Bioinformatic approaches for objective detection of water masses on continental shelves

Matthew J. Oliver, Scott Glenn, Josh T. Kohut, Andrew J. Irwin, and Oscar M. Schofield
Coastal Ocean Observation Lab, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, New Jersey, USA

Mark A. Moline
Biological Sciences, California Polytechnic State University, San Luis Obispo, California, USA

W. Paul Bissett
Florida Environmental Research Institute, Tampa, Florida, USA

Received 29 July 2003; revised 29 October 2003; accepted 23 December 2003; published XX Month 2004.

[1] As part of the 2001 Hyper Spectral Coupled Ocean Dynamics Experiment, sea surface temperature and ocean color satellite imagery were collected for the continental shelf of the Mid-Atlantic Bight. This imagery was used to develop a water mass analysis and classification scheme that objectively describes the locations of water masses and their boundary conditions. This technique combines multivariate cluster analysis with a newly developed genetic expression algorithm to objectively determine the number of water types in the region on the basis of ocean color and sea surface temperature measurements. Then, through boundary analysis of the water types identified, the boundaries of the major water types were mapped and the differences between them were quantified using predictor space distances. Results suggest that this approach can track the development and transport of water masses. Because the analysis combines the information of multiple predictors to describe water masses, it is an effective tool in detecting water masses not readily recognizable with temperature or chlorophyll alone. *INDEX TERMS:* 4283 Oceanography: General: Water masses; 4552 Oceanography: Physical: Ocean optics; 4546 Oceanography: Physical: Nearshore processes; 4842 Oceanography: Biological and Chemical: Modeling; 4899 Oceanography: Biological and Chemical: General or miscellaneous; *KEYWORDS:* remote sensing, water mass, fronts

Citation: Oliver, M. J., S. Glenn, J. T. Kohut, A. J. Irwin, O. M. Schofield, M. A. Moline, and W. P. Bissett (2004), Bioinformatic approaches for objective detection of water masses on continental shelves, *J. Geophys. Res.*, *109*, C07S04, doi:10.1029/2003JC002072.

1. Introduction

[2] Water mass analysis is an active area of research because of their potential utility for describing large-scale ocean circulation [*Warren*, 1983], assessing the impact of river plumes [*Højerslev et al.*, 1996], understanding basin-scale biogeochemistry [*Broecker and Takahashi*, 1985]. Water masses are classically defined as waters with common formation and origin having similar conservative properties such as temperature and salinity. However, it should be noted that this conservative requirement means that for temperature and salinity to remain conservative within a mass of water, the water mass cannot be in contact with the surface ocean or its source region. The introduction of the T-S diagram was the first quantitative approach to defining water masses on the basis of their conservative properties and has been a mainstay in the oceanographic community [*Helland-Hansen*, 1916]. Since that time, oceanographers have used chemical isotopes to further study the circulation of water masses in the ocean interior [*Broecker and Peng*, 1982]. In the surface ocean where temperature and salinity are not considered conservative, injections of dyes and SF_6 have been successfully used to track the circulation and subduction of surface features because the presence of SF_6 can be considered conservative compared to some of the short-timescale process in the surface ocean [*Upstill-Goddard et al.*, 1991]; however, this type of research is costly and can effectively cover only relatively small space scales. To assess the impact of broad-scale surface features, the key is to develop proxies that change over larger timescales than the processes being studied.

[3] To a certain degree, optical oceanographers have addressed the issues of water mass identification in the surface ocean by classifying them on the basis of their optical properties. Efforts by *Jerlov* [1968] classified waters into nine water types. These water types were further analyzed by *Morel and Prieur* [1977] and classified into the widely accepted Case 1 and Case 2 waters. These classifications have been an extremely useful tool. Water types are different than water masses in that water types

0148-0227/04/2003JC002072$09.00

occupy only similar predictor space while water masses occupy similar predictor and physical space [*Tomczak*, 1999]. A major objective over the last few decades has focused on understanding global and basin-scale circulation, which operate over timescales of years to thousands of years. Therefore these processes require tracers that are relatively conservative over the same timescales (i.e., salinity). However, if the timescale of interest in detecting and tracking near surface water masses is on the order of hours to days as it often is in coastal regions, optical predictors potentially provide additional dimensions of discrimination to traditional temperature and salinity analysis. This type of optical approach has been demonstrated by tracking river influence containing anthropogenic pollutants [*Højerslev et al.*, 1996]. In addition to tracking anthropogenic pollutants, the identification of frontal regions between water masses has been used to identify important areas of mixing and biological activity [*Claustre et al.*, 1994].

[4] Although simple in concept, the inclusion of optics as a water mass tag presents a problem in determining the uniqueness of a water mass. Because water mass classification has traditionally relied upon hydrographic predictors only, there exists an intuitive sense, based on a century of experience, for defining significant differences in temperature and salinity predictors before discriminating between water masses. While these discriminations are inherently subjective, the inclusion of optical predictors only confounds the already subjective interpretation. This problem is not unique to oceanography, but a fundamental problem for any scientific field that assigns categories or identifiers to a known data continuum. Therefore, if optical predictors are to be used effectively in water mass analysis and identification, an objective mathematical construct is needed for proper quantitative discrimination of water masses based on the similarity of water types [*Martin-Trayovski and Sosik*, 2003].

[5] One branch of science that has had to develop means to overcome the problems associated with assigning categories to a known continuum is the field of evolutionary and molecular biology. These problems manifest themselves in a variety of ways such as uncertainties in phylogenetic trees, species determination [*Hey*, 2001; *Wu*, 2001; *Noor*, 2002], annotations of genomes [*Meeks et al.*, 2001] and the expression of genes [*Yeung et al.*, 2001]. This problem has become more complex with technological breakthroughs such as DNA microarrays and automatic sequencers, and through necessity, the rapidly advancing field of bioinformatics has endeavored to produce several objective mathematical constructs to transform a data continuum into meaningful categories. This manuscript applies techniques developed by the bioinformatics field and adapts them for the use of objective water mass analysis and classification in a coastal region. We present a mathematical construct of a water mass classification method and apply it to the Mid-Atlantic Bight during the summer of 2001 using optical parameters measured by SeaWiFS and sea surface temperature measured by AVHRR satellite sensors.

2. Methods

[6] During the 2001 HyCODE experiment at the Long-term Ecosystem Observatory (LEO) off southern New Jersey, daily SeaWiFS and AVHRR passes were collected with an L band data acquisition system at approximately 1 km resolution over an area defined at 38.50°–41.50°N latitude and 76.00°–71.00°W longitude (Figure 1). These satellites were used as an adaptive sampling tool during the experiment so that data of the relevant hydrographic features in the region could be collected. Pixels from the single daily SeaWiFS pass were matched to the least cloud covered AVHRR pass using latitude and longitude. Morning AVHRR passes were used to avoid the effects of diurnal solar heating. Cloud removal was accomplished by adjusting the cloud coefficient in the MCSST algorithm. SeaWiFS data were processed using the DAAC algorithm. For this study, matched satellite passes from 14, 21, and 31 July and 2 August 2001 were chosen because of relatively little cloud cover. Each composite matrix of SeaWiFS and AVHRR imagery had between 75,000 and 105,000 cloud free pixels. Each composite matrix was subsampled at 6 km resolution for the analysis to increase computational speed, and to match the resolution of the surface current measurements in the region. These data were analyzed in a multistep process that identifies predominant water mass boundaries and the gradients between water masses (Figure 2).

