AN ASSESSMENT OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING APPLICATIONS TOWARD PREDICTING GULF OF MEXICO COASTAL HYPOXIA

A Thesis

by

SAMANTHA CLAIRE LONGRIDGE

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Steven F. DiMarco
Committee Members,	Jason B. Sylvan
	Antonietta S. Quigg
Head of Department,	Shari Yvon-Lewis

August 2022

Major Subject: Oceanography

Copyright 2022 Samantha C. Longridge

ABSTRACT

Observations of dissolved oxygen, salinity, temperature, and six different nutrient concentrations of the waters on the TXLA Shelf in the months of March – September in 2003 – 2014 were used in unsupervised and supervised machine learning techniques to identify driving processes of hypoxia and examine the performance of classification algorithms on predicting hypoxia on the TXLA Shelf. Unsupervised machine learning techniques, principal component analysis, and K-means clustering, successfully identified variability patterns that were associated with previously known drivers and processes of hypoxia in the region such as vertical stratification of the water column and the Mississippi River plume. The performance of eight classification algorithms (i.e., logistic regression, LDA, QDA, naïve bayes, KNN, SVM, decision tree, and random forest) on predicting hypoxia with the observations on TXLA Shelf were compared. Results showed that naïve bayes performed best on classifying hypoxia with high recall and low false positive rates. Balancing the class distribution in the training set of each algorithm significantly increased performance, indicating that classifier performance was strongly dependent on input training data. This study establishes that straightforward machine learning techniques can aid in identification of known main drivers of hypoxia and their characteristics and that those characteristics can be used to predict hypoxia on the TXLA Shelf. These techniques have the potential to evaluate hypoxia presence or absence in hydrographic data where DO is missing and can be a powerful tool used in water quality and resource management in the region. While the approaches presented in this study were specifically for the TXLA Shelf, the methodology is applicable to other coastal systems and locations with similar datasets.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. DiMarco and my committee members, Dr. Sylvan, and Dr. Quigg for their guidance and support throughout the course of this research. Thanks also to the friends and department faculty and staff, especially Dr. Wiederwohl and Dr. Yvon-Lewis who have mentored me and helped shape me throughout my time at Texas A&M.

Finally, thanks to my parents and husband for encouraging me to persevere.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor DiMarco and Professor Sylvan of the Department of Oceanography and Professor Quigg of the Department of Marine Biology.

The data analyzed for Section 3 and Section 4 were provided by Professor DiMarco.

All work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a Graduate Teaching Assistantship from Texas A&M University and two Department of Oceanography Scholarships, the Lechner Graduate Scholarship and the Luis and Elizabeth Scherck Scholarship. This work was also made possible in part by the NSF S-STEM under award #1355807 and the NOAA Ocean Acidification project (NA19OAR0170354). This work was also made possible by the Society for Underwater Technology in the U.S. (SUT-US) Scholarship, the NOAA funded Mechanisms Controlling Hypoxia project (NA09N0S4780208, NA06N0S4780198, NA03N0S4780039), and the NSF Research Experience for Undergraduates under award # 1849932.

NOMENCLATURE

TXLA	Texas-Louisiana
GOM	Gulf of Mexico
MARS	Mississippi-Atchafalaya River System
RC02	Rowe & Chapman (2002)
POM	Particulate Organic Matter
NOAA	National Oceanic and Atmospheric Administration
DO	Dissolved Oxygen
ML	Machine Learning
MCH	Mechanisms Controlling Hypoxia
R/V	Research Vessel
CTD	Conductivity/Temperature/Depth
REU	Research Experience for Undergraduates
NSF	National Science Foundation
GSW	Gibbs Seawater Package
PCA	Principal Component Analysis
PC	Principal component
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
KNN	K-Nearest Neighbor
SVM	Support Vector Machine

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

1. INTRODUCTION

1.1. Hypoxia on the Texas-Louisiana Shelf

The Texas-Louisiana (TXLA) Shelf (Figure 1.1) in the Gulf of Mexico (GOM) is home to the largest area of seasonally oxygen depleted coastal bottom waters offshore the United States. Hypoxia is defined by low oxygen levels or dissolved oxygen concentrations that are less than 2 mg/L or 1.4 mL/L and occurs in the bottom waters of the TXLA Shelf in the summer. Hypoxia on the TXLA Shelf was first reported in 1972 and has been systematically mapped since 1985 (Rabalais et al., 1994, 1996, 1999; Turner et al., 2005; Rabalais et al., 2007). Low oxygen bottom water events have been occurring for at least the last 1000 years but their size rapidly grew across the shelf in the 1950s and continues to expand (Osterman et al., 2008; Osterman et al., 2009). Average hypoxic area coverage is approximately 17,000 square kilometers of the eastern TXLA Shelf (Dale et al., 2010; Rabalais et al., 2007). Coastal systems and estuaries worldwide also exhibit similar increases in hypoxia, primarily due to increases in anthropogenic nutrient inputs (Rabalais et al., 2010; Diaz 2001; Diaz and Rosenburg, 2008). Hypoxia can harm marine organisms and can lead to negative consequences for the economy of the northwestern GOM, which is reliant on recreational and commercial fisheries resources. Additionally, future climate change and sea-level rise have the potential to alter the intensity of hypoxia globally (Rabalais et al., 2010).

Hypoxia in this region is brought on by combined effects from the nutrient and riverine load from the Mississippi-Atchafalaya River System (MARS) and associated estuaries and strong vertical stratification (i.e., stratification sufficient to prevent reoxygenation of the lower layer) that is enhanced by freshwater discharge and upwelling favorable wind stress (Forrest et al., 2010, Feng et al. 2012; Bianchi et al., 2010). This area receives ~50% of the riverine input from the

Mississippi River with added input from the Atchafalaya River that is about 30% of the total Mississippi River discharge (Etter et al., 2004, Bianchi et al., 2010). The MARS drains major agricultural and industrial regions that cover ~40% of the contiguous United States and parts of Canada (Milliman & Meade, 1983). This runoff is variable but characteristically peaks in the spring and has minimums in the late summer (Figure 1.2) (Nowlin et al., 1998). The freshwater discharge contains suspended sediments, dissolved and particulate matter, and nutrients. It flows from the Mississippi Birdfoot Delta and travels south and westward into Mexican waters.

