Bioinformatic Mapping of Ocean Biogeochemical Provinces

Matthew. J. Oliver[1,2], Oscar M. E. Schofield[1,2], Paul G. Falkowski[2,3], Andrew J. Irwin[4]

[1]Coastal Ocean Observation Lab, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA, [2]Environmental Biophysics and Molecular Ecology Lab, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA [3]Department of Geology, Rutgers University, Piscataway, NJ, USA, [4]Mathematics and Computer Science, Mt. Allison University, New Brunswick, Canada

Introduction

In 1995, Alan Longhurst introduced the concept of biogeochemical "provinces" in the oceans (Longhurst 1995, 1998; Figure 1). This province concept was based on climatological data of mixed layer depth, Brunt-Vaisala frequencies, Rossby radius of deformation, photic zone depth, and surface nutrient fields. This concept provided a framework to compare and contrast biogeochemical processes over broad regions of the global ocean. Province designations have been used to understand global distributions of primary productivity, DMSP fluxes, distributions of pelagic flora and fauna and other biogeochemically relevant parameters (Ducklow 2003; Boyd and Doney 2003; Waniek et al 2005). As a result, the province model influenced the shaping of our understanding of biogeochemical cycles of the global ocean.

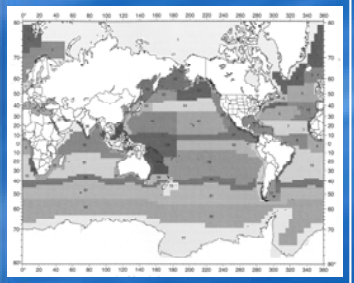


Figure 1: Longhurst's geographic distribution of provinces (From Ducklow, 2003).

What has been difficult to assess is the spatial and temporal variability in the provinces, which are known to be important on both short term (days to weeks) and long term (Pacific Decadal Oscillation [PDO], North Atlantic Oscillation [NAO]) time scales. The temporal variability of province distribution and interaction remains one of the most vexing issues in discriminating between secular changes (i.e., anthropogenically induced trends) and decadal cycles in the ocean system (i.e., natural variability). In the time since the first CZCS image was processed, atmospheric CO_2 levels have risen ~ 40 ppm, global chlorophyll concentrations have increased by 22% (Antoine et al 2005). Furthermore, global nitrate available at the ocean surface has decreased significantly in the last century (Kamykowski and Zentara 2005). While oceanic biogeochemical provinces oscillate seasonally, there appears to be a secular change in oceanic provinces, the underlying causes of which we know little about. A clearer understanding of the processes that control the distribution of oceanic provinces requires an objective method to resolve these water masses in a time-dependent manner (Platt and Sathyendranath 1999). We propose to develop and implement a biogeochemical classification scheme that overcomes the technical difficulties of fixed boundary province classification in order to objectively elucidate the time and space dependent distribution of provinces.

Goals

By implementing a biogeochemical province classification system that is time and space resolved, we hope to answer these questions"

- *How does the annual cycle affect the distribution of provinces?*
- *How much do inter-annual cycles modulate province distribution and cause deviations from the annual cycle?*
- *Do episodic events like large dust storms rapidly change the distribution of the provinces?*
- *Are there trends in regional or global primary productivity based on province distribution?*

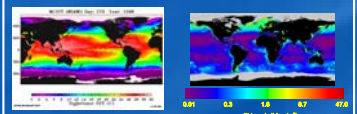
To accomplish these goals, we must push the province model beyond static climatologies.

Approach

There are two major obstacles to circumvent if we are to achieve our goals:

1. Identify data sets with synoptic coverage that contain the same information as Longhurst used.
2. Determine a way to objectively classify provinces in a dynamic way.

Solution to Obstacle #1: Combined Satellite Coverage of Sea Surface Temperature and Ocean Color



Along with climatological chlorophyll values from the CZCS sensor and ship-based measurements, the Longhurst approach to province specification includes global climatologies of mixed layer depth, Brunt-Vaisala frequencies, Rossby radius of deformation, photic zone depth, and surface nutrient fields. While all of these parameters are relevant (to varying degrees) to upper ocean biogeochemistry, there is a high degree of autocorrelation between these parameters. For example, mixed layer depth, Brunt-Vaisala frequency, Rossby radius of deformation and nutrient fields are all significantly correlated to sea surface temperature on a global scale. Furthermore, water column integrated chlorophyll concentrations, photic depth and nutrient fields are significantly correlated to ocean color. Therefore, the global time series of satellite ocean color and sea surface temperature provide a significant amount of discrimination power in determining the locations of biogeochemical provinces (Esaias et al 2000). Through the use of satellite data, we gain the temporal resolution required to infer the dynamics of the boundaries of biogeochemical provinces on seasonal, annual and inter-annual time scales.