2.1. Data and Standardization

[7] The data used from the composite matrix of AVHRR and SeaWiFS in this study were sea surface temperature (SST, °C), remote sensing reflectance measured at 490 nm ($R_{rs(490)}$) and at 555 nm ($R_{rs(555)}$) (Figure 1). Remote sensing reflectance is a quasi-inherent optical property defined as the ratio of upwelling radiance (W m^{-2} sr^{-1}) to downwelling irradiance (W m^{-2}) and has units of sr^{-1}. These data were chosen for two reasons. First, they are used in chlorophyll and primary productivity estimations. Second, a principal components analysis using the correlation matrix on the combined 4-day data set including SST and remote sensing reflectance at 412 nm, 443 nm, 490 nm, 510 nm, 555 nm and 670 nm indicated that three linear combinations described 96.6% of the variance of the data. SST, $R_{rs(490)}$ and $R_{rs(555)}$ were the largest contributors to these linear combinations. This suggests that the majority of the waters in this analysis are Case 1 and that the other remote sensing reflecting channels are highly correlated and would not add much discrimination power. Note however, the methods described in this paper are not limited to three predictors or these specific satellite products; however in this region they represented the most useful data. Work in other areas may require some similar preliminary analysis. SST, $R_{rs(490)}$ and $R_{rs(555)}$ were standardized for this analysis by subtracting their respective means and dividing by their respective standard deviations from the combined data from the 4 days. This process weighted each predictor equally for any potential water mass present.

2.2. Clustering Algorithms

[8] Four different clustering algorithms were used simultaneously in this analysis (Table 1). These algorithms were two agglomerative or hierarchical clustering algorithms, a K means and a fuzzy C means algorithm (see *Quackenbush* [2001] for a review). From the subsampled data set, each pixel (observation) was projected into three dimensional standardized predictor space. The agglomerative clustering

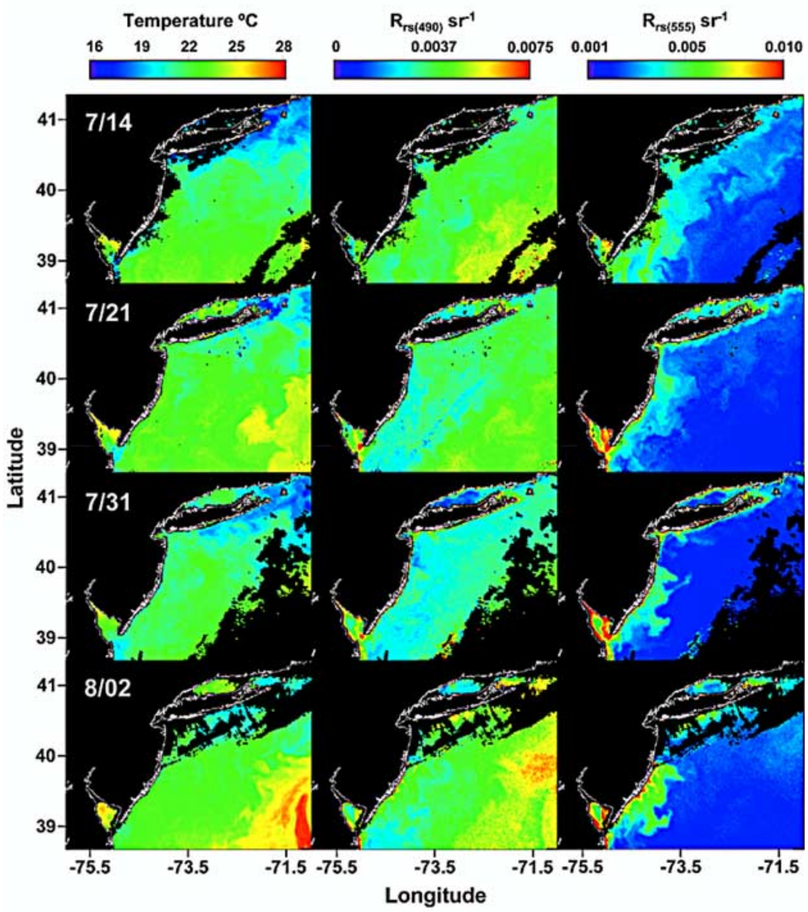

Figure 1. Temperature and reflectance maps on 14, 21, and 31 July and 2 August 2002 in this analysis. A warm-core ring is evident on 2 August as a nearshore optically dominated water mass formed nearshore. The white line is the coastline, and the black indicates land or cloud.

algorithms grouped observations in three dimensions according to their Euclidian distance in standardized predictor space. The agglomerative clustering types grouped standardized predictor data hierarchically from n to 2 clusters from closest to furthest in predictor space where n is the number of observations. The difference between how the two agglomerative clustering algorithms treated the data is based on how the data was grouped in predictor space. The first agglomerative clustering type grouped data according to complete linkage (i.e., agglomerative complete linkage (ACL)), which determined that two clusters of data ought to be joined to a single cluster based on the maximum distance between cluster edges. The second agglomerative method grouped data according to Ward's linkage (i.e., agglomerative Ward's linkage (AWL)) [*Ward*, 1963]. This method calculated the total sum of squared deviations from the cluster means, and joins clusters to minimize the increase of the total sum of squares deviation. The K means clustering algorithm is a divisive clustering algorithm, which requires a user-specified cluster number. This algorithm initialized cluster centers randomly and grouped data until the within-cluster sum of squares is minimized for the number of clusters specified [*Hartigan and Wong*, 1979]. The fuzzy C means clustering algorithm is similar to the K means clustering algorithm except that through the use of fuzzy logic and sequential competitive learning, observations are clustered [*Chung and Lee*, 1994].

[9] While there are dozens of clustering schemes, these particular algorithms were chosen on the basis of performance from the literature. *Yeung et al.* [2001] observed that on real data, using agglomerative clustering with single linkage (clusters joined into a single cluster based on the minimum distance between clusters) did not produce sensible clusters of data. Rather, the K means clustering algorithm performed very well. The ACL algorithm has been cited as very useful in producing tightly grouped clusters [*Quackenbush*, 2001]. In our opinion this is a good feature for water type identification because there is an emphasis in grouping only the most similar data. The choice of the AWL algorithm was related to previous work done by *Oliver et al.* [2004], in which a priori knowledge of the number of water masses present fit well with the results of the AWL algorithm. The fuzzy C means clustering algorithm was chosen on the basis of the results of *Chung and Lee* [1994], which showed that the competitive learning done by the fuzzy C means algorithm produced sensible clusters.

2.3. Figure of Merit

[10] A major difficulty in cluster analysis is determining how many clusters (or water types in this case) should be

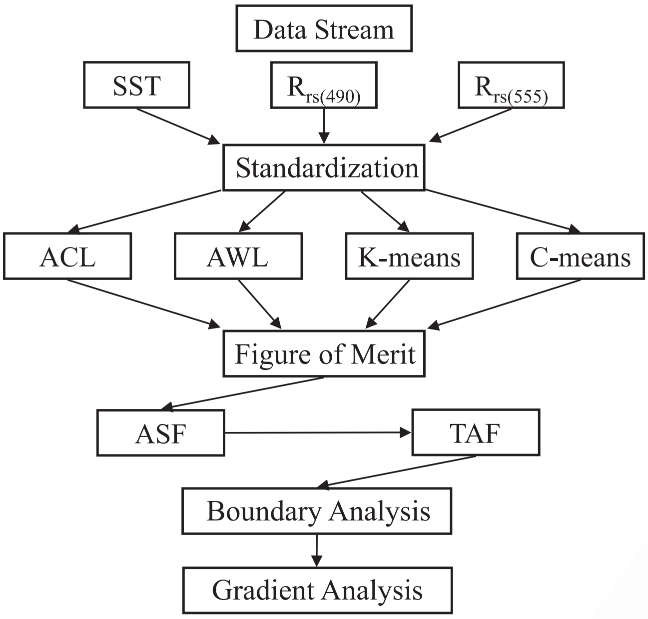

Figure 2. Flow diagram of this analysis. This analysis assimilates sea surface temperature as well as two remote sensing channels for all 4 days. The data are standardized according to the mean and variance of the combined 4-day data set to make them comparable. Water types for each day are detected using four clustering algorithms, ACL, AWL, K means, and C means. These results are combined into a Figure of Merit, where an average slope function (ASF) and threshold of acceptable flatness (TAF) are computed. These two predictors give a range of reasonable water types. For each solution for each day the boundaries are plotted, and coincident boundaries are the most prevalent, indicating similar structures found by different clustering algorithms. This indicates that the boundaries associated with this water type indicate a prevalent water mass. Finally, the predictor space distance is measured between each data point to determine how different the water is on either side of each boundary. High values indicate a very strong boundary between water masses.

