Northwest Atlantic Fisheries Organization

Serial No. N5871 **NAFO SCR Doc. 10/72**

SC WG ON THE ECOSYSTEM APPROACH TO FISHERIES MANAGEMENT – DECEMBER 2010

Preliminary Analysis for the Delineation of Marine Ecoregions on the NL Shelves

Pierre Pepin, Andrew Cuff, Mariano Koen-Alonso, and Neil Ollerhead
Northwest Atlantic Fisheries Centre
Department of Fisheries and Oceans
P.O. Box 5667
St. John's, NL, Canada A1C 5X1

Abstract

Delineating spatial management units is a necessary element for implementing ecosystem approaches to fisheries management. This study aimed to define spatially coherent ecological units (ecoregions) on the Newfoundland-Labrador (NL) Shelves using both physical and biological data which can serve as the basis for defining spatial management units (e.g. "management ecosystems"). The methods used were similar to previous studies in the US Northeast Atlantic continental Shelf and the Canadian Scotian Shelf in order to maintain consistency and a degree of comparability between results. Datasets were analyzed and classified using principal components analysis and k-means clustering. The clustering results were mapped in order to examine spatial distributions of the clusters. Results indicated that the physical variables (bathymetry, primary production, sea surface temperature) dominated the principal component analysis (PCA) signal when included in the analysis. Five major clusters (NL Shelf, Grand Banks, Southeast Shoal, Continental Slope, and nearshore) were typically identified on the classified maps. However, results varied based on the different PCA runs used in the clustering exercise. Features characterized at different spatial scales became apparent in the results with the inclusion of coral datasets. Including corals into the analysis created "patches" on the resulting maps, while the exclusion of corals from the analysis resulted in smoother and more continuous representation of the clusters on the maps.

1. Introduction

Ecosystem approaches to fisheries (EAF) are essentially place-based approaches; they aim to provide management provisions and advice encompassing multiple stocks which inhabit a common and geographically-defined area. These "ecosystem management" units, and the scale at which they are defined, ideally would capture the core of a functional ecosystem, though other considerations should also be taken into account in defining them (e.g. jurisdictional boundaries and legal issues, main fisheries and fleets, operational issues regarding surveillance and enforcement, etc). A necessary starting point in the process of defining "ecosystem management" units is the delineation of ecosystem boundaries and identification of major ecosystem subunits (ecoregions) (Fogarty and Keith, 2009, Zwanenburg et al., 2010).

The Scientific Council (SC) of the of the Northwest Atlantic Fisheries Organization (NAFO) had tasked its Working Group on Ecosystem Approaches to Fisheries Management (WGEAFM) to development a suitable framework that could allow NAFO to implement an EAF tailored to the needs and characteristics of the organization. The identification and delineation of ecoregions, as well as ecosystem-level management units, has been a topic of study by WGEAFM since its creation (NAFO 2008), but this also became a key element within the recently developed "Roadmap to EAF" (NAFO 2010).

Within WGEAFM, previously discussed work regarding ecoregion identification and delineation includes the US Northeast Atlantic Continental Shelf, and the Canadian Scotian Shelf (NAFO 2008, 2010, Fogarty and Keith, 2009, Zwanenburg et al. 2010). However, other regions within the NAFO Convention Area, like the Newfoundland and Labrador Shelves and the Flemish Cap, still remain without studies of this type.

The Newfoundland and Labrador Shelves have been identified as one of the twelve major marine biogeographic units by Fisheries and Oceans Canada (DFO 2009), but more detailed comparisons with the results from the US Northeast Atlantic Continental Shelf, and the Canadian Scotian Shelf would require the identification and delineation of ecoregions following similar-enough protocols and data.

In this context, the aim of this study is, using similar methods as previous studies (Fogarty and Keith, 2009, Zwanenburg et al., 2010) to extend the body of work on ecoregion identification and delineation to the Canadian Newfoundland and Labrador (NL) shelves (Figure 1). More precisely, this work analyzes biological and physical datasets collected from 1995 – 2007 for the purpose of delineating ecoregions using both statistical and geospatial techniques.

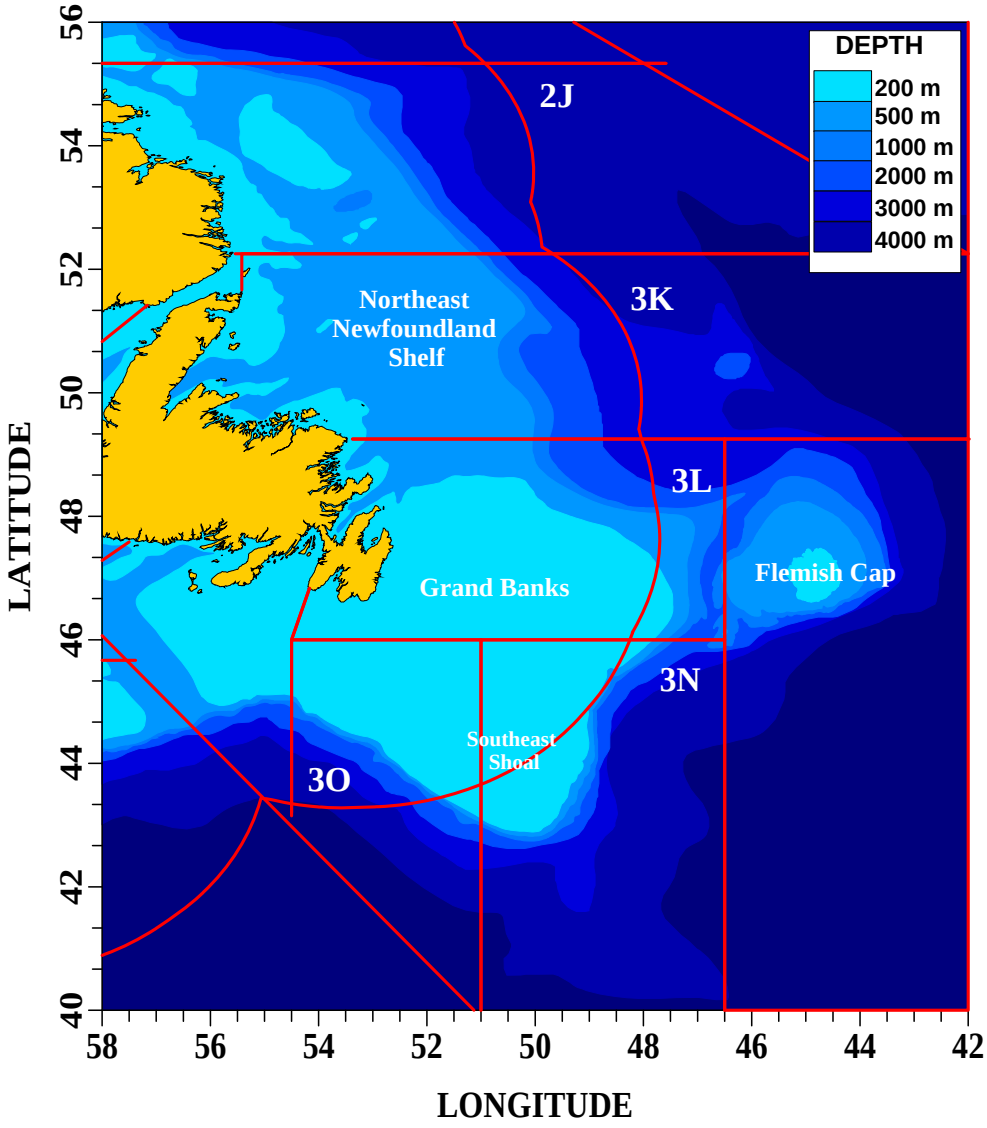

Figure 1. NAFO areas of interest for the assessment of Newfoundland Ecoregions (2J3KLNO). The major bathymetric zones showing the Northeast Newfoundland Shelf, Grand Banks, Flemish Cap as well as the Southeast Shoal region at the tail of the Grand Banks.

2. Methods

The method employed in this project (Figure 2) follows that of Fogarty and Keith (2009) and Zwanenburg et al. (2010) as closely as possible to maintain consistency among the three studies. The first step in the methods was to acquire all the datasets to be used in the analysis. Most of the data were not continuous surfaces, but rather vector point database format, and therefore had to be interpolated to a common gridded surface. Once all the data were represented as continuous surfaces they had to be made spatially comparable (i.e. perfectly overlapping cells of the same size); therefore, all raster datasets were aligned, resampled, and/or aggregated to a standard 20 km grid. Before the datasets were used in multivariate analyses, all were standardized to a common scale (mean = 0; s.d. = 1). Following these steps, the data were analyzed using principal component analysis (PCA) and then classified and mapped using k-means clustering.