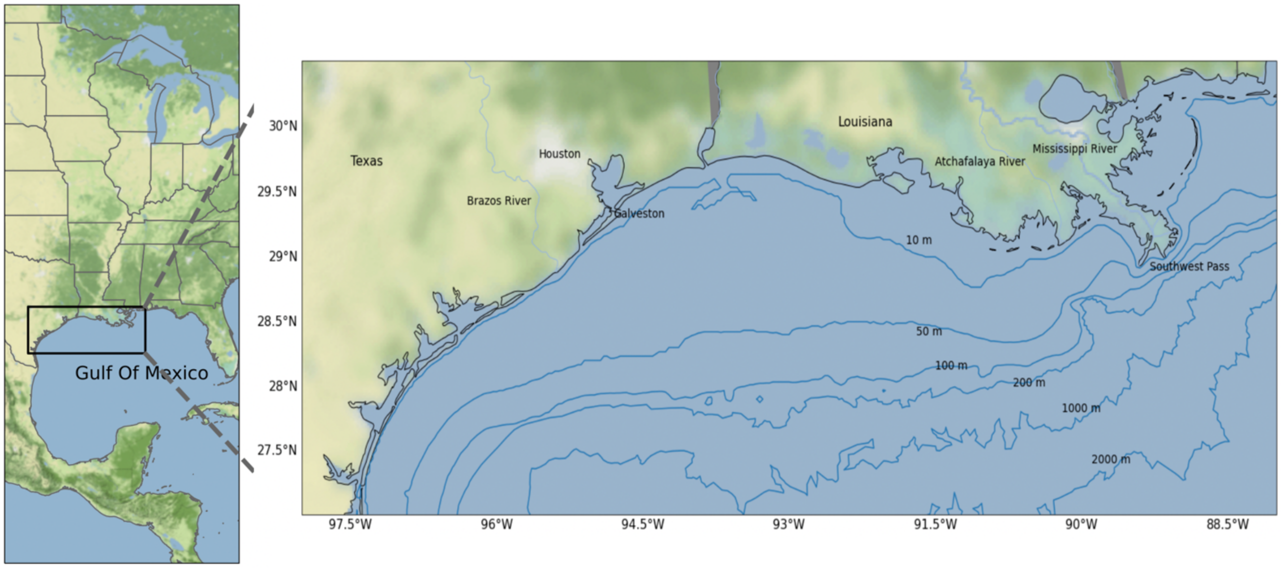


Figure1.1 The Texas-Louisiana Shelf in the northwestern Gulf of Mexico.

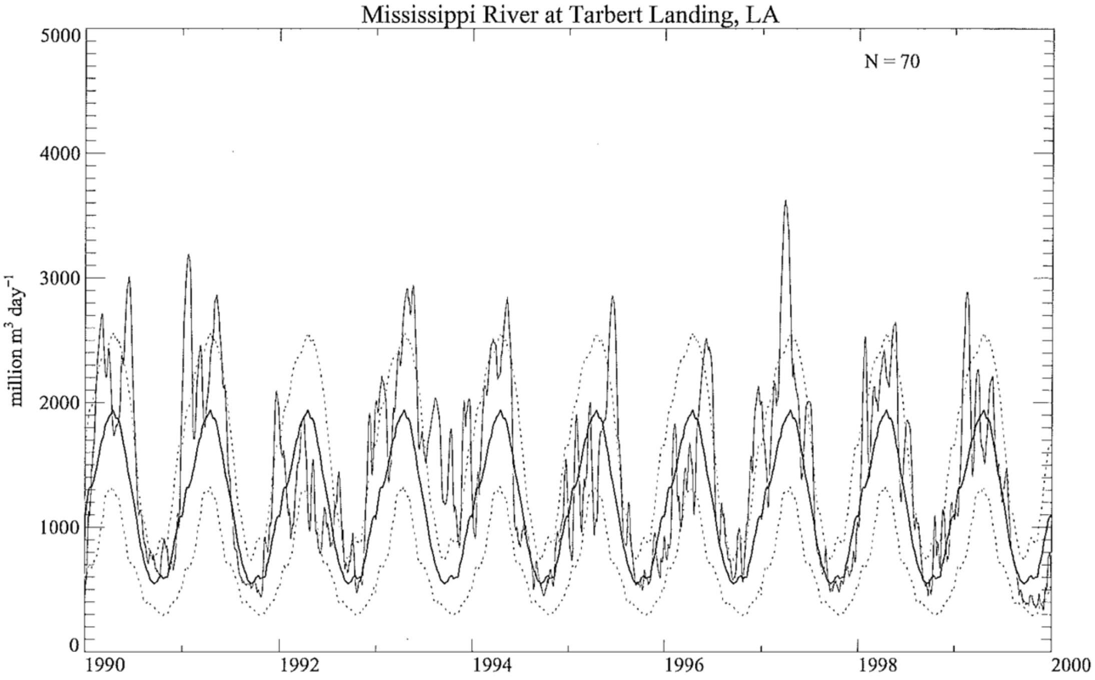


Figure 1.2 Ten-year record (1990-99) of flow in the Mississippi River at Tarbert Landing, LA. The thin solid line represents the actual flow, whereas the thicker solid line and the dotted lines represent the 70-yr mean flow and the standard deviation about the mean (Rowe & Chapman, 2002).

The effect of the riverine input is twofold in that the nutrient load fuels the biological mechanism of hypoxia and the freshwater causes stratification from the density difference between the riverine freshwater and saltier GOM ocean water (Bianchi et al. 2010). The nutrient load fuels organic carbon production such as phytoplankton blooms and fecal pellets. Heterotrophic respiration by microorganisms reduces oxygen since microorganisms use it as the most energetically favorable electron acceptor to remineralize organic carbon biomass until is it entirely depleted, after which nitrate is used. If the freshwater lens from riverine water creates a density difference between the mixed layer depth and water below the pycnocline, then oxygen removed by remineralization cannot be replaced by diffusion from the atmosphere and oxygen is drawn down below the pycnocline. If there is a partial lack of oxygen and dissolved oxygen concentrations are below 1.4 mL/L, it is hypoxic, and if oxygen is consumed to the point of depletion (i.e., dissolved oxygen concentrations = 0 mL/L), it is referred to as anoxia.