used to describe a data set as each observation could theoretically represent its own cluster. Therefore a means to analyze this structure objectively was required to identify water types in predictor space. With the advent of rapid gene sequencing and gene expression chips, the field of bioinformatics has endeavored to produce and continues to refine several algorithms that analyze gene and expression data in order to find patterns of gene expression that are linked to a variety of factors. *Yeung et al.* [2001] developed and validated one such method which essentially computes the RMS deviation between individual observations and the mean of the cluster they belong too for a given algorithm. This statistic is called the figure of merit (*FOM*). Although this algorithm was designed to calculate the difference between expression vectors of genes, 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. The *FOM* statistic was used to analyze the inherent structure defined by the clustering algorithms. The equation used in this study to calculate the *FOM* was:

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

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.

[11] Ideally, the *FOM* function will exhibit a distinct "elbow", decreasing rapidly at small k and much more slowly beyond a threshold k. This elbow represents the ideal cluster number (or number of water types in this case) for a data set because the deviation between cluster means and the individual observations in each cluster become very small. While the *FOM* statistic often show very distinct

Table 1. Description of the Four Types of Clustering Algorithms Used

Clustering Algorithm	Description
Agglomerative Complete Linkage (ACL)	Data are hierarchically grouped from n to 2 clusters. Data are grouped from closest to farthest on the basis of Euclidian distance in predictor space. The distance between clusters is measured on the basis of the maximum distance between cluster edges in predictor space.
Agglomerative Ward's Linkage (AWL)	Data are hierarchically grouped from n to 2 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.

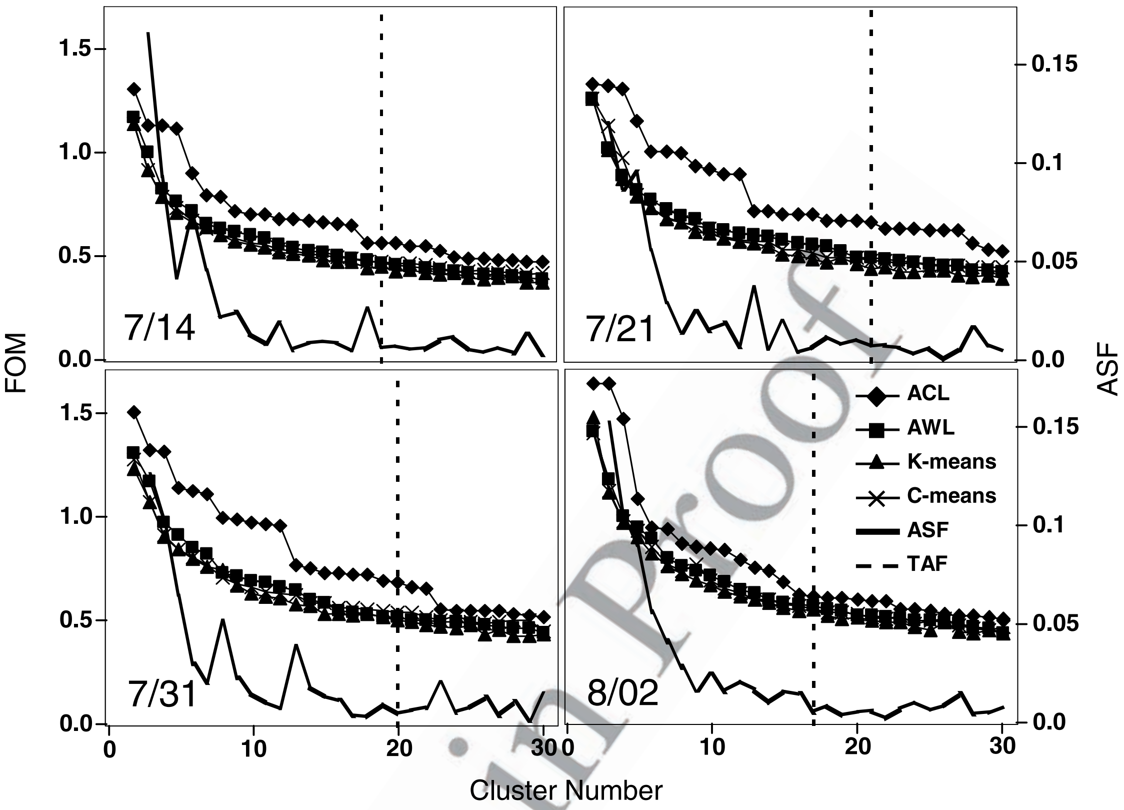

Figure 3. Figure of merit (*FOM*), average slope function (*ASF*) and threshold of acceptable flatness (*TAF*) calculation for each of the 4 days with the results of each of the clustering algorithms. A large *FOM* indicates that the variance within each cluster is comparatively large and that the cluster 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 within 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 *FOM* was less than 1% for more than three clusters.

"elbows" in simulated data sets, real data sets tend to show no distinct elbow for any of the clustering algorithms (Figure 3) [also see *Yeung et al.*, 2001, Figures 1 and 3]. In cases using real data, the *FOM* is best approximated by a power function of the number of clusters indicating that it is difficult to choose the ideal number of clusters. In this study, a threshold of acceptable flatness (*TAF*) of the *FOM* was defined by calculating the normalized average slope function ($ASF(k)$) of the *FOM* function at each cluster k for the four clustering algorithms using:

$$ASF(k) = \frac{1}{4}\sum_{c=1}^{4}\frac{FOM(c,k+1) - FOM(c,k)}{FOM_{\max}(c)} \qquad (2)$$

where $FOM_{\max}(c)$ is the maximum *FOM* value for a specific cluster algorithm c. The *TAF* was defined at the smallest cluster k where $ASF(k) < 0.01$ (<1% decrease in *FOM* relative to the maximum *FOM*) for three or more consecutive clusters. On the basis of our own observations in which k was allowed to approach n, an $ASF(k)$ value < 0.01 indicates that the variance within each cluster no longer reduces appreciably with increasing cluster number. This established an upper bound for what we believed to be reasonable cluster numbers or water type assignments by the suite of clustering algorithms. For this study, k was limited to a maximum of 30 clusters, as the *FOM* value did not change significantly after this cluster number.