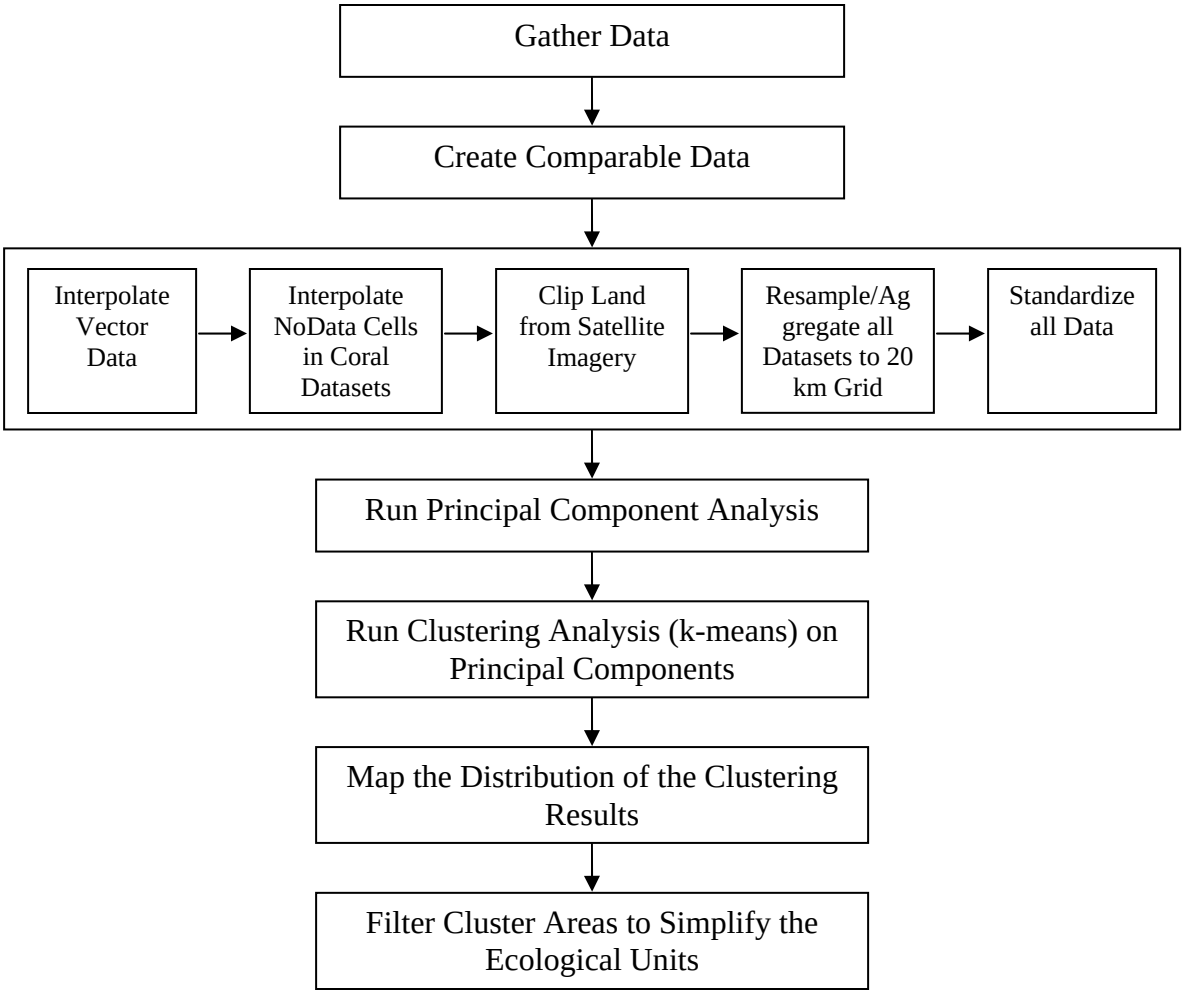

Figure 2: Flowchart of the analysis method.

2.1 Data Inputs and Processing

This section outlines the input variables used for the PCA and clustering analysis (Table 1) summarizing the initial data format (type), source, units, and temporal period. Any variables that were originally vector datasets were converted into continuous raster surfaces using appropriate interpolation methods (Goulet et al., 2010). These vector datasets were all provided by various Fisheries and Oceans Canada (DFO) surveys (Goulet et al., 2010).

The GEBCO (General Bathymetric Chart of the Oceans) bathymetry dataset was downloaded from the GEBCO website (www.gebco.net). GEBCO is composed of an international group of experts who work on the development of bathymetric datasets and operates under the auspices of the International Hydrographic Organization (IHO) and the Intergovernmental Oceanographic Commission (IOC) of UNESCO.

The geology dataset was not made available at the time this work was done and was not included in the analysis.

Sea surface temperature dataset was provided by NOAA. Sea surface temperatures were measured daily using AVHRR Satellite starting in 1985 (for more information see www.nodc.noaa.gov/SatelliteData/pathfinder4km/).

DFO's Bedford Institute of Oceanography (BIO) provided the chlorophyll *a* (Chl-*a*) and primary production (PP) datasets. The original Chl-*a* datasets were acquired from the SeaWiFS (Sea Viewing Wide Field of View Sensor) satellite sensor mounted on the Orbview-2 satellite operated by GeoEye (GeoEye, 2010). The Chl-*a* estimates are derived using the OC4.v4 algorithm (O'Reilly et al., 2000). PP estimates were derived from the Chl-*a* datasets (Platt et al. 2009). The PP image is the average over all years (98-04) and the four Chl-*a* datasets are seasonal averages (spring, summer, fall, winter) over all years (97-07). To maintain consistency, the four Chl-*a* seasonal averages were averaged to produce a single Chl-*a* dataset, matching the PP dataset.

The grid size of these datasets differ even though PP was derived from Chl-*a*. Until 2004, the Chl-*a* dataset was provided in a 1.5 km grid size and this was the data used to create the PP dataset. However, after 2004 the Chl-*a* the grid size was increased to 4 km, which is why the Chl-*a* gird size is 4 km.

Richness was estimated as the number of taxa per demersal or pelagic tow. Diversity was represented using Shannon's Evenness Index ($\left[-\sum_{i=1}^{S} p_i \ln p_i \right] / \ln S$; where *pi* is the proportion of species *i*, and *S* is the number of species).

Table 1: The variables used as input for the PCA and clustering analysis showing their original data type, source, units, and time period.

Variable	Original Data Type	Data Source	Units	Period
Physiographic Variables				
Bathymetry	*Raster*	*GEBCO One Minute Grid*	*Meters (m)*	*N/A*
Surficial Seabed Geology	*Raster*	*Natural Resources Canada*	*Mean Grain Size*	*N/A*
Biological Variables				
Zooplankton Biomass	*Vector*	*DFO Juvenile Fish (0-group) Surveys*	*g (Dry Weight) (Each sample averaged over all years)*[1]	*94-99 (Summer)*
Chlorophyll-a	*Raster*	*SeaWiFS Satellite (4 km grid)*	*mg/m^3(Seasonal average)*	*97-07*
Primary Production	*Raster*	*SeaWiFS Satellite (1.5 km grid)*	*mg/m^3/year (Cumulative)*	*98-04*
Nekton Biomass	*Vector*	*DFO Juvenile Fish (0-group) Surveys*	*kg/Standard Tow (Each sample averaged over all years)*[3]	*94-99 (Summer)*
Nekton Diversity	*Vector*	*DFO Juvenile Fish (0-group) Surveys*	*Shannon's Evenness Index (Each sample averaged over all years)*	*94-99 (Summer)*
Nekton Richness	*Vector*	*DFO Juvenile Fish (0-group) Surveys*	*Count/Standard Tow (Each sample averaged over all years)*	*94-99 (Summer)*
Demersal Fish Biomass	*Vector*	*DFO MSS*	*kg/Standard Tow*[2]	*95-07 (Spring and Fall)*
Demersal Fish Diversity	*Vector*	*DFO MSS*	*Shannon's Evenness Index*	*95-07 (Spring and Fall)*
Demersal Fish Richness	*Vector*	*DFO MSS*	*Count/Standard Tow*	*95-07 (Spring and Fall)*
Cold Water Coral Biomass	*Vector*	*DFO MSS*	*kg/Standard Tow*[2]	*02-05 (Spring and Fall)*
Cold Water Coral Presence/Absence	*Vector*	*DFO MSS*	*Individual Specimen Count*	*02-07 (Spring and Fall)*
Cold Water Coral Diversity	*Vector*	*DFO MSS*	*Shannon's Evenness Index*	*02-07 (Spring and Fall)*
Cold Water Coral Richness	*Vector*	*DFO MSS*	*Count/Standard Tow*[2]	*02-07 (Spring and Fall)*

Oceanographic Variables				
Sea Surface Temperature	*Raster*	*NOAA AVHRR Satellite Nominal 4 km Grid*	*Degrees (°C) (Annual Average)*	*85-01*
Sea Surface & Bottom Temperature	*Vector*	*DFO MEDS Surveys*	*Degrees (°C)*	*94-04 (Fall)*

1. g (dry weight) standardized for each locale (i.e. sample location) based on the ratio of total dry weight per volume filtered water.
2. kg/standard tow and count/standard tow for the multi-species survey dataset means that if the trawl was below or above 15 minutes (which is the standard trawl time) the catch numbers were linearly adjusted to 15 minutes.
3. Kg/standard tow and count/standard tow for the nekton dataset means that if the trawl was below or above 15 minutes (which is the standard trawl time) the catch numbers were linearly adjusted to 15 minutes.

Table 2 provides a comparison of the datasets used in "ecoregions" delineation projects in the US (Fogarty and Keith, 2009) and Canada (Zwanenburg et al., 2010). As highlighted in the table, not all of the variables and/or datasets were consistent among these studies. Both Scotian Shelf and US studies used 10' latitude x 10' longitude grids, which equals approximately 18 km depending on latitude. However, this report on the NL shelf used 20 km grids because this was the grid size most appropriate for the development of interpolated surfaces (Goulet et al., 2010). The different grid sizes can be made comparable by simply using the resample tool in ArcGIS but the close proximity of grid sizes should not have a substantial impact on the overall estimation of ecoregions because of their dimensions should far exceed those of the grid size.

Table 2: Comparison of data used in different projects for the delineation of regional ecosystem units (modified from Zwanenburg et al., 2010).

Variable	**U.S.A. Data** (Fogarty and Keith, 2009)	**Canadian Data Scotian Shelf** (Zwanenburg et al., 2010)	**Canadian Data NL Shelf**
Physiographic Variables			
Bathymetry	National geophysical data center, meters	CHS Atlantic bathymetry data, meters	GEBCO bathymetry, meters
Surficial Geology	Benthic grab, mean grain size	GSC surficial geology, classified sediment types	N/A
Biological Variables			
Chlorophyll	Ratio of shipboard measurements of surface to subsurface chlorophyll, dimensionless	Satellite derived estimates of chlorophyll-a using SeaWiFS, mg/m^3	Satellite derived estimates of chlorophyll-a using SeaWiFS, mg/m^3
Primary Production	Satellite derived estimates of primary production using SeaWiFS, $gC/m^2/yr$	Satellite derived estimates of primary production using SeaWiFS, $mg/m^2/yr$	Satellite derived estimates of primary production using SeaWiFS, $mg/m^2/yr$
Zooplankton	ECOMON plankton sampling, displacement volume Cc $100/m^3$	Zooplankton wet weight data from the AZMP program is substituted	DFO Juvenile (0-Group) fish surveys, dry weight (g)
Benthic Biomass	Benthic grab/sled, g/m^2	N/A	N/A
Nekton Biomass			DFO Juvenile (0-Group) fish surveys, kg/standard tow
Nekton Diversity	N/A	N/A	DFO Juvenile (0-Group) fish surveys, Shannon's diversity evenness index
Nekton Richness			DFO Juvenile (0-Group) fish surveys, # unique species/tow
Demersal Biomass	NEFSC Groundfish survey, kg/tow	DFO Groundfish Survey, kg/tow	DFO multi species surveys, kg/standard tow
Demersal Diversity	N/A	N/A	DFO multi species surveys, Shannon's diversity evenness index
Demersal Richness	NEFSC Groundfish survey, mean # species/tow	DFO Groundfish Survey, kg/tow	DFO multi species surveys, # unique species/standard tow
Deep Water Coral Presence	Benthic grab/sled	ERD coral database	DFO multi species surveys, individual specimen count
Cold Water Coral Biomass	N/A	N/A	DFO multi species survey, kg/standard tow
Cold Water Coral Diversity	N/A	N/A	DFO multi species surveys, Shannon's diversity evenness index
Cold Water Coral Richness	N/A	N/A	DFO multi species surveys, count/standard tow
Marine Mammal Presence	Aerial/shipboard sighting program	MarWhale data obtained from VDC	N/A
Sea Turtle Presence	Aerial/shipboard sighting program	MarWhale data obtained from VDC	N/A