Approach cont.

Solution to Obstacle 2: Use bioinformatic algorithm to objectively locate provinces in space and time.

The bioinformatic approach we propose to objectively classify biogeochemical provinces in the global ocean has six general steps; I) concomitant passes of ocean color and sea surface temperature (SST) are merged spatially, II) SST and Ocean Color Parameters are standardized to their relative means, iii) standardized data are projected into multidimensional parameter space and clustered by an ensemble of clustering algorithms, iv) a Figure of Merit is calculated which determines the likely number water masses/provinces, v) surface water mass/province boundaries are mapped, vi) the relative strengths of the boundaries are assessed through multidimensional gradient analysis (Figure 2, Oliver et al. 2003).

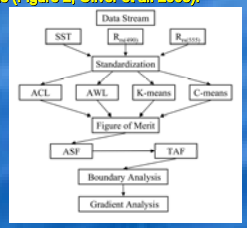


Figure 2: Flow diagram of this analysis. The following case study of the Mid-Atlantic Bight will detail the method.

Case Study: Mid-Atlantic Bight

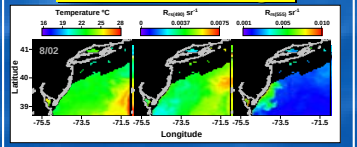


Figure 3: Temperature and reflectance maps on 8/02 2002 in this analysis.

Principal components analysis indicated SST, $R_{rs(490)}$ and $R_{rs(555)}$ were the largest contributors to the variance in the data set (96%).

Data were standardized by subtracting their respective mean and dividing by their respective standard deviation to equally weight SST, $R_{rs(490)}$ and $R_{rs(555)}$.

Cluster Analysis

Clustering Algorithm	Description
Agglomerative Complete Linkage (ACL)	Data are hierarchically grouped from *n* to 1 clusters. Data are grouped from closest to furthest based on Euclidian distance in predictor space. The distance between clusters is measured based on the maximum distance between cluster edges in predictor space.
Agglomerative Ward's Linkage (AWL)	Data are hierarchically grouped from *n* to 1 clusters. Data are grouped at each step to minimize the variance of the clusters.
K-Means	Data are divided from 1 to *k* clusters where *k* is the number of clusters requested by the user. To form *k* clusters, *k* cluster centers are randomly initialized in predictor space. Data are then assimilated into cluster centers as to minimize the within cluster sum of squares.
Fuzzy C-Means	Similar to K-means, except this algorithm clusters initial cluster centroids through competitive learning.

Figure of Merit

A major difficulty in cluster analysis is determining how many clusters (or provinces) should be used to describe a data set as each observation could theoretically represent its own cluster. The Figure of Merit (FOM) algorithm was designed to calculate the difference between expression vectors of genes (Yeung et al 2001); here it is used to analyze the inherent structure of clusters in predictor space detected by the clustering algorithms. In this case, "gene" expression vectors were standardized values of SST, $R_{rs(490)}$ and $R_{rs(555)}$ at each pixel.

$$FOM(c,k) = \sqrt{\frac{1}{n}\sum_{i=1}^{3}\sum_{j=2}^{k}\sum_{l=1}^{m_j}\left(\bar{a}_{ij} - a_{ijl}\right)^2}$$

where c is one of the four clustering algorithms, n is the total number of observations, i =1-3 indexes the three variables measured at each pixel, j is the cluster number, k is the number of clusters each data set was divided into, l is a specific observation of the total number of pixels m in cluster j, a_{ijl} is the specific standardized observation of predictor i in cluster j, and $\bar{a}_{ij}$ is the mean for each cluster.

- *This function is essentially a measure of the variation within clusters as a function of cluster number (Figure 4).*

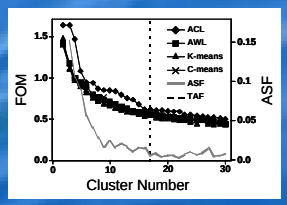


Figure 4: Figure of Merit (*FOM*), Average Slope Function (*ASF*) and Threshold of Acceptable Flatness (*TAF*) calculation for each of the four days with the results of each of the clustering algorithms. A large *FOM* indicates that the variance within each cluster is comparatively large and that the cluster centroid is a generally poor predictor of the other data points within each cluster. A small *FOM* indicates that the cluster centroid better predicts the other members of its cluster, and that the variance with in the cluster is comparatively small. *ASF* is the average percent change of the four clustering algorithms compared to the maximum *FOM*. *TAF* was defined when the average change in the *ASF* was less than 1% for more than three clusters, and represents the maximum possible number of different water types.