2.4. Boundary Analysis

[12] One major difference between the clustering of a gene data set and a water mass data set is that clusters defined in a water mass data set occupy predictor space represented by standardized SST, $R_{rs(490)}$ and $R_{rs(555)}$ and physical space represented by latitude and longitude while a gene data set has no physical space representation. Water mass definitions vary slightly, so for the purposes of this analysis, our definition of a water mass is that it must occupy physical space, and water with similar properties in separate physical spaces represent different water masses. The spatial attributes of water masses provide additional useful information not generally associated with genes, and provide a useful means in delineating the physical boundaries between waters that have similar properties identified by the cluster analysis. 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 similar water types. Because each of the clustering algorithms is slightly different, the bound-

aries described at any specific cluster number k between water types may be different. However, it was clear that different clustering algorithms often had similar boundary solutions at different cluster numbers. This is because different water types were differentiated at slightly different cluster numbers because of differences in the clustering algorithms. 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 $2 \leq k \leq TAF$. To detect the most common water mass boundaries for any cluster number, the cluster number gradient in latitude and longitude space was computed using:

$$\nabla C_{xykc} = \sqrt{\left(\frac{C_{xykc} - C_{x+\Delta x, ykc}}{\Delta x}\right)^2 + \left(\frac{C_{xykc} - C_{y+\Delta y, xkc}}{\Delta y}\right)^2} \quad (3)$$

where x is longitude, y is latitude, C_{xykc} is the cluster number assignment for k clusters for c clustering algorithm and ∇C_{xykc} is the magnitude of the cluster number gradient vector. Where ∇C was nonzero, it was replaced with a logical value of 1 to indicate the presence of a boundary using:

$$b_{xykc} = \begin{cases} 1 \text{ if } \nabla C_{xykc} \neq 0 \\ 0 \text{ if } \nabla C_{xykc} = 0. \end{cases} \quad (4)$$

where b_{xyck} is the logical boundary value for a given longitude and latitude for the given cluster algorithm for k clusters. Although it is nonsensical to calculate gradients of categorical data, this method effectively detects the boundaries of the water masses. A 2-D histogram was constructed of high-frequency boundaries for each of the 4 days using:

$$B_{xy} = \frac{\sum_{c=1}^{4} \sum_{k=2}^{TAF} b_{xyck}}{4(TAF - 1)} \times 100\% \quad (5)$$

where B_{xy} is the frequency that a boundary (0–100%) at a given longitude and latitude. This 2-D histogram describes the most common physical boundaries between similar water types defined by the clustering algorithms. The presence of a high-frequency boundary was interpreted as a boundary between separate water masses.

2.5. Gradient Analysis

[13] In addition to determining where the major water mass boundaries are, the relative strengths of these boundaries were also estimated. Theoretically, water types could be distinctly separated in predictor space, but still be relatively close to each other in predictor space. In this case a boundary on a physical map between these water types would be drawn frequently between these distinct water types, while their differences would still be relatively minor. The purpose of the gradient analysis was to determine how different water types were in predictor space in relation to geographic space. The relative strength of the boundaries was defined as:

$$D_{x\to x+\Delta x} = \sqrt{\left(SST'_x - SST'_{x+\Delta x}\right)^2 + \left(R'_{rs(490)x} - R'_{rs(490)x+\Delta x}\right)^2 + \left(R'_{rs(555)x} - R'_{rs(555)x+\Delta x}\right)^2} \quad (6)$$

$$D_{y\to y+\Delta y} = \sqrt{\left(SST'_y - SST'_{y+\Delta y}\right)^2 + \left(R'_{rs(490)y} - R'_{rs(490)y+\Delta y}\right)^2 + \left(R'_{rs(555)y} - R'_{rs(555)y+\Delta y}\right)^2} \quad (7)$$

$$\nabla G(x, y) = \sqrt{\left(\frac{D_{x\to x+\Delta x}}{\Delta x}\right)^2 + \left(\frac{D_{y\to y+\Delta y}}{\Delta y}\right)^2} \quad (8)$$

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

2.6. Current Structure of the Region

[14] Surface current maps, measured by an HF radar system, provide a dynamical context in which to evaluate the placement of water mass boundaries. The long-range HF radar system used here was first deployed in 2001 [*Kohut and Glenn*, 2003], and consists of four remote transmit/receive sites along the coast of New Jersey and a central processing site in New Brunswick, New Jersey. Using the scatter of radio waves off the ocean surface each remote site can measure the surface current component moving toward or away from the site [*Barrick et al.*, 1977]. Information from all four remote sites is then geometrically combined at the central site to provide a total vector current map. The systems are operating at a frequency of about 5 MHz, which provides range out to 200 km offshore, a total vector grid resolution of 6 km and a surface current averaged over the upper 2.5 m of the water column. Each current map is a three hour average. For this analysis, the 3-hour data were averaged for 21 and 31 July and 2 August. Current data for 14 July were not yet available. If a particular range cell did not have at least 60% coverage over each day, the current vector in that range cell was not used in the analysis. A simple drifter experiment, which modeled 48 drifters along a boundary on 31 July, was used to determine if local advective processes could explain the changes in the boundary location during these days. This exercise attempts to predict the frontal location 51 hours later on 2 August. The current field was interpolated to the position of each drifter. The three hour average current maps were assimilated sequentially. At hourly intervals, the location of the drifter was evaluated and a new vector

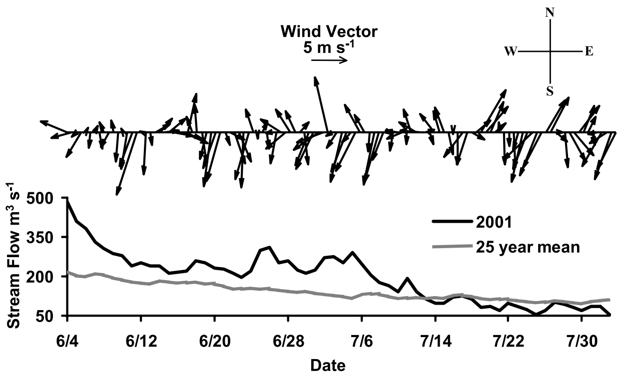

Figure 4. Wind record from the RUMFS field station and Hudson River flow recorded at Waterford, New York, during the study time period. From 14 July to 2 August there were three upwelling favorable events that may have sustained phytoplankton growth nearshore. The elevated streamflow during this particular year recorded at Waterford, New York, may have initiated the formation of a Hudson River-derived water mass during the 4-day study period. It has been reported that water outflow from this area takes 40 days to reach the Southern New Jersey shore [*Yankovski and Garvine*, 1998].

was assigned to the drifter. At three hour intervals a new current map was assimilated.

3. Results

[15] This study focused on a series of four composite images of SST, $R_{rs(490)}$ and $R_{rs(555)}$ from 14 July to 2 August 2001. During this period, a phytoplankton bloom developed in the northern portion of the study site and dispersed alongshore to the south (M. A. Moline et al., Episodic forcing and the structure of phytoplankton communities in the coastal waters of New Jersey, submitted to *Journal of Geophysical Research,* 2003, hereinafter referred to as Moline et al., submitted manuscript, 2003). Offshore, part of a Gulf Stream warm-core ring was observed on August 2 as it propagated from east to west (Figure 1). The phytoplankton bloom may have been associated by terrestrial runoff and was sustained by several upwelling events. Outflow from the Hudson River, one of the largest sources of terrestrial runoff in this region, measured at the Waterford, New York, site prior to the satellite passes was up to a factor of 2 larger than the 25 year mean during that time period (Figure 4). *Yankovski and Garvine* [1998] have shown that the time lag of these outflows to reach the study area is approximately 40 days, which coincides with the time with a large outflow from the Hudson River of this study (approximately 4 June). In addition, this time period had several upwelling favorable wind patterns on or around 19, 26, and 30 July. These upwelling wind events are regular in this region and stimulate phytoplankton growth [*Schofield et al.*, 2002; Moline et al., submitted manuscript, 2003].

3.1. Evaluation of the Figure of Merit

[16] For each of the days, *FOM* was calculated from $k = 2$ to 30 clusters for the four clustering methods (Figure 3). These *FOM* functions were generally decreasing with increasing cluster number in all cases and were similar to those found by *Yeung et al.* [2001] in that no distinct "elbow" was obvious. In all *FOM* cases, the ACL clustering algorithm was slightly higher than the other three clustering algorithms. While not producing exactly the same *FOM* statistic, the AWL, K means and C means clustering algorithms were very similar within days. *FOM* curves between days were similar in shape, however they differed slightly in magnitude. The $ASF(k)$ function for these days showed the most rapid decrease occurred where $k < 10$. In addition, all of the $ASF(k)$ functions display erratic changes in value where $10 < k < 15$. For $k > 15$, the $ASF(k)$ functions in all 4 days flattened noticeably. The *TAF* value for 14, 21, and 31 July and 2 August were 19, 20, 24, and 20 clusters, respectively. These values served as the upper bound for the boundary analysis.