Additionally, in the summer, the region's mean alongshore winds are reversed and flow upcoast creating conditions favorable for upwelling and advection of water from Mexico onto the TXLA Shelf (Cochrane & Kelley, 1986; Feng et al., 2014; Nowlin et al., 2005; Wiseman et al., 1997). The winds are also weaker in the summer and fronts occur less frequently (Nowlin et al., 2001). The reversed winds and currents keep the riverine input from flowing 'downcoast' meaning from Louisiana toward the west (Cochrane & Kelley, 1986; Cho et al., 1998). This allows additional time for respiration processes to continue in both the sediment and water column (Rowe & Chapman, 2002). The destruction of the halocline stratification occurs with the onset of fall storms that reoxygenate the bottom waters to eliminate hypoxic conditions (Wiseman et al., 1997; Bianchi et al. 2010; Rabalais et al. 2007).

1.2. Additional Influences of Hypoxia

Since this region is a river-dominated ocean margin, the importance of the nutrient load brought on by the river system has been stressed as a primary cause for hypoxia. It has been shown that river discharge and nutrient load are positively correlated with hypoxic area in the GOM; Rabalais et al. (2002b) concluded that the main cause of hypoxia is marine phytoplankton production driven by nitrogen loading from the Mississippi River. As a consequence of this, it was suggested that policies and regulations should be defined and implemented to reduce nitrogen load to mitigate the size of the associated hypoxic zone (Rabalais et al., 2002a). However, other studies have concluded that while, nutrients, specifically nitrogen, sourced from the MARS have taken the full blame for hypoxia occurrences, it was not actually the case. Rowe & Chapman (2002) suggested other sources of material besides the nitrogen load from the

MARS could cause oxygen consuming processes that contribute to hypoxia in the region. Additionally, nitrate has typically been regarded as the most important nutrient that controls phytoplankton growth in the northwestern GOM, but it has been shown that phosphate can act as the limiting nutrient (Ammerman & Sylvan, 2004; Sylvan et al., 2006; Sylvan et al., 2007). In the western region of the TXLA shelf, DiMarco et al., (2012) found that discharge from the Brazos River created favorable conditions for hypoxia to form locally and along the Texas coast in the summer of 2007. These findings and additional studies indicate that the MARS is not the sole source of nutrients that can cause hypoxia and that the process of limiting nutrients and riverine input utilization is more complex than initially thought (Dortch & Whitledge, 1992; Sharples et al., 2017).

The relative importance of the river system, nutrients, and stratification (physical water column structure) on hypoxia formation has been shown to be significant in hypoxia formation and strength (Forrest et al., 2011; Zhang et al., 2015). There are other physical processes at play that contribute to the complexity and variation of hypoxia in the GOM. The amount of river discharge, wind mixing, and advection modulate the strength of hypoxia (Wiseman et al., 1997; Feng et al., 2012; Forrest et al., 2011). DiMarco et al. (2010) showed that topography and currents play a crucial role in managing intensity, distribution, and timing of hypoxia. It has also been shown that a minimum vertical stratification (i.e., stability frequency of ~ 40 cycles/hr) is needed for bottom water oxygen levels to be classified as hypoxic (Belabbassi, 2006; Bianchi et al., 2010; Dale et al., 2010), and that water below the pycnocline needs to be quiescent with a slow replacement time (i.e. the time needed to replace the amount of water in the Gulf of Mexico) for hypoxia to occur (Rowe & Chapman, 2002).

The northwestern GOM river dominated margin (Bianchi et al., 2010) is a highly dynamic system. Rowe & Chapman in 2002 (referred to as RCO2) proposed that there were three distinctive zones of hypoxia that have their own controlling mechanisms depending on different physical and biochemical processes (Figure 1.3).