Oceanographic Variables			
Bottom Temperature	N/A	BIO's Hydrographic Database, ^{0}C	DFO MEDS Surveys, ^{0}C
Bottom Temperature Span	N/A	BIO's Hydrographic Database, ^{0}C	N/A
Sea Surface Temperature	Satellite SeaWiFS, ^{0}C	BIO's Hydrographic Database, ^{0}C	Satellite SeaWiFS, ^{0}C
Sea Surface Temperature Span	Satellite SeaWiFS, ^{0}C	BIO's Hydrographic Database, ^{0}C	N/A
Water Column Stratification	Shipboard hydrographic measurements, sigma-t units	BIO's Hydrographic Database, mixed layer depth (m)	N/A

Zooplankton, nekton, demersal, and bottom temperature raster datasets were all interpolated using ordinary kriging (Goulet et al., 2010). Note that the nekton biomass dataset had extreme outliers removed before interpolation (Goulet et al., 2010). The cell size of the interpolated surfaces was dependent on the spatial distribution of samples from the original datasets, ranging from 2 km to 20 km. Both the zooplankton and nekton datasets were interpolated to a 20 km grid to match the resolution of the surveys, but the demersal and bottom temperature datasets were denser and interpolated to grids of 2 km or 2.5 km depending on the dataset (Goulet et al., 2010). To maintain consistency between the raster and vector datasets, all surfaces were converted to a 20 km grid using the aggregate tool in ArcGIS (ESRI, 2008). The value assigned to each 20 km cell was calculated by taking the mean values of all the original 2 km, or 2.5 km, cells within the larger 20 km cell. Certain datasets (sea surface temperature, primary production, Chl-*a*, and bathymetry) could not be brought to a 20 km grid using the aggregate tool as they were not an integer factor of 20, which is a requirement of the aggregate tool (ESRI, 2008). Therefore, to calculate the mean values of these variables within each 20 km cell the datasets were resampled, using the nearest neighbour algorithm, to create a raster that was an integer factor of 20. The factor value of 20 that was closest to the original cell size was used, for example, the sea surface temperature raster had a cell size of ~4.8 km so it was resampled to a 5 km cell size. Once this was completed, the resampled raster was aggregated into 20 km cells using the aggregate tool.

Another important process that had to be done with the vector datasets was to "clip" out the land from the datasets. Including the land in these datasets would create biases in the statistics of measurements. First the land polygon was converted to a raster grid of the same cell size as the interpolated data (20 km). A reclassification was then done on the land raster to change the raster value of land to no data and the value of water to 1. Doing this creates a water mask that can be multiplied by the satellite images to produce a product that only represents the portion of the images taken over the ocean. To ensure that water mask actually represented water along the coastal areas, a 20 km buffer from the coastline was performed on the land polygon.

The coral datasets were not originally interpolated because of the large number of observations that recorded no coral catch. Therefore the coral raster datasets were actually mean values within each 20 km x 20 km cell (Goulet et al., 2010). However, there were cells within these raster datasets that had a no data value, as opposed to zero values, and would result in a no data value in the PCA analysis. Therefore, a focal mean based on a 3x3 cell window was calculated for each no data cell in order to fill in these gaps and create a continuous surface.

All the data were standardized to a mean of 0 and standard deviation of 1 ([x – mean (x)] / s.d.; Zwanenburg et al. 2010). The Raster Calculator inside ArcGIS was used to perform this calculation on the raster datasets.

2.2 Analysis Methods

Principal components analysis (PCA) is often performed on high-dimensional data to eliminate redundancy, find patterns, emphasize variance within the variables, and improve interpretation (ESRI, 2008). Essentially, PCA transforms the data in multivariate space to a new multivariate space whose axis are rotated so that the greatest variance is explained by the first principal component, the second principal component (orthogonal to

the first) explains the second greatest variance, and so on. The first three or four principal components typically explain the most variance and by analyzing only these components, one reduces the number of dimensions without much loss of information. PCA was first performed on the normalized datasets using all sixteen available variables (all runs exclude geology as these data are not yet available). A second run of PCA was performed excluding the nekton and zooplankton datasets. These datasets had the smallest number of samples and the smallest spatial extent. By excluding these variables one can examine how these datasets may have influenced the analysis. Based on the results of the first two PCA analyses we determined that the sea surface temperature, primary production, and bathymetry dominated the signal. Therefore, a third run of PCA excluded these datasets along with excluding again the nekton and zooplankton datasets. The third run of PCA was done again (3B), this time also removing the only other remaining physical variable, bottom temperature. A fourth run of PCA was done excluding just bathymetry, sea surface temperature, and primary production while leaving in the nekton and zooplankton datasets (Table 3). Finally, a fifth run excluded information pertaining to corals; the high ratio of absence to presence data in the coral datasets (Goulet et al., 2010) may be a strong factor in determining the amount of variance explained by the PCA analysis.

Table 3: Variables used (marked with an X) in the three different runs of PCA.

Variables	Run #1	Run #2	Run #3	Run #3B	Run #4	Run #5
Bathymetry	X	X				X
Surficial Seabed Geology	N/A	N/A	N/A	N/A	N/A	N/A
Zooplankton Biomass	X				X	
Chlorophyll-a	X	X	X	X	X	X
Primary Production	X	X				X
Nekton Biomass	X				X	
Nekton Diversity	X				X	
Nekton Richness	X				X	
Demersal Fish Biomass	X	X	X	X	X	X
Demersal Fish Diversity	X	X	X	X	X	X
Demersal Fish Richness	X	X	X	X	X	X
Cold Water Coral Biomass	X	X	X	X	X	
Cold Water Coral Presence/Absence	X	X	X	X	X	
Cold Water Coral Diversity	X	X	X	X	X	
Cold Water Coral Richness	X	X	X	X	X	
Sea Surface Temperature	X	X				X
Bottom Temperature	X	X	X		X	X

The PCA results were used in a k-means clustering procedure to classify the data. K-means clustering is an unsupervised classification technique, meaning there is no prior knowledge on what information classes exist in the data. An information class is a similar grouping of values that are known to belong to a specific class, for example, in satellite imagery classification a class may be defined as a meaningful grouping of locations representing real world objects such as water or forest. A cluster, on the other hand, is simply a statistical grouping in the data with similar attribute values in multivariate space with no knowledge on what that cluster represents in the real world. These clusters must be interpreted into meaningful classes by the user.

The raster outputs of the first four principal components from the ArcGIS PCA analysis were used as input into k-means clustering using the algorithm of Legendre (2001). The number of clusters we investigated ranged from 2 to 10. Legendre's (2001) algorithm provides the optimal number of clusters as determined by the Calinksi-Harabasz (C-H) statistic, and the count of observations within each cluster. The C-H statistic is calculated for different number of clusters using the following equation (Legendre, 2001):

$$\text{C-H} = [R^2/(K-1)]/[(1-R^2)/(n-K)]$$

R^2 = (SST – SSE)/SST
SST = total sum of squared distances
SSE = sum of squared distances of the objects to their group's own centroids
K = number of groups

The number of clusters that yields the highest C-H criterion corresponds to the most compact set of clusters, or optimal number of groups (Legendre, 2001). The output from the k-means clustering was then mapped to visualize the distribution of clusters from each run.

3. Results

Table 4 to Table 9 present the results of the different PCA runs (see Table 3 for the input variables used in each run). Figure 3 to Figure 7 show plots of the first 3 principal components against each other for PCA runs #2, #3, #3B, #4, and #5. PCA run #5 provided the most rapid rise in explained variance. This run did not include the coral datasets; therefore, the high ratio of absence to presence data in the coral datasets (Goulet et al., 2010) may be a strong factor in determining the amount of variance explained by the PCA analysis. PCA runs #3 and #3B also provided a rapid rise in explained variance as these runs had the fewest variables and little or no physical variables.

Physical variables (i.e. primary production, sea surface temperature, and bathymetry) dominated the signal on the first two principal components in runs #1, #2, and #5 (Table 4, , Table 5, Table 9, Figure 3, and Figure 7). In run #1, variables related to corals and demersal surveys loaded on PC3, with fish more abundant in areas of low coral abundance and presence, and with demersal biomass and diversity in opposition to one another (Table 4). There were strong loadings of bathymetry on PC2 and PC3 of run #2 suggests some non-linearity in the influence of this variable on the overall multivariate variance (Table 5). We also note a positive loading of coral variables on PC2 and a negative loading on PC3, again suggesting that non-linear relationships among variables may be at play.

In PCA run #3 (Table 6) and #3B (Table 7) the first principal component (PC) is dominated by coral percent presence, diversity, and richness, while the second PC represents the influence of demersal biomass, diversity, and bottom temperature, with biomass correlated with bottom temperature and in opposition to diversity (Table 6 and Table 7). A possible interaction between coral distributions and catches from demersal surveys is most apparent on PC3 and PC4.

PCA run #4 had all the variables included except sea surface temperature, primary production, and bathymetry, and yielded the same general pattern as runs #3 and #3B (Table 8). We also note a weaker contribution of nekton variables on PC2, with biomass, diversity and richness loading inversely to those of demersal biomass.

PCA run #5 represented the information with the greatest spatial coverage. Sea surface temperature and primary production were positively correlated and in opposition to bathymetry on PC1 (Table 9). Bathymetry had a strong negative loading on PC2, as did primary production, while there was a weak positive loading of demersal richness. PC3 was dominated by positive loadings of demersal biomass and bottom temperature and a somewhat weaker negative loading of demersal diversity.