3.2. Location and Strengths of Common Water Mass Boundaries

[17] The *FOM* analysis of the water types defined by the four clustering algorithms indicated that the "ideal" number of water types (clusters) was in the range of $2 \leq k \leq TAF$. For each c and k, k water types were defined that had boundaries described by equations (3) and (4) in physical space. Equation (5) is the frequency of these boundary observations across all c and k. A boundary frequency map (B_{xy}) was computed for each of the 4 days (Figure 5). In general, water mass boundaries become more defined from 14 July to 2 August. The most frequent boundaries are associated with strong optical or temperature fronts. Figure 6 illustrates the boundary frequency differences between the 4 days. As a function of total boundaries drawn on a map, high-frequency boundaries ($B_{xy} > 60\%$) were more spatially common on 31 July and 2 August compared to 14 and 21 July. Also, low-frequency boundaries ($0\% < B_{xy} < 20\%$) are more common on 31 July and 2 August compared to 14 and 21 July. These two conditions cause the 31 July and 2 August B_{xy} maps to appear more cleanly defined. In contrast, medium-frequency boundaries ($20\% < B_{xy} < 60\%$) were more common on 14 and 21 July compared to 31 July and 2 August, causing the 14 and 21 July maps to appear more cluttered. On 21 and 31 July and 2 August, when boundaries are more distinct, the major water masses are associated with the nearshore plume, shelf water, and water east of the shelf break front and the warm-core ring.

[18] The objective of the cluster analysis was to describe the inherent structure and separation of water types in predictor space, which was then mapped in the form of boundaries in Figure 5. The purpose of the gradient analysis was to determine how different water types were in predictor space in relation to geographic space. Figure 7 is the application of equations (6), (7), and (8) to evaluate the relative strengths of the boundaries between water masses. Because each pixel is slightly different from its neighbors, the gradient is never zero. The median value for this gradient calculation for this study is approximately 10, with a standard deviation of about 10. Therefore a strong gradient has a value in excess of 20 for this study. On 14 and 21 July gradients between water masses defined in the boundary analysis are relatively weak indicating that the water types found in these days are fairly similar. In

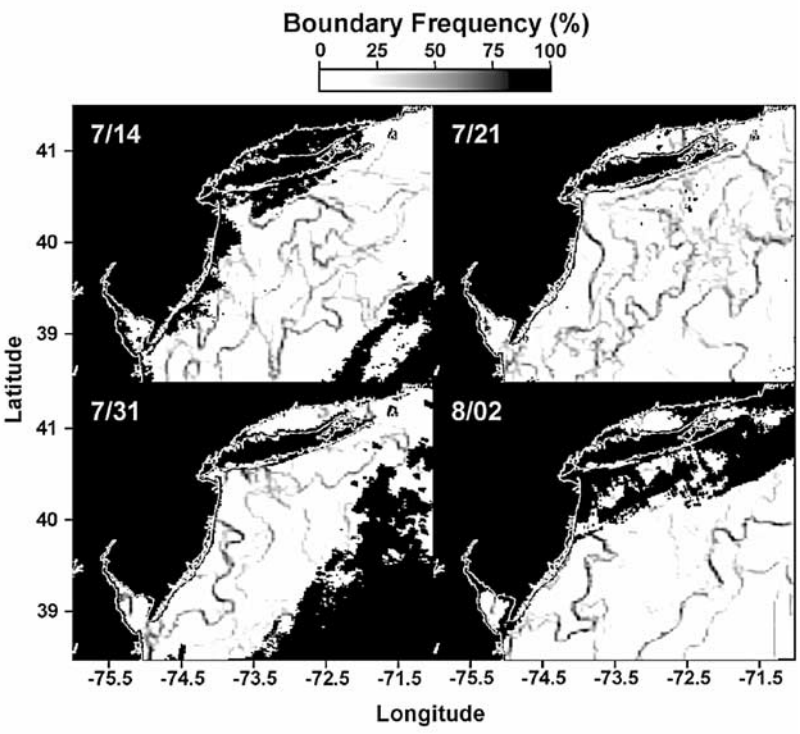

Figure 5. High-frequency boundary locations as calculated from equation (5). The contrast indicates how often a particular pixel was designated as a boundary. The most frequent boundaries represent water types that are easily separable in predictor space. Boundaries become more distinct from 14 July to 2 August.

contrast, strong gradients were found associated with the nearshore optical front. These relatively strong gradients are coincident with the high-frequency boundaries described in Figure 5 indicating that these particular water types are structurally distinct and very different. In addition, strong gradients were detected near clouds which may be a result of inadequate cloud masking.

3.3. Surface Current Structure, Gradient Strengths, and Boundary Locations

[19] The seasonal mean flow in the summer time in this region is along shore toward the south [*Kohut and Glenn*, 2003], which was generally observed in the 3-hour average flow on 21 and 31 July and 2 August. However, the flow structure on these dates was highly variable. The current fields in Figure 8 represent the flow field at the time of the satellite over pass with the spatial mean subtracted from it. This was done to visually enhance the fine-scale current structure associated with the water mass boundary gradients. Generally speaking, gradients were associated with physical features in the flow fields such as horizontal sheer, indicating that these features were strongly influenced by advective processes. However, the strength of the gradient was not related to the strength of the horizontal sheer, nor were all horizontal sheers associated with gradients.

[20] To determine if the apparent movement of the boundary was associated with physical advection, a simple simulated drifter experiment was performed (Figure 9). 48 modeled drifters were placed along the frontal boundary on 31 July and sequentially assimilated the surface current fields in hourly time steps. The predicted position of the major boundary feature was generally in good agreement with the location of the boundary on 2 August. The predicted boundary has a more pronounced ‘‘hammer-head’’ appearance much like that of the boundary on

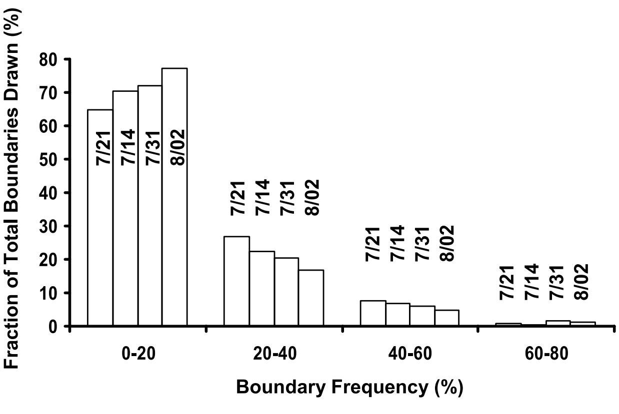

Figure 6. The boundary frequency calculated by equation 5 related to the total number of boundaries drawn. The days with more disorganized boundaries (14 and 21 July) have less low-and high-frequency boundaries and more medium-frequency boundaries. This causes the disorganized look on these days and indicates that the clustering algorithms had a difficult time coming to similar solutions. Days 31 July and 2 August had more low-frequency and high-frequency boundaries and low medium-frequency boundaries indicating that the clustering algorithms were in agreement more often and that water types were consistently distinguished.