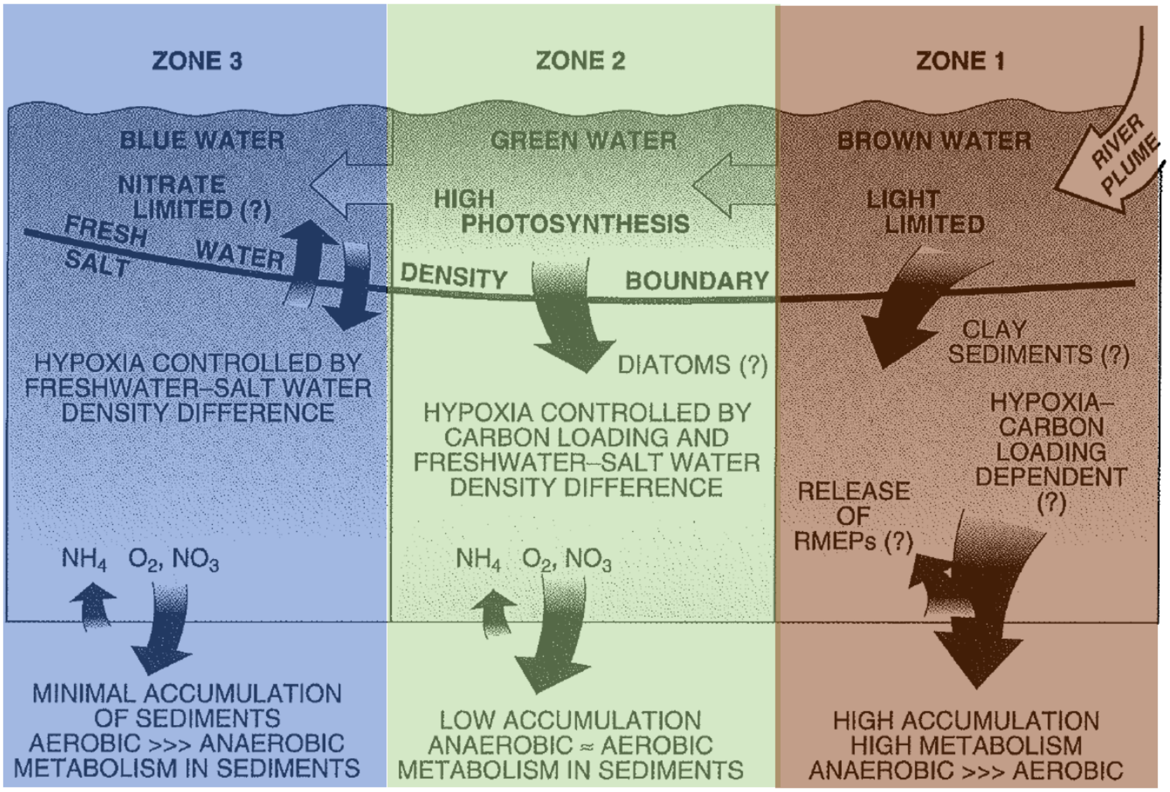

ZONE 3
ZONE 2
ZONE 1
BLUE WATER
GREEN WATER
BROWN WATER
RIVER PLUME
NITRATE LIMITED (?)
HIGH PHOTOSYNTHESIS
LIGHT LIMITED
FRESH WATER
SALT
DENSITY BOUNDARY
CLAY SEDIMENTS (?)
DIATOMS (?)
HYPOXIA CONTROLLED BY FRESHWATER–SALT WATER DENSITY DIFFERENCE
HYPOXIA CONTROLLED BY CARBON LOADING AND FRESHWATER–SALT WATER DENSITY DIFFERENCE
HYPOXIA– CARBON LOADING DEPENDENT (?)
RELEASE OF RMEPs (?)
NH_4 O_2, NO_3
NH_4 O_2, NO_3
MINIMAL ACCUMULATION OF SEDIMENTS AEROBIC >>> ANAEROBIC METABOLISM IN SEDIMENTS
LOW ACCUMULATION ANAEROBIC ≈ AEROBIC METABOLISM IN SEDIMENTS
HIGH ACCUMULATION HIGH METABOLISM ANAEROBIC >>> AEROBIC

Figure 1.3 Colored RC02 zone model paradigm describing physical and biochemical processes controlling hypoxia on the Texas-Louisiana shelf (Rowe & Chapman, 2002).

Zone 1 is located nearest to the river mouth, where sediment load is high and light availability is reduced, thereby, reducing phytoplankton production. The particulate organic matter (POM), including particulate organic carbon and clay, is deposited and aerobic metabolism draws down oxygen, resulting in hypoxia and the production of ammonium, sulfides, reduced iron, and manganese species. In zone 2, which is adjacent to zone 1 (Lahiry, 2007), there is less turbidity since the POM has flocculated out and therefore the water column is not light limited. Nutrient concentrations are still high compared to the open ocean, but not as high as they are closest to the river mouth. Phytoplankton can bloom, and the general paradigm of hypoxia can occur. As one moves further from the mouth of the rivers offshore into zone 3, the nutrient load in the surface decreases because it was used up in zone 2 and this region is controlled mostly through the strong stratification caused by the freshwater and saltwater density difference. Therefore, it is most affected by river input and changes in freshwater volume. Primary production in zone 3 is driven by regenerated nitrogen. This zone can be described as low salinity water stripped of its nutrients in the euphotic zone and can extend from the river mouth westward and beyond the Texas coast.

The defined edges of these zones are not spatially static but can change with river discharge and biological processes (Rowe & Chapman, 2002; Dale et al., 2010). Lahiry (2007) defined the edges of these RC02 zones in three oceanographic cruises conducted in April, June, and August of 2004 using salinity changes and hydrographic data. Kim et al. (2020b) further explored the RC02 zone model by using a box model and Kim et al (2020a) utilizes the zone hypothesis to explore the nutrient and salinity relationships to define the biological productivity associated with the Mississippi and Atchafalaya Rivers.

1.3. Modeling Hypoxia – the need for Machine Learning

Hypoxia on the shelf is a matter of concern and with the potential consequences for water quality and fisheries that are significant in the region; therefore, accurate diagnostics and prediction of the distribution and intensity of hypoxia is needed. Both the timing of hypoxia development and the spatial extent can vary dramatically. When considering the large suite of processes that influence oceanic water quality, including weather and climate variability, this poses challenges for water quality and marine resource management (Boesch et al., 2001). Long- and short-term prediction is necessary due to the large spatiotemporal variability.

Many efforts have been made to model hypoxia and its influences in this region. Hindcasts and predictions of hypoxia and hypoxic extent from Scavia et al. (2003) and Turner et al (2005) rely on statistical relationships of nutrient load, freshwater flux, and enhanced primary productivity. Hetland & DiMarco (2008) discuss their shortcomings and point out that their predictive abilities are not greater than the direct correlation between Mississippi River Discharge and areal hypoxia extent. Additionally, Hetland & DiMarco (2008) used a realistic hydrodynamic model of the TXLA Shelf and found that their results suggest that different biological processes are responsible for hypoxia in different physical regions proposed by RC02. Also, Forrest et al. (2010) used multivariate statistical regression to model hypoxic areal extent. Statistical models have been developed for forecasting hypoxic volume, such as the NOAA Ensemble Hypoxia Forecast which is a compilation of different models to forecast the size of the hypoxic zone. Although these forecasts are regularly published online every June and receive media attention, they are limited in that they predict overall hypoxic zone size but provide no information about the evolution of the spatial distribution of hypoxia over time.