Table 4: PCA Run #1 results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**1.286**	**0.939**	**0.810**	**0.657**	**0.341**	**0.280**
Explained Variance (Percent)	**25**	**18**	**15**	**13**	**6**	**5**
Cumulative Variance	**25**	**43**	**58**	**71**	**77**	**82**
Bathymetry	-0.276	-0.711	-0.330	0.422	-0.030	-0.013
Bottom Temp	-0.119	0.163	0.228	0.369	-0.006	-0.166
Coral % Presence	-0.143	0.280	-0.370	0.119	0.235	0.086
Coral Biomass	-0.055	0.124	-0.102	0.113	-0.152	0.925
Coral Diversity	-0.123	0.302	-0.348	0.192	0.114	-0.190
Coral Richness	-0.149	0.335	-0.413	0.181	0.193	-0.058
Chlorophyll-a	-0.073	-0.131	-0.022	0.149	-0.125	-0.026
Demersal Biomass	-0.070	0.074	0.330	0.398	0.356	-0.018
Demersal Diversity	0.126	0.062	-0.348	-0.132	-0.535	-0.156
Demersal Richness	-0.220	0.258	0.030	0.177	-0.543	-0.190
Nekton Biomass	0.033	-0.017	-0.173	-0.154	-0.009	0.009
Nekton Diversity	-0.136	0.079	0.221	0.194	-0.188	0.049
Nekton Richness	-0.128	0.075	0.170	0.134	-0.245	0.062
Primary Production	0.554	-0.155	-0.140	0.425	-0.103	-0.002
Sea Surface Temp	0.642	0.187	0.013	0.260	-0.007	0.005
Zooplankton Biomass	-0.138	0.086	0.202	0.180	-0.202	0.037

Table 5: PCA Run #2 results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**1.241**	**0.930**	**0.747**	**0.572**	**0.325**	**0.280**
Explained Variance (Percent)	**27**	**20**	**16**	**12**	**7**	**6**
Cumulative Variance	**27**	**47**	**63**	**75**	**82**	**88**
Bathymetry	-0.333	-0.658	-0.538	0.188	-0.042	-0.007
Bottom Temp	-0.080	0.127	0.102	0.556	-0.249	-0.106
Coral % Presence	-0.155	0.333	-0.341	-0.011	0.312	0.055
Coral Biomass	-0.052	0.135	-0.122	0.076	-0.190	0.950
Coral Diversity	-0.130	0.349	-0.347	0.079	0.094	-0.187
Coral Richness	-0.160	0.392	-0.396	0.040	0.225	-0.074
Chlorophyll-a	-0.073	-0.131	-0.100	0.097	-0.088	-0.041
Demersal Biomass	-0.029	0.022	0.177	0.647	0.134	0.023
Demersal Diversity	0.094	0.108	-0.273	-0.329	-0.593	-0.105
Demersal Richness	-0.189	0.253	-0.007	0.202	-0.604	-0.168
Primary Production	0.558	-0.186	-0.394	0.194	-0.029	-0.024
Sea Surface Temp	0.672	0.134	-0.144	0.167	0.012	0.000

Table 6: PCA Run #3 results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**0.803**	**0.552**	**0.299**	**0.257**	**0.163**	**0.151**
Explained Variance (Percent)	**33**	**23**	**13**	**11**	**7**	**6**
Cumulative Variance	**33**	**56**	**69**	**80**	**87**	**93**
Bottom Temp	0.126	-0.563	-0.244	-0.095	0.200	-0.107
Coral % Presence	0.495	0.084	0.314	0.051	-0.198	0.224
Coral Biomass	0.194	-0.038	-0.179	0.954	0.007	-0.047
Coral Diversity	0.508	0.015	0.100	-0.180	0.232	-0.300
Coral Richness	0.576	0.058	0.230	-0.072	0.014	-0.035
Chlorophyll-a	-0.004	-0.060	-0.051	-0.032	-0.715	-0.691
Demersal Biomass	-0.008	-0.649	0.143	0.035	0.267	-0.222
Demersal Diversity	0.155	0.425	-0.592	-0.090	0.373	-0.343
Demersal Richness	0.296	-0.255	-0.611	-0.175	-0.380	0.448

Table 7: PCA Run #3B results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**0.796**	**0.414**	**0.274**	**0.249**	**0.157**	**0.147**
Explained Variance (Percent)	**37**	**19**	**13**	**12**	**7**	**7**
Cumulative Variance	**37**	**56**	**69**	**81**	**88**	**95**
Chlorophyll-a	-0.009	-0.066	-0.085	-0.105	-0.940	-0.303
Coral % Presence	0.505	-0.067	0.255	0.166	-0.155	0.413
Coral Biomass	0.191	-0.080	-0.586	0.771	-0.019	-0.048
Coral Diversity	0.509	-0.061	0.147	-0.134	0.135	-0.480
Coral Richness	0.582	-0.070	0.223	0.018	-0.014	-0.022
Demersal Biomass	-0.063	-0.728	-0.103	-0.081	0.225	-0.422
Demersal Diversity	0.184	0.647	-0.295	-0.143	0.150	-0.418
Demersal Richness	0.270	-0.164	-0.644	-0.567	0.001	0.389

Table 8: PCA Run #4 results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**0.812**	**0.724**	**0.317**	**0.268**	**0.256**	**0.154**
Explained Variance (Percent)	**28**	**25**	**11**	**9**	**9**	**5**
Cumulative Variance	**28**	**53**	**64**	**73**	**82**	**87**
Coral % Presence	-0.464	0.191	0.215	-0.250	0.073	0.057
Bottom Temperature	-0.208	-0.397	0.070	0.483	-0.141	-0.015
Coral Richness	-0.544	0.196	0.197	-0.120	-0.064	-0.020
Coral Biomass	-0.196	0.011	-0.111	0.218	0.939	-0.033
Coral Diversity	-0.483	0.150	0.139	0.032	-0.187	-0.097
Chlorophyll-a	-0.015	-0.075	-0.097	-0.103	-0.018	-0.969
Demersal Biomass	-0.085	-0.471	0.427	0.337	-0.001	-0.071
Demersal Diversity	-0.073	0.394	-0.479	0.441	-0.132	-0.103
Demersal Richness	-0.352	-0.203	-0.536	0.187	-0.180	0.162
Nekton Biomass	0.012	0.224	-0.029	-0.015	0.009	0.003
Nekton Diversity	-0.110	-0.330	-0.207	-0.293	0.033	0.026
Nekton Richness	-0.105	-0.259	-0.276	-0.349	0.043	0.029
Zooplankton Biomass	-0.120	-0.309	-0.223	-0.283	0.021	0.049

Table 9: PCA Run #5 results of the first six principal components. Cells shaded in grey highlight the higher eigenvector scores.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	**1.213**	**0.839**	**0.571**	**0.373**	**0.156**	**0.126**
Explained Variance (Percent)	**36**	**25**	**17**	**11**	**5**	**4**
Cumulative Variance	**36**	**61**	**78**	**89**	**94**	**98**
Bathymetry	-0.397	-0.822	0.158	-0.106	0.256	0.208
Bottom Temperature	-0.065	0.165	0.545	-0.294	0.172	-0.182
Chlorophyll-a	-0.087	-0.158	0.089	-0.090	-0.628	-0.516
Demersal Biomass	-0.035	0.113	0.666	0.114	0.222	-0.293
Demersal Diversity	0.127	-0.072	-0.386	-0.626	0.390	-0.504
Demersal Richness	-0.146	0.224	0.158	-0.692	-0.316	0.506
Primary Production	0.551	-0.448	0.170	-0.092	-0.379	-0.034
Sea Surface Temperature	0.699	-0.052	0.155	-0.051	0.256	0.246

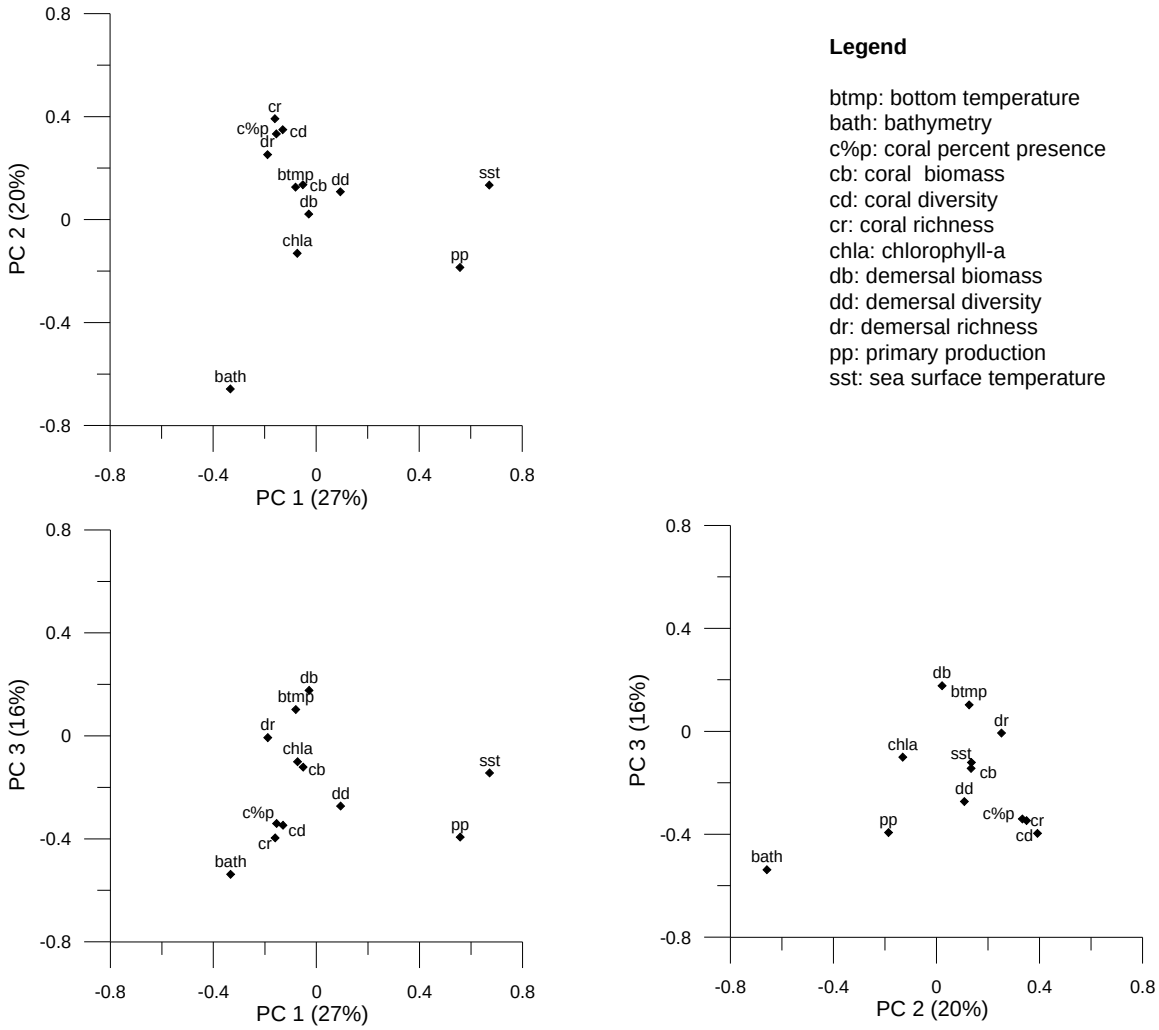

Figure 3: Plots of the first 3 principal components from run #2.