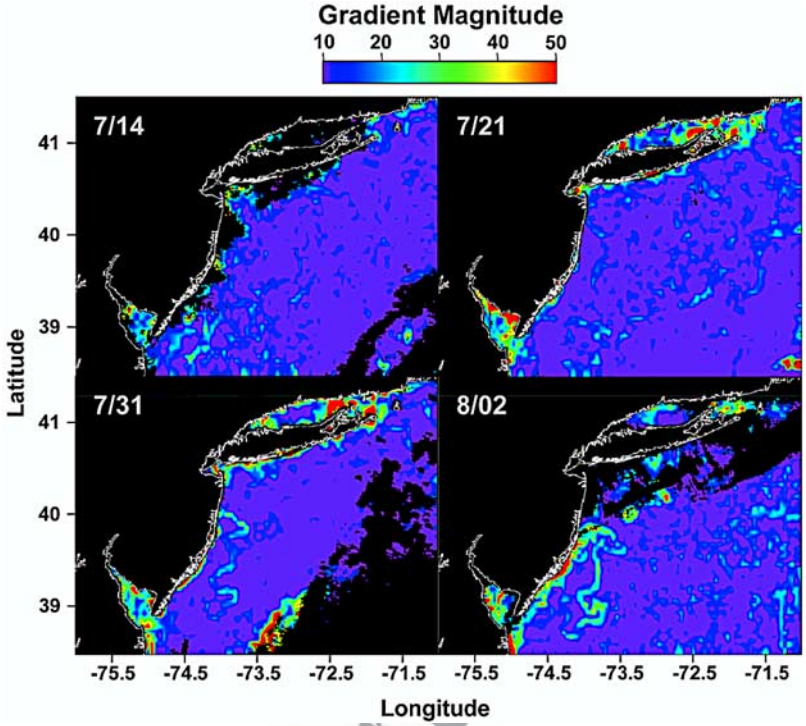

Figure 7. The gradient defined by equations (6), (7), and (8). The gradients are a relative measure of how different adjacent water masses are. Because no two adjacent pixels are equal, the gradient is never zero. The background gradient value for this study is approximately 10, with a standard deviation of approximately 10. Gradient values larger than 20 in this study are considered to be significant. Stronger gradients were evident in days 31 July and 2 August. This indicates that the water types on either side of the boundary are markedly different. However, strong gradients are not necessarily coincident with high- or medium-frequency boundaries because two water types may be readily distinguishable in predictor space but still be relatively close to one another.

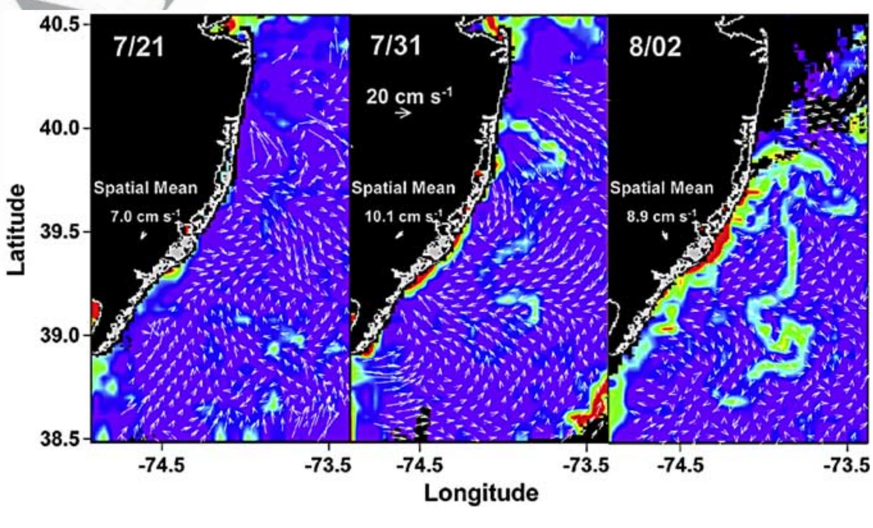

Figure 8. Boundary gradients overlaid with surface current fields with the surface current spatial mean subtracted for visual clarity. Areas with larger gradients are coincident with convergent and divergent areas, indicating that local current structure accounts for the gradient locations. However, not all convergent areas had gradients associated with them.

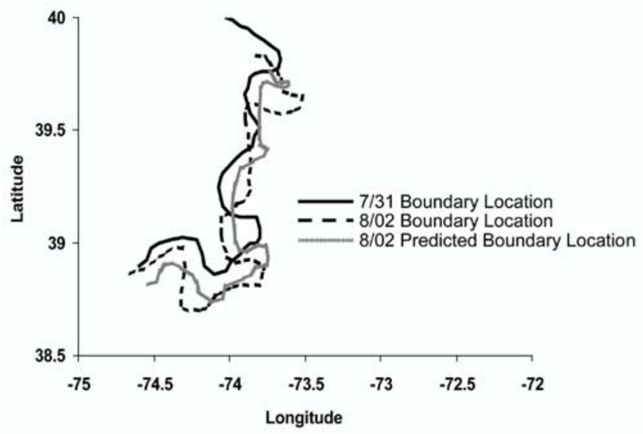

Figure 9. Results of simulated drifter experiment. The predicted location of 48 drifters on 2 August based on the initial position of the 31 July boundary by assimilating the CODAR measured surface currents generally approximates the location and shape of the boundary on 2 August. This indicates that the apparent movement of the boundary can be generally attributed to local advective processes. Also, this indicates that water masses in this area can be tracked effectively.

2 August. In addition the northern protrusion of the front moved southward, approximating its location on 2 August. Because the predicted position of the boundary region approximates the location of the boundary on 2 August, it suggests that local advection processes are largely responsible for changes between 31 July and 2 August.

4. Discussion

[21] AVHRR and ocean color satellite products are used to measure or infer several ocean processes. These include the tracking of the Gulf Stream [*Auer*, 1987], the modeling of Gulf Stream rings [*Glenn et al.*, 1990] and to estimate global ocean primary production [*Behrenfeld and Falkowski*, 1997]. New production in an ocean system has also been estimated through the combination of AVHRR and ocean color [*Sathyendranath et al.*, 1991]. To estimate new production, water types were defined intuitively, to which an idealized biomass profile was assigned. Conceivably, errors could be introduced in this type of approach if the way in which water types were defined was incorrect. *Karabashev et al.* [2002] addressed the water type problem through K means cluster analysis of SeaWiFS data; however, the number of clusters chosen (k = 20) was subjective.

[22] More recently, *Martin-Traykovski and Sosik* [2003] show very convincingly that there exist distinct optical water types in the Mid-Atlantic Bight region, and that they can be successfully discriminated. Their study developed a feature-based classification based on remote sensing reflectance in three wave bands and used a training set of data with known water types to develop classifiers. The method was evaluated on the ability of the classifiers to properly classify pixels into the correct categories. A goodness of fit measure was used as a measure for determining how variable the water is within each water mass. This method works very well if some a priori knowledge about the water types or water masses present is available. The *FOM* approach builds on this technique and does not require a training set of data, or prior knowledge of the water masses present, as it strictly looks for inherent structure in the data. Additionally, the method allows for the estimation of the strengths of the fronts between water types in physical space and temporal changes in boundary locations due to local advective processes. The *Martin-Traykovski and Sosik* [2003] method provides a solid foundation for water mass classification from space and complements this effort as the methods could be run in conjunction to elucidate water mass characteristics based on derived satellite products.

[23] In general, the water masses detected in this study were a nearshore plume, a water mass over the continental shelf separated by the shelf break front, water offshore the shelf break front and a warm-core ring. As for their origins, we can only speculate as satellites only detect their surface expressions. The nearshore water mass is most likely from the Hudson River, but it could also be upwelled water driven by southwest winds (S. M. Glenn et al., Biogeochemical impact of summertime coastal upwelling in the Mid-Atlantic Bight, submitted to *Journal of Geophysical Research,* 2003) The origin of the shelf water is from glacial melt along the southern Greenland coast that flows south to the MAB as a buoyant coastal current [*Beardsley and Winant*, 1979; *Chapman and Beardsley*, 1989]. Beyond the shelf break, water masses and the warm-core ring reflect the Gulf Stream and or the Sargasso Sea.

