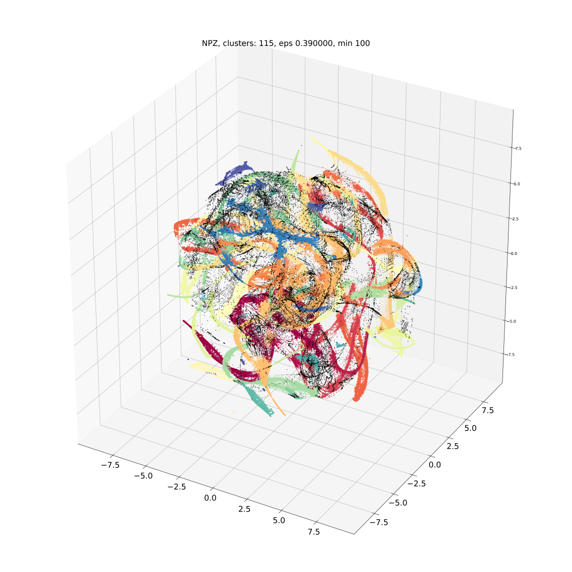
Ester et al. 1996

Densitybased spatial clustering of apps with noise



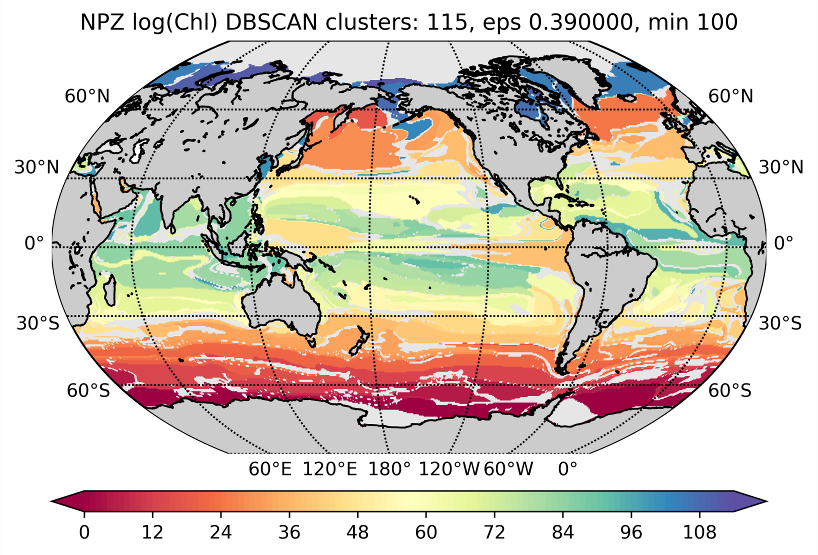
2D ‘elbow’ check in connectedness+cover

Density-based spatial clustering of apps with noise



Clustering nutrients, phytoplankton and zooplankton (NPZ)

Eco-Provinces



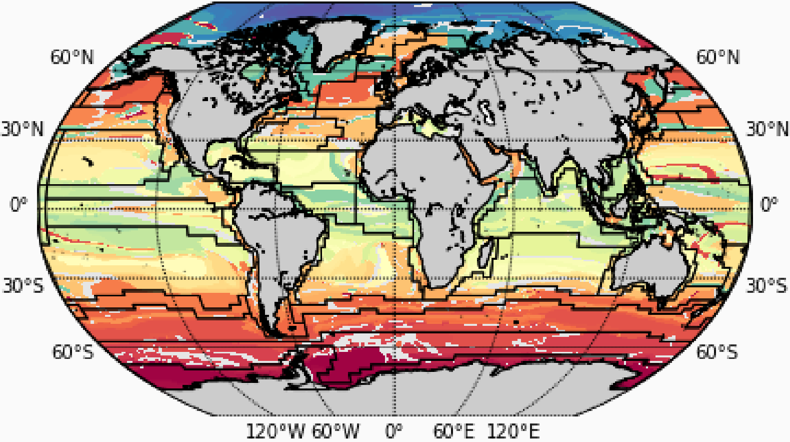
Clustering nutrients, phytoplankton and zooplankton (NPZ)

Aggregated Eco-Provinces: AEPs

Global utility of eco-provinces

Goal: Aggregation/nesting for regional to global insight

Current: Longhurst provinces are 'gold standard'