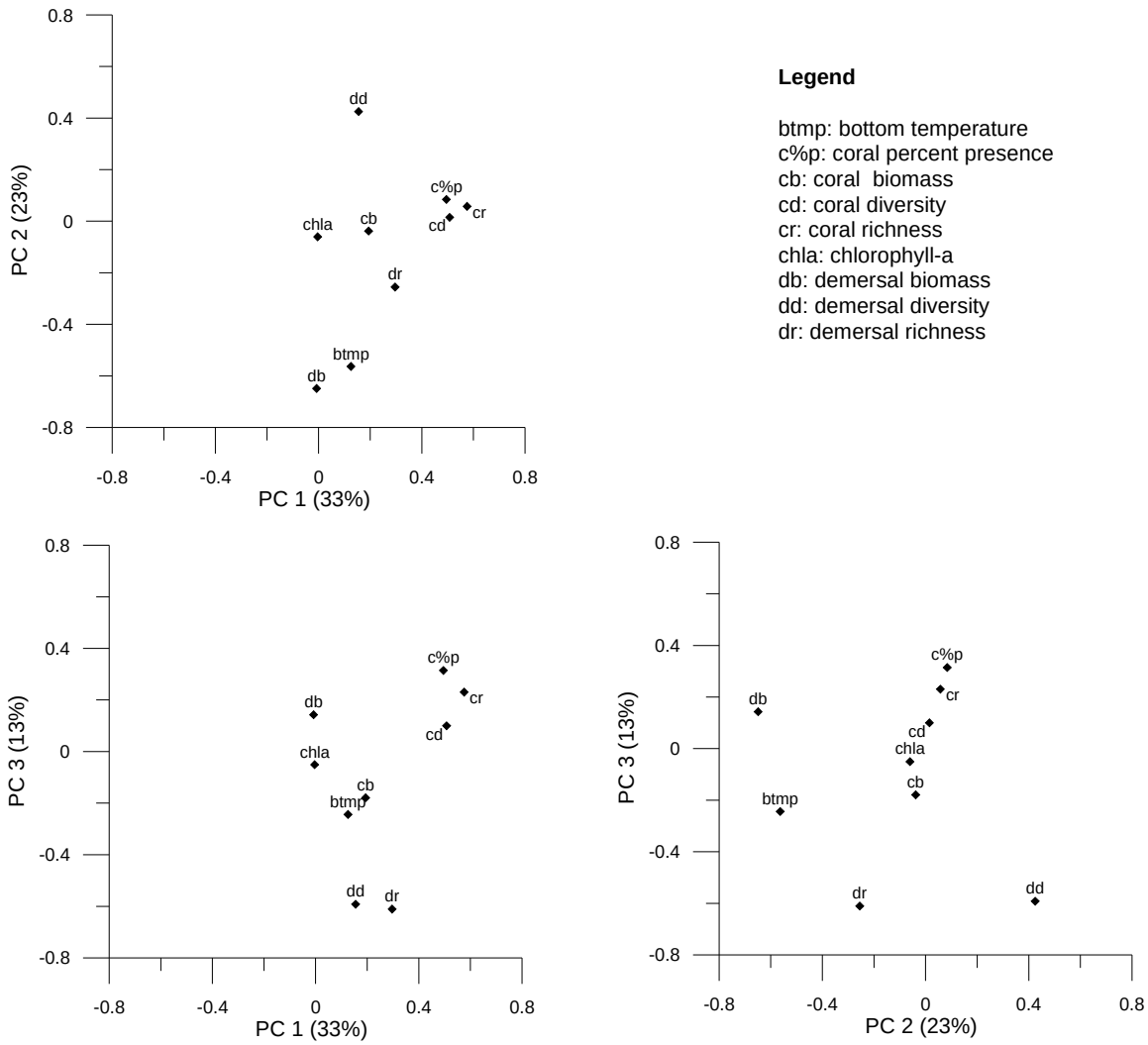

Figure 4: Plots of the first 3 principal components from run #3.

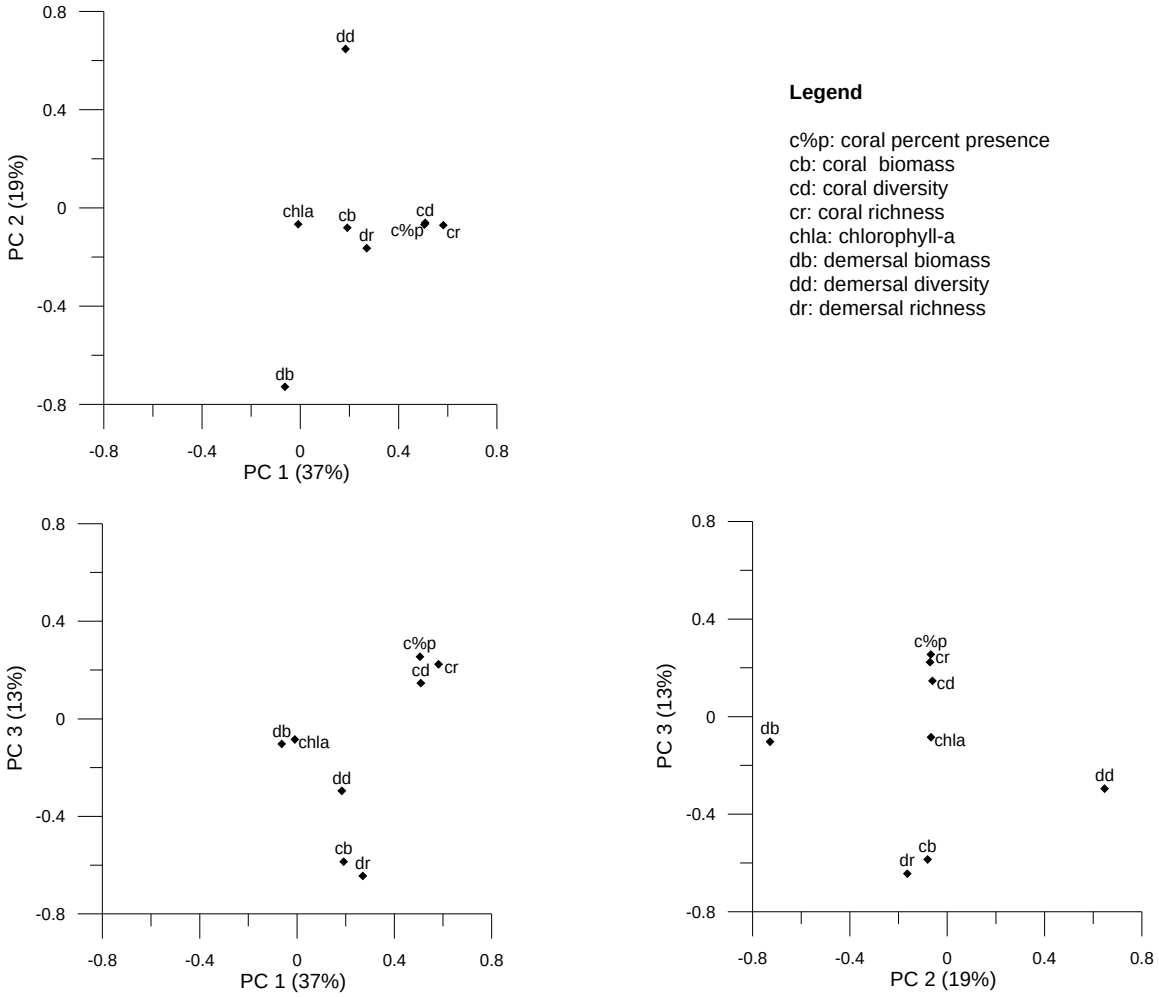

Figure 5: Plots of the first 3 principal components from run #3B.

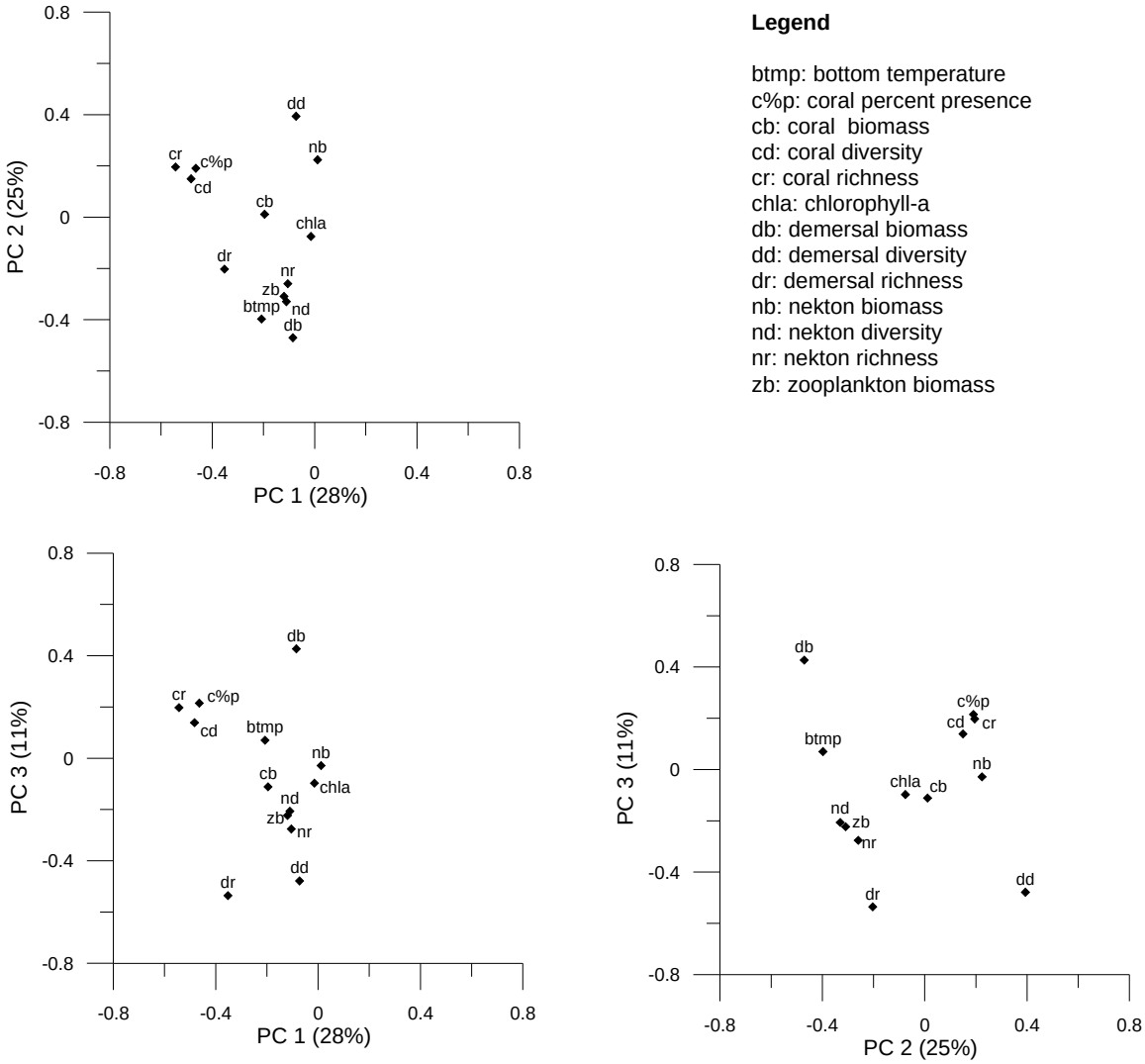

Figure 6: Plots of the first 3 principal components from run #4.