- *In this study, an average slope function (ASF) was used to determine what the average decrease in the FOM was as clusters were added.*

$$ASF(k) = \frac{1}{4}\sum_{c=1}^{4}\frac{FOM(c,k+1) - FOM(c,k)}{FOM_{\max}(c)}$$

where is the maximum value for a specific cluster algorithm *c*.

- A Threshold of Acceptable Flatness (*TAF*) was defined at the smallest cluster *k* where < 0.01 (< 1% decrease in relative to the maximum) for three or more. consecutive clusters.

- *The use of the ASF and TAF established an upper bound for what we believed to be reasonable cluster numbers or water type assignments by the suite of clustering algorithms.*

Boundary and Gradient Analysis

Clusters defined in a data set occupy predictor space represented by standardized SST, $R_{rs(490)}$ and $R_{rs(555)}$ and physical space represented by latitude and longitude.

The mapping of defined water types for any cluster number k and clustering algorithm c into physical space (this case in dimensions of latitude and longitude) defines physical boundaries between provinces.

Because of this, a physical space representation of the clusters was used to determine which boundaries occurred most often by constructing a 2-d histogram for boundaries at $2 \le k \le TAF$ (Figure 5A).

The purpose of the gradient analysis was to determine how different water types were in predictor space in relation to geographic space (Figure 5C).

The relative strength of the boundaries was defined as:

$$D_{x\to x+\Delta x} = \sqrt{\left(SST_x - SST_{x+\Delta x}\right)^2 + \left(R_{rs(490)\,x} - R_{rs(490)\,x+\Delta x}\right)^2 + \left(R_{rs(555)\,x} - R_{rs(555)\,x+\Delta x}\right)^2}$$

$$D_{y\to y+\Delta y} = \sqrt{\left(SST_y - SST_{y+\Delta y}\right)^2 + \left(R_{rs(490)\,y} - R_{rs(490)\,y+\Delta y}\right)^2 + \left(R_{rs(555)\,y} - R_{rs(555)\,y+\Delta y}\right)^2}$$

$$\nabla G(x,y) = \sqrt{\left(\frac{D_{x\to x+\Delta x}}{\Delta x}\right)^2 + \left(\frac{D_{y\to y+\Delta y}}{\Delta y}\right)^2}$$

where *SST* is standardized sea surface temperature, $R_{rs(490)}$ is standardized $R_{rs(490)}$, $R_{rs(555)}$ is standardized $R_{rs(555)}$, $D_{x\to x+\Delta x}$ is the standardized predictor space distance between x and $x+\Delta x$ (Longitude), $D_{y\to y+\Delta y}$ is the standardized predictor space distance between y and $y+\Delta y$ (Latitude), and ∇ gradient in predictor space with respect to and . While the boundary analysis determines likely locations of water mass boundaries, *PSA(x,y)* describes the strength of boundaries through simultaneous analysis of SST, $R_{rs(490)}$ and $R_{rs(555)}$.

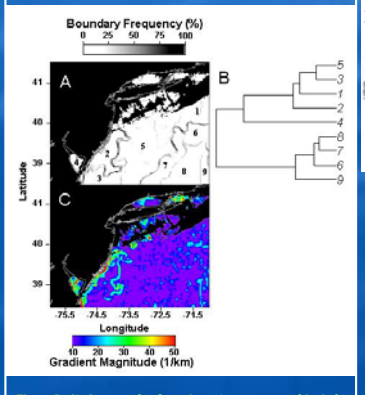


Figure 5: A) An example of provinces /water masses objectively classified in the Mid-Atlantic Bight. Boundaries show the presence of a near-shore upwelling region and the shelf-break front. B) The horizontal branch lengths between classified regions in the relatedness tree are proportional to the difference in ocean color and SST between regions. C) The magnitude of parameter space gradient indicates how different provinces are across a boundary. The strongest gradients are between regions 2 and 5, indicating the waters on either side of the boundary are very different; as opposed to the shelf break front boundary between regions 5 and 7, which is weak in comparison.

Validation of Boundaries

This algorithm was independently validated by ship-board salinity (Figure 8) and nutrient (Figure 7) transects.