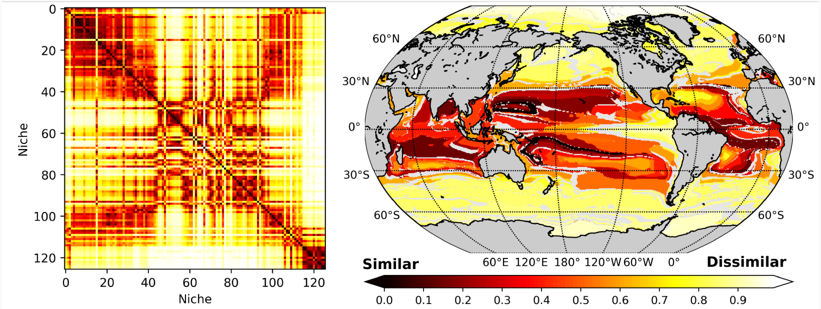


Longhurst provinces on eco-provinces (Longhurst et al. 1995)

Quantifying ecological "distance"

Bray-Curtis dissimilarity:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

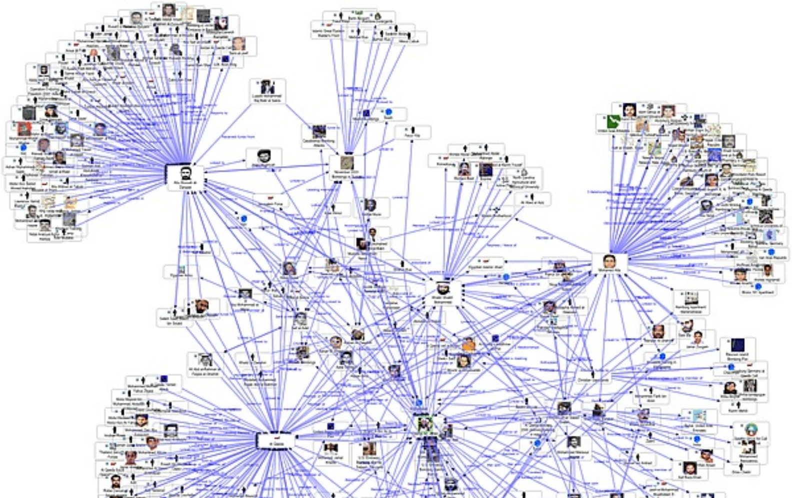


Sonnewald et al., Science Advances, 2020, Bray and Curtis, 1957

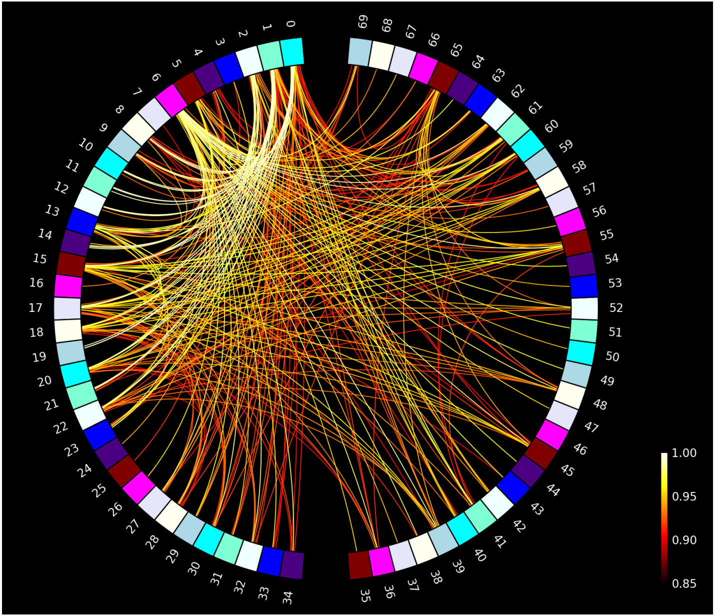
Leveraging complexity: Graph Theory

"A branch of mathematics used to represent relations and networks. Widely used in network analysis. A graph consists of a set of points (nodes or vertices) and the pairwise links between them (arcs or lines)."