[24] This approach to water mass classification has five basic steps: i) project predictors measured for each water parcel into standardized predictor space; ii) use a suite of clustering algorithms to detect clusters in multidimensional predictor space data which are analogous to water types; iii) use the *FOM* statistic to determine a reasonable range of how many water types exist; iv) map water types into geographic space and determine the most frequent boundaries between water masses; v) evaluate the difference between water types in predictor space as a measure of the difference or gradient between defined water masses. What this analysis provides are means that validate and add mathematical rigor to intuition about the water masses present in this study. The remaining portion of the paper will discuss the factors that must be considered when interpreting the water mass boundaries and gradients calculated by this analysis.

4.1. Standardization of Variables

[25] The three predictors were standardized to their respective means and standard deviations so that the variation observed in each predictor gets equal weight in this analysis. Without this standardization, temperature alone would have dominated the results because it is numerically on the order of 10^1 units while R_{rs} is numerically on the order of 10^{-3} units. However, in doing this the water mass boundaries and gradients can only be compared within the group that was standardized, in this case the 4 days presented here. This is an important consideration in interpreting the results of the algorithm. Large gradients and frequent boundaries surround the obvious optical load seen on 31 July and 2 August in $R_{rs(555)}$ because it represented a large change in optical predictors compared to all of the data in this analysis. While this bloom is a distinct feature for those 4 days, if the question were whether this feature is distinct compared to a seasonal

trend or annual trend, the 4-day data set would need to be standardized to the mean and variability of the season or year. The same principle applies for a comparison of these images to images taken in another location or in reference to larger regions. For example, for a comparison of the gradients in this image to dynamics in another coastal region, the mean and variability of both regions would have to be included for proper comparison. While this nearshore optical load may be very distinct in the context of these 4 days in this particular region, its distinctness seasonally or annually in this region may be different depending on the inherent mean and variability of the system.

[26] While standardization of the variables is important for interpretation of the results, it is also important to note that standardization of the data does not guarantee that the data are normally distributed. Examining Figure 1, one can see that the temperature and the $R_{rs(490)}$ are fairly normally distributed (i.e., the area with high values is approximately equal to area with low values, and the majority of the area is covered with midrange values). In the case of $R_{rs(555)}$, most of the area is covered with low values and only a small area nearshore is covered with high values. This means that the data have a slightly skewed distribution. Therefore, in predictor space, despite standardization of this particular data set, there is a larger range of data along the $R_{rs(555)}$ axis, thus waters with high $R_{rs(555)}$ values in this study are more easily discriminated in parameter space.

4.2. Predictor Space Structure, Frequent Boundaries, and Gradients

[27] The suite of clustering algorithms was used to detect the inherent structure or water types in predictor space represented in four composite data sets of SST, $R_{rs(490)}$ and $R_{rs(555)}$. For increased computational speed clusters were defined from 2 to 30, however it is mathematically possible to define *n* water types where each observation is unique. This is the challenge associated with categorizing a known continuum of data; it is difficult to determine how different an observation of SST, $R_{rs(490)}$ and $R_{rs(555)}$ should be before it is considered a separate water type. The *FOM* statistic provides a mean to address this problem. While not providing a definitive answer as to how many water types existed in this data set, it did reduce the range of possibilities from *n* water types to 2-*TAF* water types. The geographic distribution of water types detected by the clustering algorithms between 2 and *TAF* is illustrated in Figure 5. The significance of high-frequency boundaries in this figure is that they represent consistent divisions of water types detected by more than one clustering algorithm at more than one cluster number (k). In essence, the four clustering algorithms vote by majority of what data in predictor space determine the dominant water types. However, because this technique uses the similarity of solutions by different clustering algorithms to determine dominate boundaries of water masses, the dissimilar solutions, which represent the low-frequency boundaries in Figure 5, represent somewhat of a ‘‘forced’’ result due to low signal.

[28] While boundaries may be consistently reflecting recognizable water types in predictor space by the clustering algorithms, the frequency of boundaries is not necessarily related to the gradients separating the water masses. For example, on 14 July several high-frequency boundaries were present indicating that the clustering algorithms were finding consistent structure in predictor space indicating discrete water types. However, gradient analysis of that same day indicates that while distinct water types are present in the data set, the differences between them are relatively small. This is different than 31 July and 2 August when the most frequent boundary also reflected a strong gradient. Therefore, for complete interpretation of water mass characteristics, both frequency of boundaries and gradient strengths must be considered. For example, a high-frequency water mass boundary is calculated on 21 July at approximately 40°N, 73°W which is the same frequency as the water mass boundary calculated for the nearshore ‘‘hammer-head’’ shape on 31 July and 2 August (Figure 5), however the gradient calculated for this boundary (Figure 7) is weak compared to gradients found on 31 July and 2 August. This result indicates that the boundary on 21 July is separating distinct water types in predictor space, however the water masses represented by these water types are not nearly as different as the water masses separated along the ‘‘hammer-head’’ shape on 31 July and 2 August. A distinct frontal region cam be inferred on 21 July in this area, but the water masses that are meeting at this front are not as different as ones encountered elsewhere in this analysis.

4.3. Current Structure, Boundaries, and Gradients

[29] The measured current structure associated with the boundaries and gradients indicate that physical features in the current field such as convergent zones and horizontal sheers are generally associated with water mass boundaries. This suggests that the physical processes are driving the propagation of the frontal region, as opposed to spurious changes in the optics due to changes in biomass or SST due to solar sea surface warming. Furthermore, it has been shown that optical properties are highly related to spatial physical dynamics in this region [*Oliver et al.*, 2004; *Schofield et al.*, 2002]. However, it should be noted that the current resolution (6 km) averaged over three hours might be too coarse to resolve all pertinent currents that are shaping these complex fronts. The drifter simulation (Figure 9) from 31 July to 2 August shows that the positions of water mass boundaries in this study are also related largely to local advective processes. The predicted boundary location of the 31 July boundary on 2 August using assimilated CODAR fields is very similar to the observed boundary position on 2 August. The current magnitudes and directions are sufficient to explain not only the general location of the water mass boundary, but also how some of the specific features form such as the protrusion of the northern horn of the ‘‘hammer-head’’ shape. Discrepancies between the predicted location of the boundary on 2 August and the actual location of the boundary on 2 August may be due to local vertical sheers. The CODAR system measures the current velocity of approximately the top meter of the water column, while the boundary location is responding to the integrated depth averaged current. Despite this, these results suggest that at least over the short term in this coastal region, water masses can be identified and tracked.

[30] Presently, ocean observatories are being developed world wide and the water mass analysis presented here is an efficient way to assimilate observational data and objectively

describe prevalent water types in a system as well as describe the strengths of the boundaries between them. From an operational standpoint, this can be a powerful tool in determining sampling strategies for specific experiments. Depending on the variables of interest, this type of analysis can be used when the position of water masses defined by other predictors or many predictors are more cryptic and nonintuitive. With the development of remote sensing optical inversion algorithms that detect functional groups of phytoplankton, this analysis can be used to detect clusters of communities and identify ecotones. These ecotone regions often have higher primary and secondary production leading to higher fish production [*Pingree et al.*, 1974]. In addition, this type of analysis can be used in understanding the biogeochemistry of a particular water mass and be able to track it in the context of an observing system.