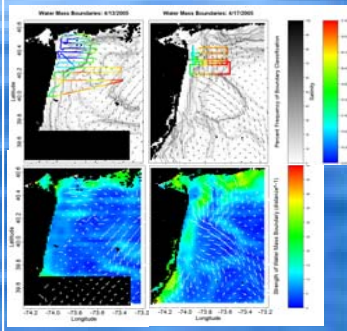


Figure 6: The bioinformatic approach using SST and Ocean to detect provinces/water masses near the Hudson River was validated with ship based salinity tracks for two separate days. Salinity changes are concomitant with objectively detected boundaries from space, thus validating their presence. The nearshore boundaries appear to be the strongest. Also, surface currents measured by CODAR show boundaries detected by this method are in convergent and divergent areas.

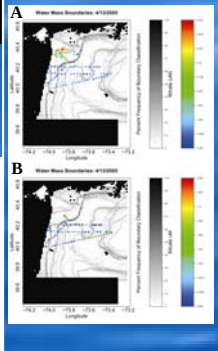


Figure 7: Independent test of boundary analysis with nitrate as an independent predictor. Black areas indicate areas of cloud contamination or land. A) This analysis indicates that nitrate levels in the "bulge" are very high while on the other side of the bulge boundary they are typical of open ocean values. B) In addition, there are relatively high values of silicate present in an old river plume that is being advected offshore.

Conclusions

In our regional analysis described above, we believe that we have a solution for the major obstacles that must be resolved before a dynamic, objective elucidation of biogeochemical provinces can be employed by:

1. Using the available SST and ocean color satellite databases.
2. Employ a version of a bioinformatic algorithm to objectively and statistically determine the location and dimensions of ocean provinces.

References

Antoine, D., A. Morel, H. R. Gordon, V. F. Banzon, R. H. Evans 2005. Bridging ocean color observations of the 1980s and 2000s in search of long-term trends. J. Geophys. Res., 110, C06009, doi:10.1029/2004JC002620.

Boyd, P. W. and S. C. Doney, 2003. Impact of Climate Change and Feedback Processes on the Ocean Carbon Cycle, in Ocean Biogeochemistry, Ed. M. Fasham, Springer, New York.

Ducklow, H. 2003. Biogeochemical Provinces: Towards a JGOFS Synthesis, in Ocean Biogeochemistry, Ed. M. Fasham, Springer, New York.

Esaias, W. E., R. L. Iverson, and K. Turpie. 2000. Ocean province classification using ocean colour data: observing biological signatures of variations in physical dynamics. Glo. Ch. Biol. 9:39-55.

Kamykowski, D. and S. Zentara. 2005. Changes in world ocean nitrate availability through the 20th century. Deep-Sea Res. I (available on-line and in press)

Longhurst, A. R. 1995. Seasonal cycles of pelagic production and consumption. Prog. Oceanogr. 36:1907-1928.

Longhurst, A. R. 1998. Ecological geography of the sea. Academic, San Diego, 398pp.

Oliver, M. J., Kohut, J. T., Irwin, A. J., Glenn, S. M., Schofield, O., Moline, M. A., Bissett, W. P. 2003. Bioinformatic Approaches for Objective Detection of Water Masses on Continental Shelves. J. Geophys. Res. 109, C07S04,doi:10.1029/2003JC002072.

Platt, T. and S. Sathyendranath. 1999. Spatial structure of pelagic ecosystem processes in the global ocean. Ecosystems, 2:384-394.

Waniek, J. J., D. E. Schulz-Bull, J. Kuss, T. Blanz. 2005. Long time series of deep water particle flux in three biogeochemical provinces of the northeast Atlantic. J. Mar. Sys. 56:301-318.

Yeung, K. Y., D. R. Haynor and W. L. Ruzzo, Validating clustering for gene expression data, Bioinformatics, 17(4), 309-318, 2001.

The support of the National Ocean Partnership Program (N00014-07-1-1019), the Office of Naval Research CDMCP and HyCODE programs (N00014-97-0767, N00014-99-0195) and the NSF EEGLE program (OCE-9727941, OCE-9727942) are gratefully acknowledged. We especially thank Mike Crowley and Jen Boesh for providing cloud masking and processing of the satellite data. The other ONR supported researchers are thanked for patience/humor when working in the Jersey swamps. We toast our fellow COOL faculty, students, and research staff (http://marine.rutgers.edu/cool/) with good beer. Finally, the continuing support from the great state of New Jersey is acknowledged.