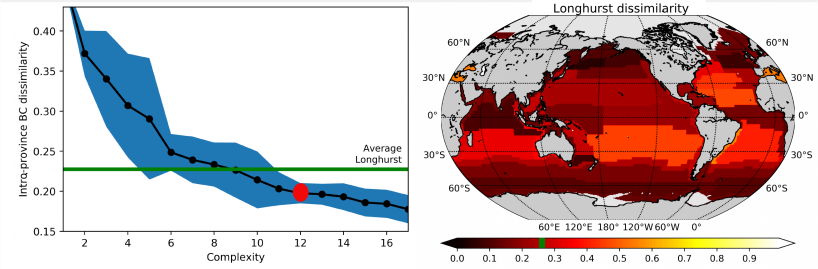
Oxfordreference.com



Complexity: Eco-province graph

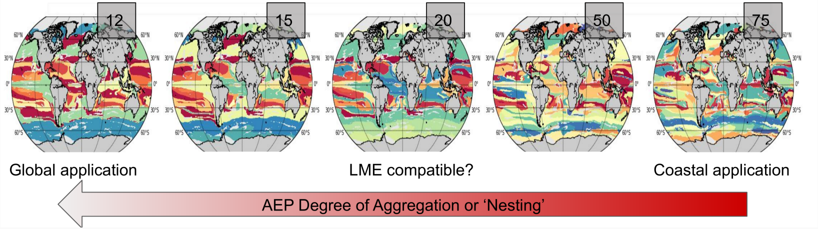


Complexity: Global application benchmark

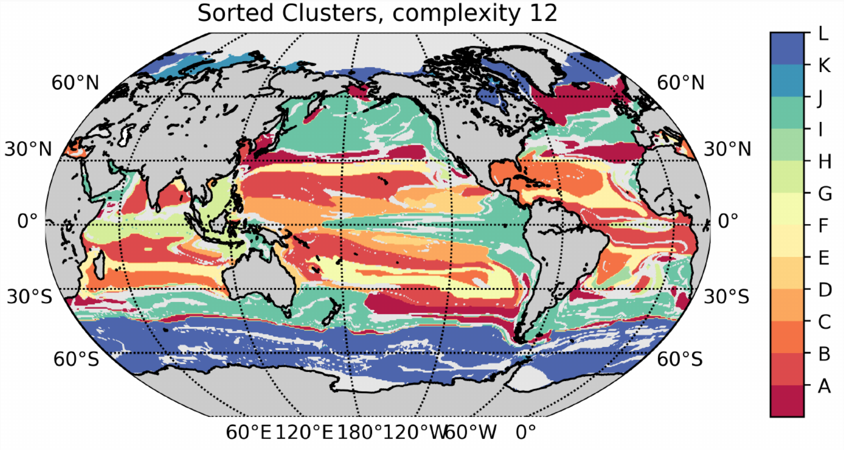


Sonnewald et al. Science Advances, 2020

Complexity: Choosing aggregation level

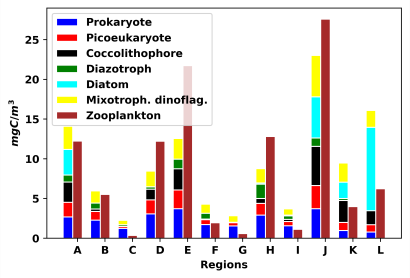


Aggregated Ecological Provinces



Sonnewald et al. Science Advances, 2020

Interpret patterns

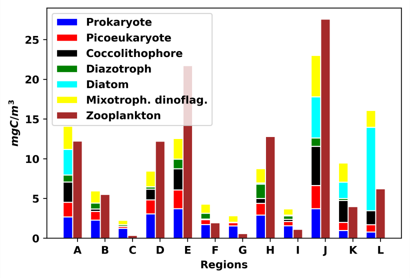


- Similar biomass/chl but different community structure
- Biomass is a poor predictor of zooplankton: Trophic cascades?

Model 'cartoon' of real ecosystem: Combine with in-situ observations?

Sonnewald et al. Science Advances, 2020

Interpret patterns



- Similar biomass/chl but different community structure
- Biomass is a poor predictor of zooplankton: Trophic cascades?

Model 'cartoon' of real ecosystem: Combine with in-situ observations?

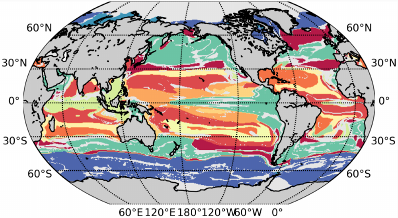
Sonnewald et al. Science Advances, 2020

Global biochemical provinces

Objectively uncovered eco-provinces in global ocean.

→ Aggregation for global to regional applications.

- Method:
    - Probabilistic t-SNE and DBSCAN
    - AEP: Graphs and dissimilarity
- Similar Chl;different ecology
- Ecological impact on zoolankton abundance



Systematic AGregation of Eco-provinces method:
github.com/maikejulie/plottingAEPs

Sonnewald et al. Science Advances, 2020

Thank you!