5. Conclusion

[31] The goal of this study was to determine if specific water types could be identified and mapped as distinct water masses in a coastal region using satellite data from AVHRR and SeaWiFS, and whether the measured surface current field supported the boundaries and gradients in these maps. Because of the episodic and dynamic nature of coastal regions, optical discriminators were added to a water mass analysis to resolve water types that would not be resolved only by a single suite of parameters. To do this tools were adapted from the field of bioinformatics to constrain the number of water types in this study. On the basis of the boundary and gradient analysis, water types based on temperature and remote sensing reflectance could be mapped and that the relative differences between them could be estimated. Furthermore, the boundaries and gradients were generally colocated with features in the current field. Simulated drifter experiments show that the location of these boundaries is largely a result of local advective processes. This suggests that the predictors used in this experiment change slow enough to act as effective tracers of water masses over short timescales.

[32] **Acknowledgments.** The support of the National Ocean Partnership Program (N00014-97-1-1019), the Office of Naval Research COMOP and HyCODE programs (N00014-97-0767, N00014-99-0196) and the NSF EEGLE program (OCE-9727341, OCE-9727342) are gratefully acknowledged. We especially thank Mike Crowley and Jen Bosch 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://www.marine.rutgers.edu/cool) with good beer. Finally, the continuing support from the great state of New Jersey is acknowledged.

References

Auer, S. J. (1987), Five-year climatological survey of the Gulf Stream system and its associated rings, *J. Geophys. Res.*, *92*, 11,709–11,726.
Barrick, D. E., M. W. Evans, and B. L. Weber (1977), Ocean surface currents mapped by radar, *Science*, *198*, 138–144.
Beardsley, R. C., and C. D. Winant (1979), On the mean circulation in the Mid-Atlantic Bight, *J. Phys. Oceanogr.*, *9*, 612–619.
Behrenfeld, M. J., and P. G. Falkowski (1997), Photosynthetic rates derived from satellite-based chlorophyll concentration, *Limnol. Oceanogr.*, *42*, 1–20.
Broecker, W. S., and T. H. Peng (1982), *Tracers in the Sea*, 690 pp., Lamont-Doherty Geol. Obs., New York.
Broecker, W., and T. Takahashi (1985), Sources and flow patterns of deep-ocean waters deduced from potential temperature, salinity, and initial phosphate concentration, *J. Geophys. Res.*, *90*, 6925–6939.
Chapman, D. C., and R. C. Beardsley (1989), On the origin of shelf water in the Middle Atlantic Bight, *J. Phys. Oceanogr.*, *19*, 384–391.
Chung, F. L., and T. Lee (1994), Fuzzy competitive learning, *Neural Networks*, *7*, 539–551.
Claustre, H., P. Kerherve, J. C. Marty, L. Prieur, C. Videau, and J. H. Hecq (1994), Phytoplankton dynamics associated with a geostrophic front-Ecological and biogeochemical implications, *J. Mar. Res.*, *54*, 711–742.
Glenn, S. M., G. Z. Forristall, P. Cornillon, and G. Milkowski (1990), Observations of Gulf Stream Ring 83-E and their interpretation using feature models, *J. Geophys. Res.*, *95*, 13,043–13,063.
Hartigan, J. A., and M. A. Wong (1979), Algorithm AS 136: A K-means clustering algorithm, *Appl. Stat.*, *28*, 100–108.
Helland-Hansen, B. (1916), Nogen hydrografiske metoder form, *Skand. Naturf. Mote.*, 357–359.
Hey, J. (2001), The mind of the species problem, *Trends Ecol. Evol.*, *16*, 326–329.
Højerslev, N. K., N. Holt, and T. Aarup (1996), Optical measurements in the North Sea-Baltic transition zone. I. On the origin of the deep water in the Kattegat, *Cont. Shelf Res.*, *16*, 1329–1342.
Jerlov, N. J. (1968), *Optical Oceanography*, Elsevier Oceanogr., New York.
Karabashev, G., M. Evdoshenko, and S. Sheberstov (2002), Penetration of coastal waters into the eastern Mediterranean Sea using the SeaWiFS data, *Oceanol. Acta*, *25*, 31–38.
Kohut, J. T., and S. M. Glenn (2003), Improving HF radar surface current measurements with measured antenna beam patterns, *J. Atmos. Oceanic Technol.*, *20*, 1303–1316.
Martin-Traykovski, L. V., and H. M. Sosik (2003), Feature-based classification of optical water types in the Northwest Atlantic based on satellite ocean color data, *J. Geophys. Res.*, *108*(C5), 3150, doi:10.1029/2001JC001172.
Meeks, J. C., J. Elhai, T. Thiel, M. Potts, F. Larimer, J. Lamerdin, P. Predki, and R. Atlas (2001), An overview of the genome of *Nostoc punctiforme*, a multicellular symbiotic cyanobacterium, *Photosyn. Res.*, *70*, 85–106.
Morel, A., and L. Prieur (1977), Analysis of variations in ocean color, *Limnol. Oceanogr.*, *22*, 709–722.
Noor, M. A. F. (2002), Is the biological species concept showing its age?, *Trends Ecol. Evol.*, *17*, 153–154.
Oliver, M. J., S. Glenn, J. T. Kohut, A. J. Irwin, O. M. Schofield, M. A. Moline, and W. P. Bissett (2004), Bioinformatic approaches for objective detection of water masses on continental shelves, *J. Geophys. Res.*, *109*(1), C07S04, doi:10.1029/2003JC002072, in press.
Pingree, R. D., G. R. Forster, and G. K. Morrison (1974), Turbulent convergent tidal fronts, *J. Mar. Biol. Assoc. U.K.*, *54*, 469–479.
Quackenbush, J. (2001), Computational analysis of microarray data, *Natl. Rev. Gen.*, *2*, 418–427.
Sathyendranath, S., T. Platt, E. P. W. Horne, W. G. Harrison, O. Ulloa, R. Outerbridge, and N. Hoepffner (1991), Estimation of new production in the ocean by compound remote sensing, *Nature*, *353*, 129–133.
Schofield, O., T. Bergmann, W. P. Bissett, F. Grassle, D. Haidvogel, J. Kohut, M. Moline, and S. Glenn (2002), Linking regional coastal observatories to provide the foundation for a national ocean observation network, *J. Oceanic Eng.*, *27*, 146–154.
Tomczak, M. (1999), Some historical, theoretical and applied aspects of quantitative water mass analysis, *J. Mar. Res.*, *57*, 275–303.
Upstill-Goddard, R. C., A. J. Watson, J. Wood, and M. I. Liddicoat (1991), Sulphur hexafluoride and helium-3 as sea-water tracers: Deployment techniques and continuous underway analysis for sulphur hexafluoride, *Anal. Chim. Acta*, *249*, 555–562.
Ward, J. H. (1963), Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, *58*, 236–244.
Warren, B. A. (1983), Why is no deep water formed in the North Pacific?, *J. Mar. Res.*, *41*, 327–347.
Wu, C. I. (2001), The genic view of the process of speciation, *J. Evol. Biol.*, *14*, 851–865.
Yankovski, A. E., and R. W. Garvine (1998), Subinertial dynamics on the inner New Jersey shelf during the upwelling season, *J. Phys. Oceanogr.*, *28*, 2444–2458.
Yeung, K. Y., D. R. Haynor, and W. L. Ruzzo (2001), Validating clustering for gene expression data, *Bioinformatics*, *17*, 309–318.

W. P. Bissett, Florida Environmental Research Institute, 4807 Bayshore Blvd., Suite 101, Tampa, FL 33611, USA.

S. Glenn, A. J. Irwin, J. T. Kohut, M. J. Oliver, and O. M. Schofield, Coastal Ocean Observation Lab, Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Rd., New Brunswick, NJ 08901, USA. (oliver@imcs.rutgers.edu)

M. A. Moline, Biological Sciences, California Polytechnic State University, San Luis Obispo, CA 93405, USA.