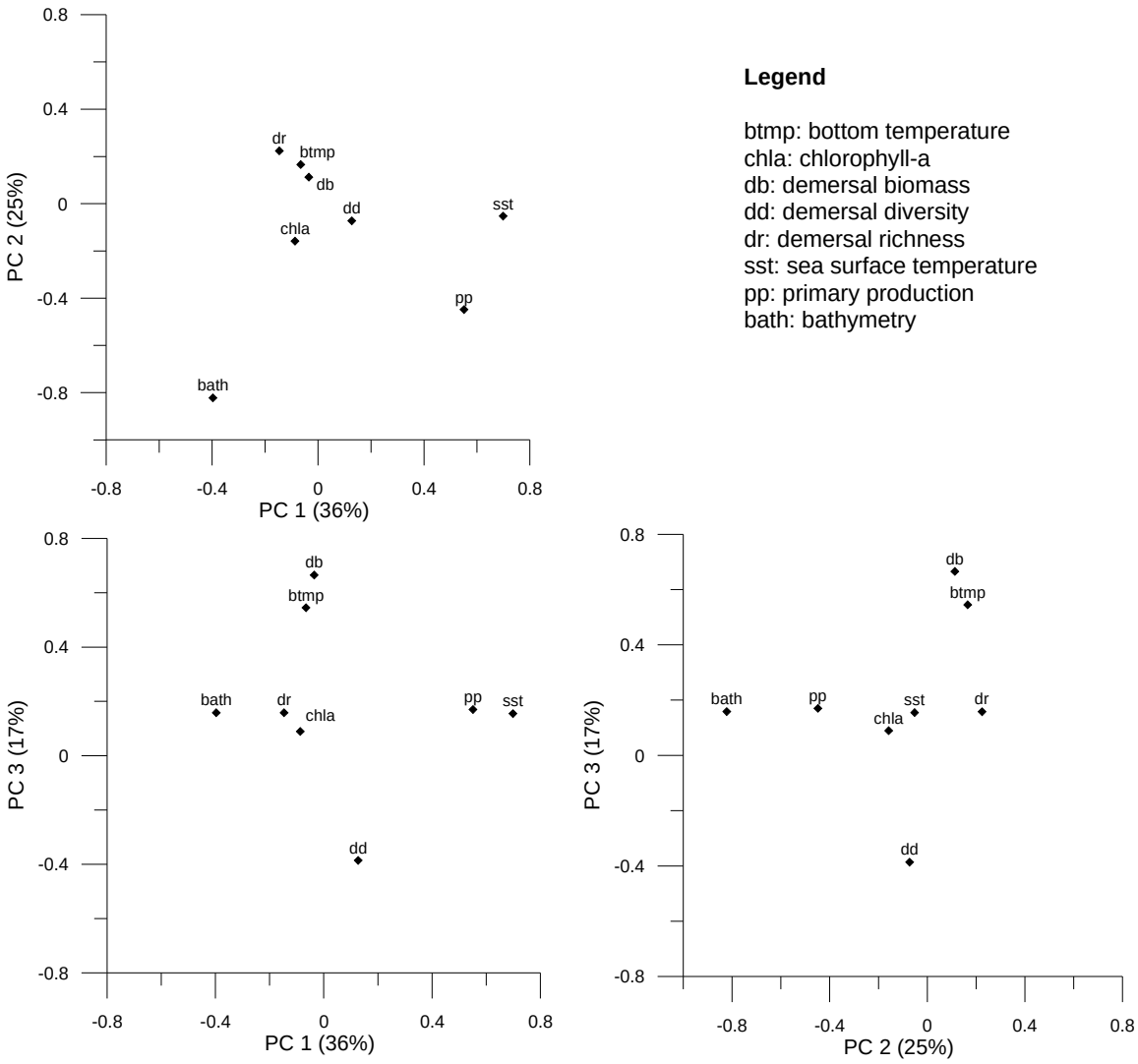

Figure 7: Plots of the first 3 principal components from run #5.

The first four PCs were used in an unsupervised k-means clustering algorithm to classify each cell to a specific cluster. The optimal number of clusters for the first four PCs for each PCA run was determined using the Calinski-Harabasz statistic (Legendre, 2001). The optimal number of clusters for PCA run #3 and #3B was six, five for run #2 and #5, and two for run #4 (Figure 8). The results of PCA run #1 were not mapped because the spatial extent was relatively small as a result of the inclusion of the nekton and zooplankton datasets, and the optimal number of cluster is only two which does not convey much information. The spatial distribution of k-means clustering results for four, five, and six clusters for runs 2 to 5 are mapped in Figure 10 to Figure 13. Presenting this range of clusters was chosen to reflect the jump from three to four clusters in Figure 8 and the maximum number of optimal clusters.

The numerical assignment of cluster classes are not comparable between different runs of k-means clustering (e.g. run #2 and run #3) or between different numbers of clusters within the same k-means run (e.g. run #2 with five clusters and run #2 with six clusters) (Figure 9). This is because k-means is an unsupervised classification algorithm with no prior knowledge on what information classes the clusters may represent. However, by examining the spatial distribution of the clusters between different k-means runs or between different numbers of cluster for the same k-means run, one can infer that certain clusters represent the same class on different maps. For example, in Figure 10 cluster 1 for the k-means run with four clusters, cluster 2 for the run with five clusters, and cluster 1 for the run with six clusters, all seem to cover the same area and are likely representing the same class.

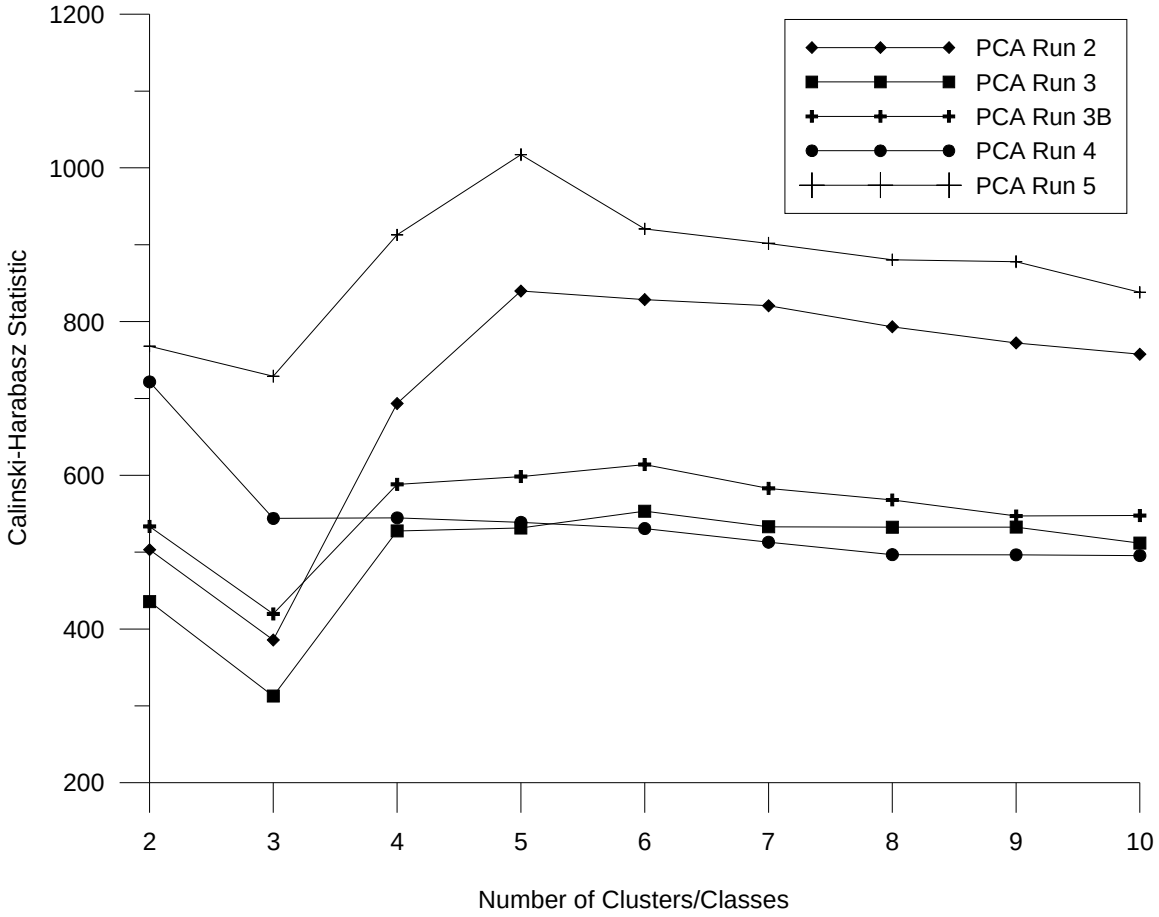

Figure 8: Plot of the Calinski-Harabasz statistic for each number of clusters. Maximum values indicate the optimal number of clusters.