Even with advanced numerical methods, accurate prediction of coastal dissolved oxygen (DO) variability is challenging and computationally expensive. Recently, machine learning (ML) algorithms have become more efficient, computationally cheaper, and faster than traditional hydrodynamic and biogeochemical numerical models. These algorithms can also represent spatial distribution and variability. Machine learning techniques can easily replicate nonlinear phenomena from a sufficiently large dataset (i.e., dependent upon the complexity of the system the user is aiming to predict and the learning algorithm used) with an appropriate number of features and target quantities. These ML capabilities have been demonstrated for predicting hypoxia and other biogeochemical and water quality parameters in other estuaries, coastal systems and freshwater bodies in several studies, including Corpus Christi Bay, TX, San Luis Obispo Bay, CA, Chesapeake Bay on the eastern US Coast, Lake Huron, Hong Kong marine waters, and Tampa Bay, FL. (Chang et al., 2013; Coopersmith et al., 2011; Deng et al., 2021; Guo et al., 2021; Ross & Stock et al., 2019; Valera et al., 2020; Yu et al., 2021; Yu et al., 2020). Coopersmith et al. (2011) produced reasonable estimates of hypoxic probability in Corpus Christi Bay using previous day DO, salinity, water temperature, wind speed and wind direction as input forcings in a k-nearest neighbor model. Valera et al. (2020) showed that random forest regression and support vector regression models accurately reproduced nearshore and offshore DO with high accuracy. Even with testing and application of ML techniques to predict variables in range of coastal environments, it is still relatively limited particularly for DO concentrations on the continental shelf. Specifically, ML methods for identifying the dominant process that control hypoxia has not yet been applied in the northwestern GOM. With ML methods' flexibility to represent nonlinearity, spatial variability, and seasonal changes in DO, ML methods provide a unique opportunity for new insights into hypoxia in the GOM.

1.4. Research Goals and Questions

Considering ML's capabilities, RC02's proposal of different controlling regimes, and the lack of ML methods applied to hypoxia on the TXLA Shelf, here we explore if machine learning methods can aid in identification of the dominant hypoxia controlling processes and classify hypoxia from hydrographic characteristics.

The two fundamental guiding questions for my research are:

1) Can machine learning methods be used to identify the dominant oceanographic processes that drive hypoxia?
2) Can hydrographic characteristics (i.e., salinity, temperature, nutrients) be used as a predictive tool for identifying hypoxia on the TXLA Shelf?

The following hypotheses stem from the project's research questions.

H1) Machine learning methods can identify and quantify differences in hydrographic data that are related to oceanographic processes across the Texas-Louisiana Shelf.

H2) Hydrographic characteristics that are associated with hypoxia can be used as a predictive tool for determining the presence of hypoxia.

H1 will be addressed by applying two unsupervised ML methods to identify the drivers of hypoxia on the TXLA Shelf using a large dataset made up of hydrographic data from 31 research cruises. **H2** is addressed by exploring the capabilities of supervised ML classification techniques to predict hypoxia. Here we apply eight supervised ML classification techniques and examine their performance in classifying hypoxia on the TXLA Shelf. This thesis analyzes the capability of ML techniques to identify driving processes of hypoxia and examines the performance of classification techniques on predicting hypoxia on the TXLA Shelf.

1.5. Outline

This thesis is organized into five sections. Section 1 provides the introduction and background information of hypoxia in the northwestern GOM, and the motivation and need for ML. It also includes the guiding research questions, hypotheses, and plan for testing the hypotheses. Section 2 describes the data and methods used to resolve the research questions. Section 3 describes and discusses the results from the Research Question 1 (above). Section 4 describes and discusses results from the Research Question 2 (above). Section 5 provides an integrative discussion, conclusions, and recommendations for future work.

2. DATA AND METHODS

2.1 Mechanisms Controlling Hypoxia Data (MCH)

Hydrographic data including salinity, temperature (ºC), dissolved oxygen (mL/L), nitrate (µmol/L), nitrite (µmol/L), total nitrogen (µmol/L), silicate (µmol/L), phosphate (µmol/L), urea (µmol/L), and ammonia (µmol/L) were obtained from a previously conducted study, the Mechanisms Controlling Hypoxia on the Louisiana Shelf (MCH) project (DiMarco & Zimmerle, 2017; DiMarco, 2012; DiMarco, 2021). The MCH Project was funded by NOAA from 2003-2016 and consisted of an integrated observational and numerical modeling approach to better understand the Interactions of the physical, biological, and geochemical processes and their variability across the entire TXLA Shelf. This information contributes to a comprehensive description of the mechanisms that control hypoxia in the northern Gulf of Mexico. Environmental and oceanographic observations were recorded on 31 process-oriented research cruises and resulted in more than 120 towed transects, ~5,000 CTD casts, ~30,000 water samples, and more than 50,000 km of ship flow-through system data.

The 31 cruises occurred in the months of March – September in the years of 2003 – 2014 aboard four different research vessels including R/V Gyre, R/V Pelican, R/V Blazing Seven, and the R/V Manta. Physical and biogeochemical data were collected by bottle and CTD. Sampling stations associated with these cruises span from the Mississippi River Delta to as far southwest as Corpus Christi Bay (Figure 2.1).