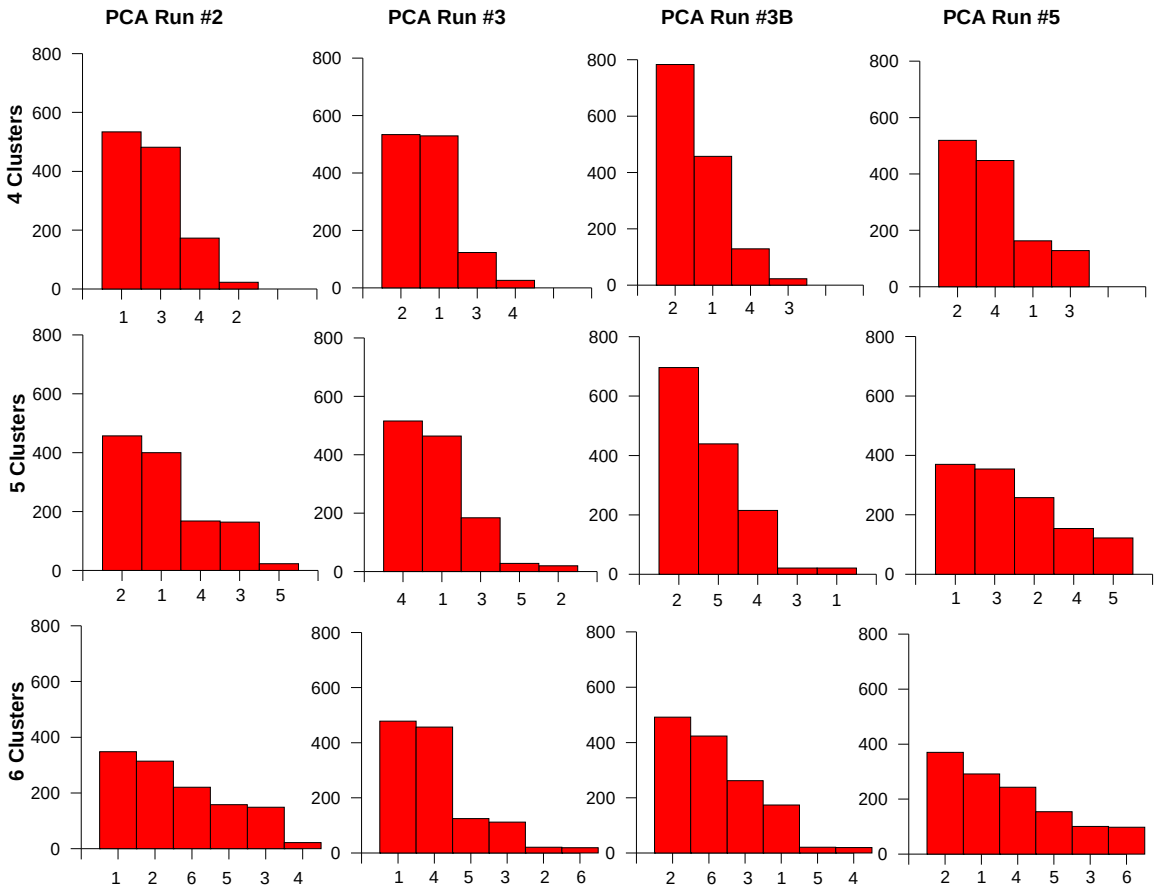

Figure 9: Frequency of occurrence for each cluster from the k--means clustering analysis. PCA runs #2, #3, #3B, and #5 are shown. PCA run #4 is not shown as the optimal number of clusters was two. The clusters are ordered from highest occurrence to lowest. Note that the cluster numbers are not comparable between the different runs or for different number of clusters as k-means clustering is an unsupervised classification.

There are some common outcomes to all analyses that differ in the spatial extent of the different clusters identified from the analysis among PCA runs (Figure 10 to Figure 13). First, the area of the northeast Newfoundland Shelf and the southeast Shoal cluster together irrespective of the variables included in the analysis, except in runs #2 and #5 with 5 or 6 clusters (Figure 10, Figure 13). These latter analyses reveal when the relative contribution of physical variables over biological indicators becomes most apparent in the PCA analysis by distinguishing the northeast Newfoundland Shelf from the Southeast Shoal. Second, the area of the northern and western Grand Banks tends to form a distinct cluster which includes the coastal areas of the northeast Newfoundland Shelf except when a large number of clusters are used in runs #2, #3B and #5 in which coastal areas appear as a distinct cluster. Finally, the area of the continental slope also forms a distinct cluster in all runs. The spatial extent of the cluster is narrowest off Labrador and northern Newfoundland and broadest in the area of Orphan Knoll, the nose of the Grand Banks and Flemish Pass. However, many rasters on the continental shelf (single rasters and small groups) were assigned to this cluster. This may reflect in the presence or absence of corals in some parts of the Newfoundland region, as demonstrated by the absence of rasters assigned to the slope cluster when corals are removed from the analysis in run#5 (Figure 13). In all analyses, allowance of 6 clusters results in greater resolution of the marginal areas near the coast and along the continental slope.

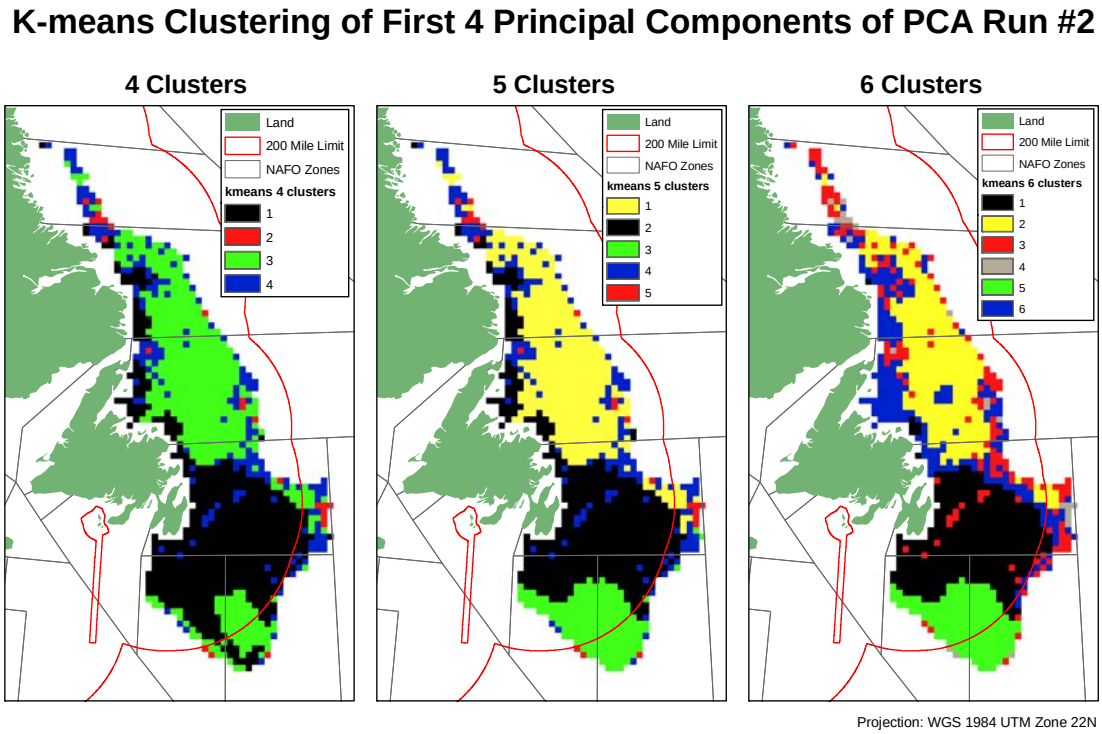

Figure 10: Map of k-means clustering results of the first four principal components of PCA run #2 with nekton and zooplankton datasets removed.

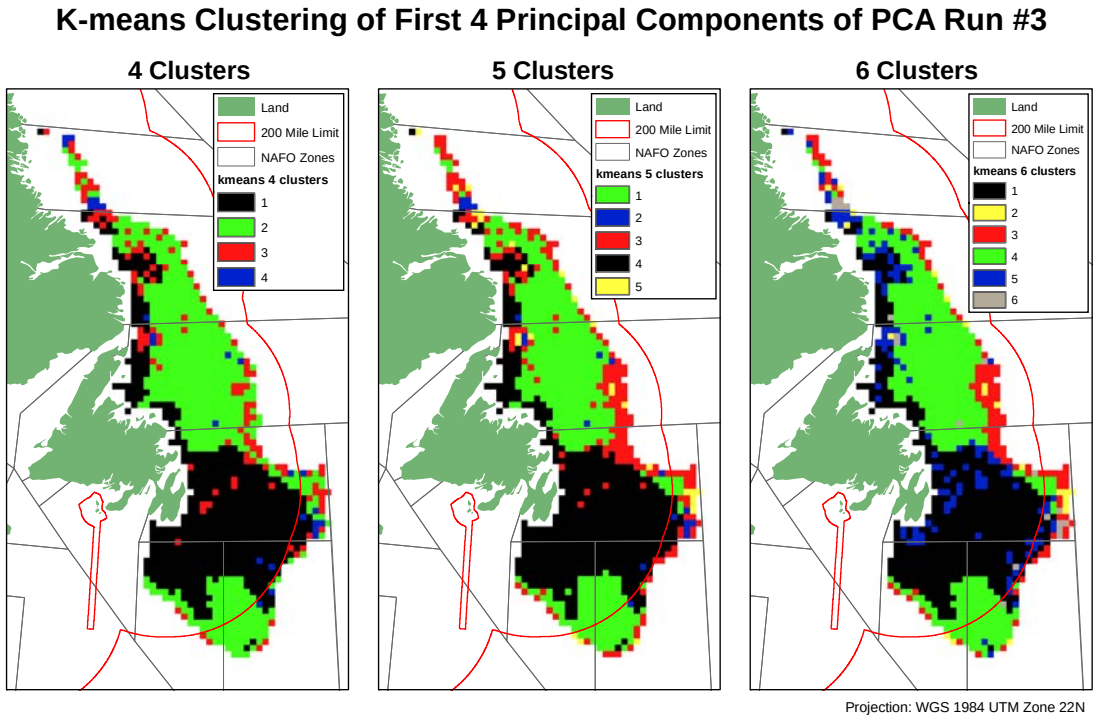

Figure 11: Map of k-means clustering results of the first four principal components of PCA run #3 with nekton, zooplankton, and physical datasets (except bottom temperature) removed.