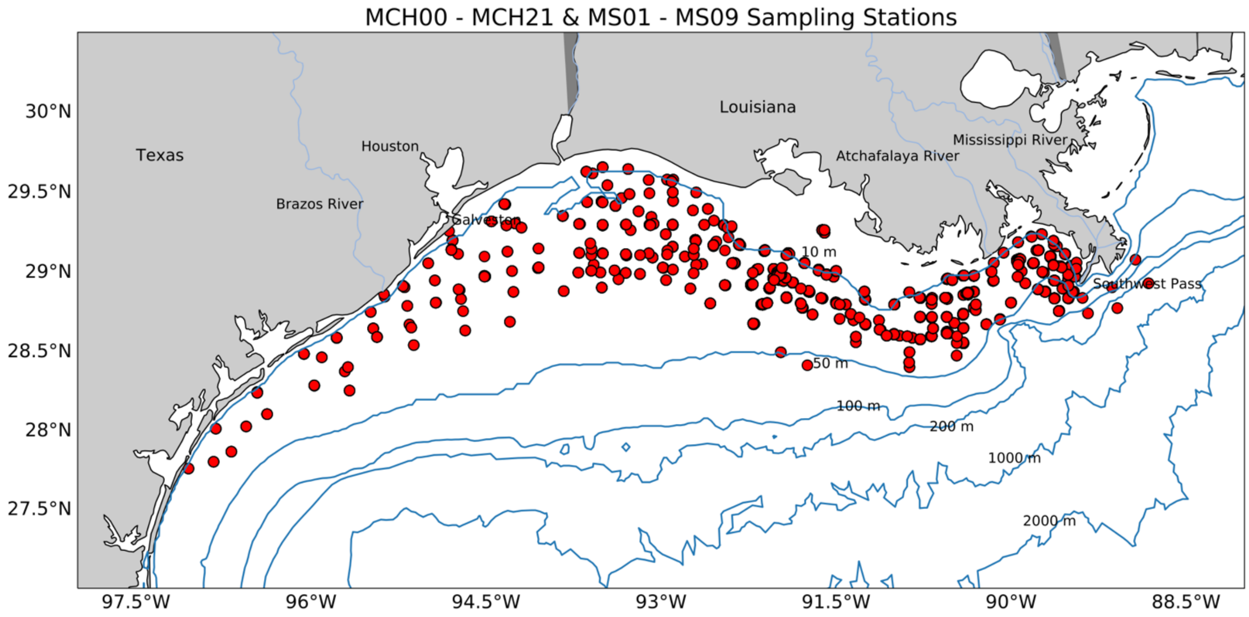


Figure 2.1 MCH sampling station locations.

2.1.1 MCH Preprocessing

Data from the 31 MCH cruises were compiled into one dataset and rows that were missing information were not included. To maximize the amount of data included, missing values in salinity and in DO were substituted with values from the CTD and DO probe on the niskin rosette respectively. Eight samples collected on the MCH21 cruise at the mouth of the Atchafalaya River were excluded from the analysis and were deemed as outliers since they were not representative of the population of the coastal area. Additionally, 99 ammonia and 49 urea values were greater than 10 µmol/L and were excluded since these were suspect for the GOM. After applying these restrictions, 4,886 sampling points remained in the data set. A table of sampling stations with their associated dates and number of stations included in this project can be seen in Table 2.1.

Hypoxia Cruise	Date	Number of Stations	Number of Stations Included in this Project
MCH00	14-16 September 2003	36	0
MCH01	2-7 April 2004	59	53
MCH02	26 June – 1 July 2004	62	59
MCH03	21-25 April 2004	56	55
MCH04	23-27 March 2005	50	50
MCH05	20-26 May 2005	82	81
MCH06	8-14 July 2005	74	37
MCH07	18-23 August 2005	90	88
MCH08	23-29 March 2007	41	39
MCH09	17-20 July 2007	23	23
MCH10	6-10 September 2007	30	29
MCH11	16-18 April 2008	21	21
MCH12	17-20 July 2008	29	23
MCH13	7-10 April 2009	35	11
MCH14	28-31 July 2009	9	9
MCH15	6-10 April 2010	4	4
MCH16	15-23 August 2010	8	8
MCH17	25-30 April 2011	5	5
MCH18	20-23 June 2011	7	7
MCH19	16-20 August 2011	6	6
MCH20	25-30 April 2012	5	5
MCH21	7-12 August 2012	5	4
MS01	14-17 June 2012	19	18
MS02	3-7 August 2010	65	27
MS03	24-28 June 2011	53	33
MS04	8-14 August 2011	56	52
MS05	10-16 June 2012	56	55
MS06	15-21 August 2012	64	18
MS07	20-25 June 2013	75	3
MS08	4-10 August 2013	69	14
MS09	17-23 June 2014	45	4

Table 2.1 Mechanisms Controlling Hypoxia Project research cruises during years 2003 – 2014.

2.2 Ancillary Data

Additional data were used in the second part of this study to test machine learning models (Section 4). These supporting data includes hydrographic data from three research cruises conducted as a part of Texas A&M's Oceanography Department Research Experience for Undergraduates (REU) *Observing the Ocean* in 2016 2017, and 2018. It also includes hydrographic data collected from NSF's Ocean Acidification research cruises (XR01 and XR02).

2.2.1 Research Experience for Undergraduates 2016, 2017, & 2018 (REU)

Hydrographic data were collected by the Texas A&M's Oceanography Research Experience for Undergraduates (REU) *Observing the Ocean* in June in the years 2016, 2017 and 2018. These data are publicly available at: https://geo.gcoos.org/tamu_reu_observing_the_ocean/about/ . Sampling stations of these cruises can be seen in Figure 2.2. Plots of salinity, temperature, dissolved oxygen, nitrate, nitrite, silicate, phosphate, urea, and ammonia can be seen in the Appendix.

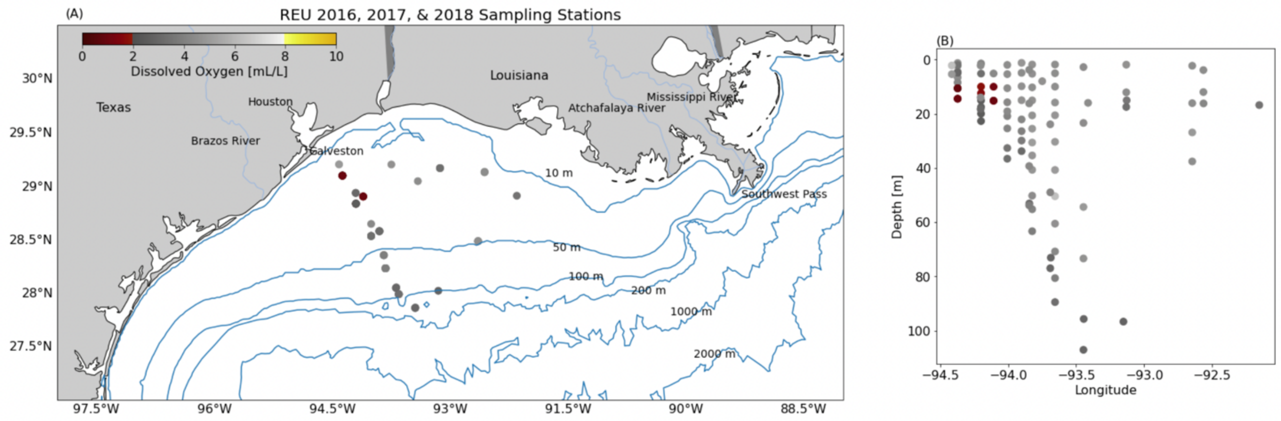


Figure 2.2 (A) REU 2016, 2017, & 2018 sampling stations colored with DO [mL/L]. (B) Vertical view of sampling stations.

2.2.2 Ocean Acidification Data (XR01 & XR02)

Hydrographic data were obtained from a previously conducted study, the Ocean Acidification Project (Award #: NA19OAR0170354). This project included two cruises XR01 and XR02 aboard the R/V Pelican. XR01 occurred in April 2021 and XR02 occurred in August

2021. Physical and biogeochemical data were collected by bottle and CTD. XR01 and XR02 temperature was derived using the Gibbs Seawater (GSW) package in MATLAB (McDougal & Barker (2011). Sampling stations of these cruises can be seen in Figure 2.3 and Figure 2.4. Plots of salinity, temperature, dissolved oxygen, nitrate, nitrite, total nitrogen, silicate, phosphate, urea, and ammonia can be seen in the Appendix.

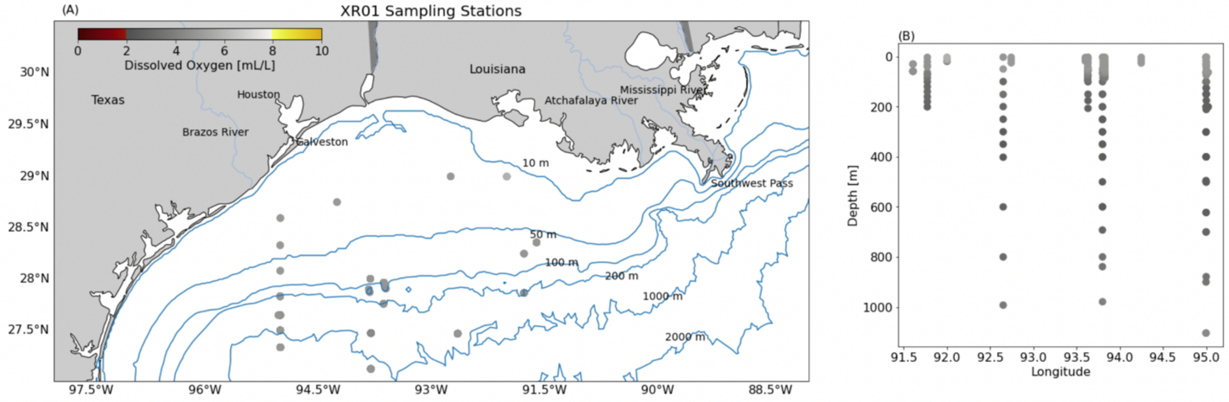


Figure 2.3 (A) XR01 sampling stations colored with DO [mL/L]. (B) Vertical view of sampling stations.

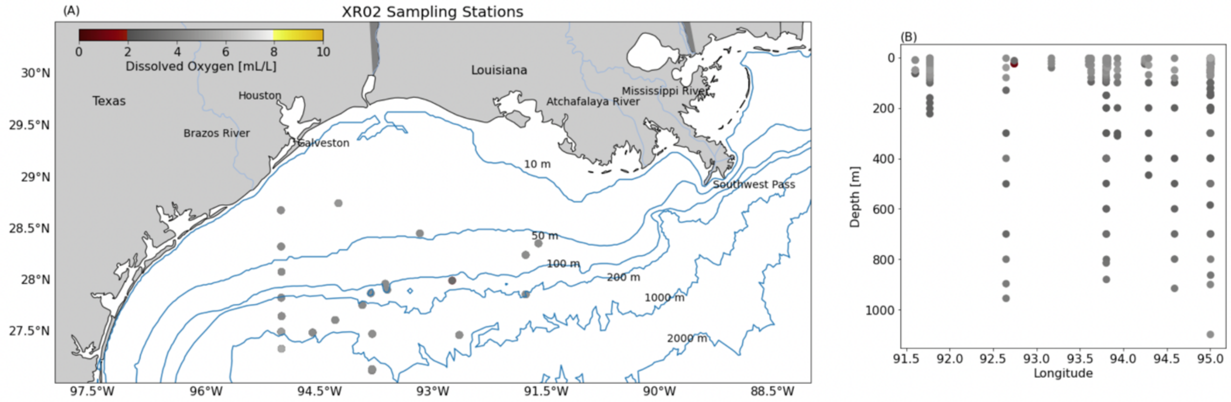


Figure 2.4 (A) XR02 sampling stations colored with DO [mL/L]. (B) Vertical view of sampling stations.

2.2.3 Ancillary Data Preprocessing

The same assessment applied to that the MCH data were applied to the ancillary data. To maximize data included, missing salinity and DO values were substituted with CTD salinity and DO probe values from units on the deployed niskin rosette. Rows of the datasets with missing variables were eliminated. Additionally, it was checked that ammonia and urea values did not exceed more than 10 μmol/L. After applying these restrictions and combining the REU, XR01 and XR02 data, there was 473 sampling points in the dataset.

2.3. Unsupervised Machine Learning Methods

Prior to applying unsupervised ML methods, the MCH dataset was standardized (i.e., all variables demeaned and normalized by their standard deviation) in MATLAB. Standardizing variables is typical for many unsupervised ML methods because they are affected by the scales of different variables. Principal Component Analysis (PCA) and K-Means Clustering are the techniques selected for the analyses. We decided to focus on these two techniques, as opposed to other methods for their straightforward approach and potential wider applicability which includes characterization of data, investigation of variable relationships, and anomaly detection.