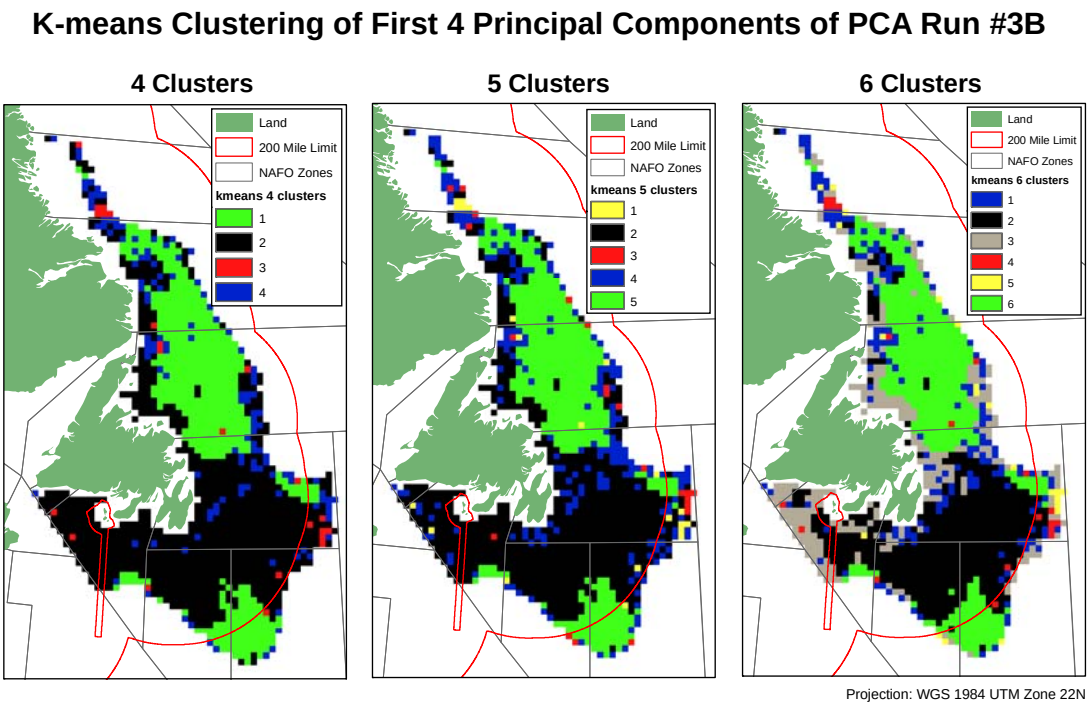

Figure 12: Map of k-means clustering results of the first four principal components of PCA run #5 with nekton, zooplankton, and coral datasets removed.

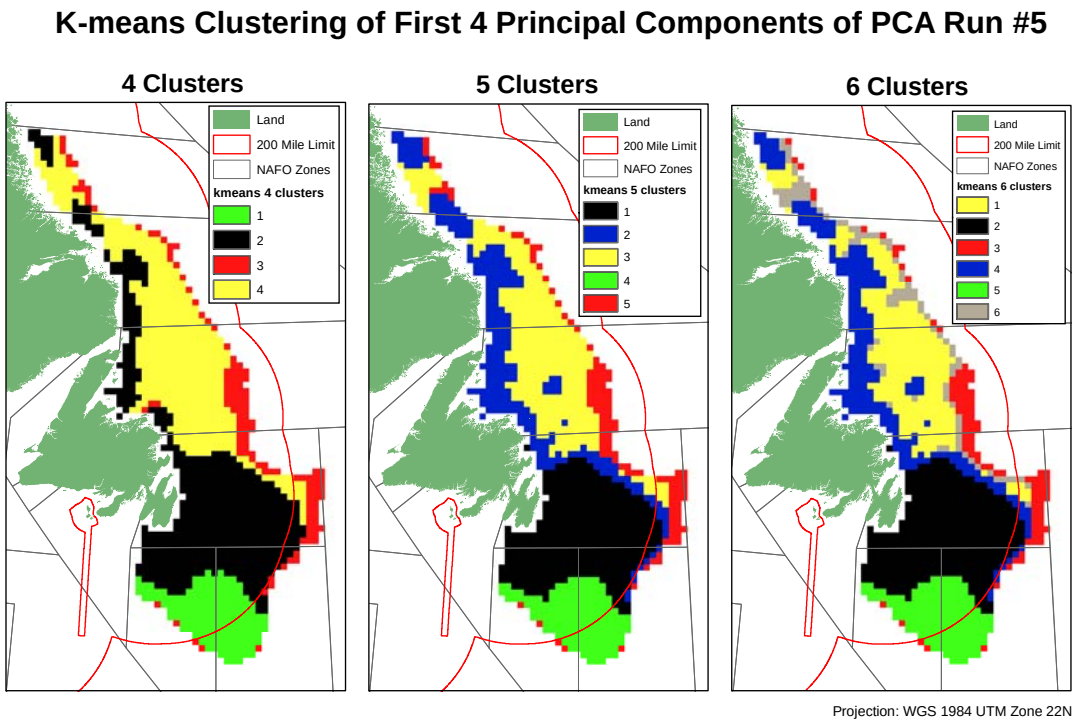

Figure 13: Map of k-means clustering results of the first four principal components of PCA run #5 with nekton, zooplankton, and all coral datasets removed.

4. Discussion

This research has applied, as closely as possible, the same methods used to define Ecoregions on both the Canadian Scotian Shelf (Zwanenburg et al., 2010) and US Northeast Continental Shelf (Fogarty and Keith, 2009). For the most part, the same variables were used with the exception of a few (see Table 2 for a direct comparison). There is a certain amount of uncertainty in the process used to identify ecoregions as a result of the methods used and limitations of the data. The first source of uncertainty comes from the interpolation process. Certain datasets (demersal fish, nekton fish, zooplankton, and bottom temperature) were interpolated using the kriging method (Goulet et al., 2010) and, as with any interpolation method, the interpolated values are basically educated guesses of unknown values based on certain models. The results come with a certain amount of uncertainty and this is represented as a root mean squared error in the Goulet et al. (2010) report. There were also issues with varying spatial and temporal resolutions of the datasets. The nekton and zooplankton datasets covered the smallest area and had the coarsest spatial resolution (Goulet et al., 2010). Therefore, the remainder of the datasets had to be resampled to fit the limited spatial resolution of the nekton and zooplankton datasets. The were also limitations in the years covered by each of the datasets (Table 1), therefore, data covering longer time periods will be more representative of broadscale trends than those collected over a shorter .

The PCA results can be summarized by a few obvious trends. Bathymetry, primary production, and sea surface temperature were strong driving variables when included in the PCA analysis. These variables essentially reflect physical attributes of ocean circulation features in the region that are measured with relatively little uncertainty when averaged over long periods of time. When these variables were absent, the coral and demersal datasets dominated the PCA signal.

The clustering results of PCA run #2 (with only nekton and zooplankton datasets removed) and PCA run #5 (with nekton, zooplankton, and coral datasets removed) clustered the NL shelf and the southeast shoal as different clusters. In contrast, PCA runs #3 (nekton, zooplankton, SST, PP, bathymetry removed) and PCA run #3B (same as run #3 with bottom temperature removed) clustered them as having similar attributes. The contrast between runs #2, #3 and #3B highlights the influence of the physical variables in determining the difference between these two areas in k-means clustering results. In all runs, the Grand Banks remained its own class separate from the NL shelf and the southeast shoal. The “patchy” class along the edge of the clustering results of all PCA runs, except run #5, occurred on the continental slope, and is most likely caused by the variables associated with the description of corals. This was lost in run #5, where coral variables were removed, which likely reflects the influence of different spatial scales of coral habitats relative to those describing the physical features of the environment and the temporally averaged distribution of demersal fish and phytoplankton.

A critical aspect of these analyses comes from the relatively simplistic description of the biological variables using biomass, diversity and richness that resulted in the northeast Newfoundland Shelf and the southeast Shoal appearing to have the same attributes. Although the two areas are similar in their general metrics, the species found in each part of the continental shelf differ substantially. This raises a cautionary aspect to the definition of ecoregions using summary variables. Once clusters are identified based on multivariate analyses, approaches based on ecosystem, community or taxonomic diversity must be considered to determine if all elements of a cluster are based on functionally similar communities. Creating GIS layer(s) that reflect metrics of taxonomic similarity (or dissimilarity) could serve to add an important element in the definition of ecoregions that would include aspects of ecosystem structure that reflects the patterns of biodiversity across the region of interest. The metrics of taxonomic similarity should be based on the results of separate (unconstrained) ordination analyses of community structure derived from interpolated maps of the distribution of individual species. What weight is given to dominant versus rare or keystone taxa would have to reflect all aspects of the conservation objectives associated with the definition of ecoregions.

There are a number of future research paths that could be taken here. One option is to try and smooth the clustering results using the “Boundary Clean” and “Filter” tools in ArcGIS; this would be consistent with the approach taken by Fogarty and Keith (2009). The inputs for the PCA and clustering analysis could also be changed to examine the influence on results when including and excluding different variables. Demersal datasets were binned into three multi-year intervals (Goulet et al., 2010); however, these datasets were not examined here. These datasets could be examined to assess the influence of change over time of the demersal survey variables. Finally, there are also other unsupervised classification algorithms that could be used to classify the PCA results. Different clustering

procedures can influence how objects are grouped into clusters and how a cluster is defined (Legendre and Legendre, 1998). ArcGIS also provides an unsupervised classification technique whereby the ISO Cluster tool creates a signature file containing information on group assignment and the Maximum Likelihood tool takes this data creates a classified map. The latter enables the user to set a minimum class size, which could eliminate some of the smaller clusters and aggregate them into larger clusters, thereby simplifying the result.

References

DFO. (2009). Development of a Framework and Principles for the Biogeographic Classification of Canadian Marine Areas. DFO Canadian Science Advisory Secretariat. Science Advisory Report, 2009/056, 17pp.

ESRI, Inc. (2008). ArcGIS Version 9.3.1. Desktop Help for ArcGIS release 9.3. Environmental Systems Research Institute (ESRI), Redlands, CA, USA.

Fogarty, M.J. and Keith, C. (2009). Delineation of Regional Ecosystem Units on the U.S. Northeast Continental Shelf. NEFSC Discussion Paper.

GeoEye Inc. (2010). http://www.geoeye.com/CorpSite/

Goulet, P., Walsh, M., Cuff, A., Devillers, R., and Edinger, E. (2010). Data analysis towards GIS-based ecoregion delineation in the NL shelf, 1995 – 2007. Final Report for Fisheries and Oceans Canada, Marine Geomatics Lab, Memorial University of Newfoundland, 43pp.

Legendre, P. (2001). Program K-means user's guide. Département de sciences biologiques, Université de Montréal. 11 pages.

Legendre, P. and Legendre, L. (1998). Numerical Ecology (2nd Edition). Elsevier Science, Amsterdam, The Netherlands.

NAFO. (2010). Report of the NAFO Scientific Council Working Group on Ecosystem Approach to Fisheries Management (WGEAFM). Serial No. N5815. NAFO Scientific Council Summary Document, 10/19, 101pp.

O`Reilly, J.E., et al. (2000). SeaWiFS postlaunch calibration and validation analysis, Part 3. In Hooker, S.B. and Firestone, E.R. (Eds.), SeaWiFS Postlaunch Technical Report Series, Volume 11. NASA Goddard Space Flight Center.

Platt, T., Sathyendranath, S., Forget, M-H., White III, G.N., Caverhill, C., Bouman, H., Devred, E., Son, S.H. (2008). Operational estimation of primary production at large geographical scales. Remote Sensing of Environment, 112, 3437 – 3448.

Zwanenburg, K., Horsman, T., and Kenchington, E. (2010). Preliminary analysis of biogeographic units on the Scotian Shelf. NAFO Scientific Report Doc. 10/06.