Application of PCA is quite common in ocean sciences and is typically favored for use on larger datasets due to its dimensionality reduction abilities. PCA increases interpretability and minimizes information loss by creating uncorrelated variables (principal components / modes). The PCA in this study decomposes the data matrix Y into the following form:

$$Y_j^{(k)} = \sum_{i=1}^{N} Z_i^{(k)} a_{ij}$$

where Z are the scores (data in the principal component space), a are the eigenvectors. j represents columns, k represents rows, N is the number of variables and modes. Eigenvectors (a) and eigenvalues (λ) are obtained by solving the covariance matrix C.

$$Ca_i = \lambda_i a_i$$

The data used in Section 3 of this study has dimension 4886 x 10 (salinity, temperature, dissolved oxygen, nitrate, nitrite, total nitrogen, ammonia, urea, phosphate, and silicate), representing the spatially varying values at 1,570 sampling stations over 11 years (2004 – 2014). Principal component analysis was implemented in MATLAB. Major variance of the dataset is well represented in the major principal components. Our analysis below shows that the first 4 PCA modes account for ~75% of the total variance. Following North's Rule of Thumb (North et al., 1982), we determined to truncate the principal modes after mode 4 (Figure 2.5) and only focus on the first 4 modes in the results and discussion.

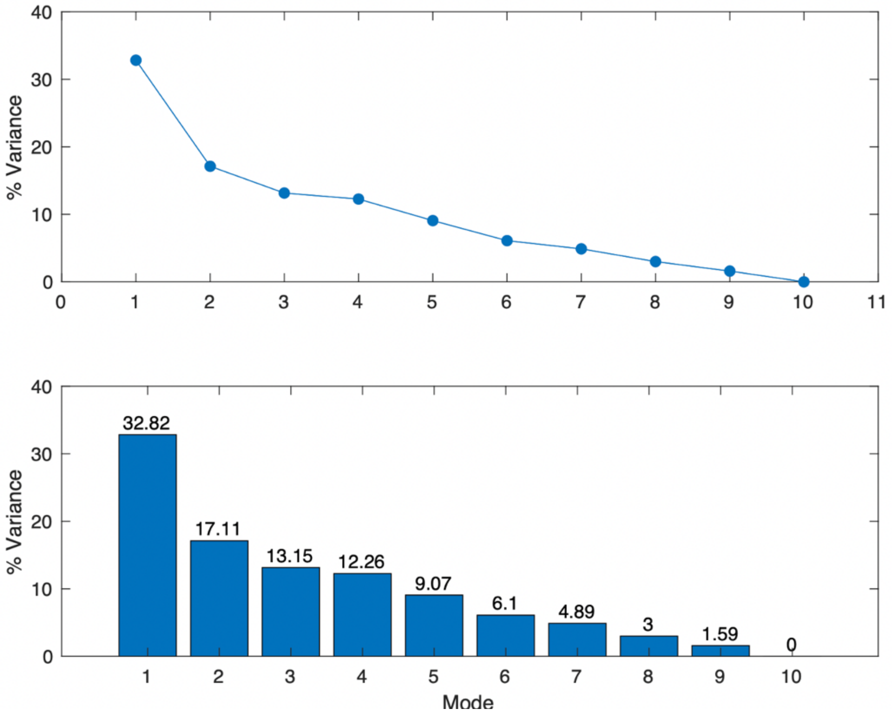


Figure 2.5 Principal Component Modes % of Variance.

K-means clustering was selected as the other form of unsupervised ML since it differs from PCA in that it looks for homogeneous subgroups among observations. K-means partitions observations into groups in which each observation belongs to the cluster with the nearest mean. It also minimizes the variance within each cluster. This technique will allow us to categorize observations based on their hydrographic characteristics to describe different zones of drivers of hypoxia in the GOM. The same standardized data matrix (4886 x 10) used in PCA was also used in K-means in MATLAB. To determine the optimal number of clusters to use, silhouette scoring (i.e. a measure of how close each point is in one cluster to points in the neighboring cluster used to determine optimal number of clusters to use) was run iteratively in MATLAB.

2.4 Supervised Machine Learning Methods

To address the second hypothesis of this study, we focused on classification algorithms to test if hydrographic characteristics can be used as a predictive tool for determining if hypoxia will be present. We chose to assess the performance of eight classification algorithms: logistic regression (Cox, 1958), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) (Fisher,1936; Rao, 1948), naïve bayes (Bayes, 1764), k-nearest neighbors (KNN) (Fix & Hodges, 1989; Cover & Hart, 1967), support vector machine (SVM) (Boser et al 1992), decision tree (Morgan & Sonquist, (1963), and random forest (Breiman ,1996; Breiman, 2001). All algorithms were implemented in R. These algorithms were chosen due to their straightforward interpretability of results and wide range of applicability.

Before supervised ML techniques could be applied to the MCH dataset, it needed to be prepared for classification. Total nitrogen was excluded because it is positively correlated with individual concentrations of nitrate (r = 0.74) and nitrite (r = 0.66). DO was converted to type: 'hypoxic' or 'not hypoxic' in the MCH dataset (4886 x 9).

2.4.1 K-Fold Cross Validation

To assess classification performance, we utilized k-fold cross validation, a resampling method that uses different portions of the data (folds) to test and train a model on different iterations. In a prediction problem such as classification, the model is given a dataset of known data on which training is run (training set) and a data set of unknown data is used to test the model (test set). K-fold cross validation randomly partitions the dataset into k equal sized subsamples. Of the k subsamples, a single subsample is retained for testing and the remaining subsamples are used as training data. The cross-validation process is then repeated k- times, meaning each subsample is used exactly once as the test set.

Cross-validation is important when estimating how accurate a predictive model will perform in practice. The advantage of using cross-validation is that it allows for a more robust measure of the model's ability to predict classification on new data that was not used in training. Cross-validation allows the user to identify problems with the model such as overfitting (i.e., the model fits too closely to the training set and accounts for too much noise, decreasing its ability to predict new data) or selection bias where the model is not trained with properly randomized data. Without cross-validation, measures of accuracy and prediction may not be representative of the model's true capability to predict.

In this case, we determined that five folds was appropriate for the MCH dataset. A larger number of folds means that each model is trained on a larger training set and tested on a smaller test fold. Given that the MCH data set is a 4,886 x 9 matrix, five folds means that each model is trained on 75% of the data (3,664 data points) and tested with 25% (1,221 data points). Five-fold cross validation was applied to each classification algorithm. The average of all five fold's recall, precision, specificity, accuracy, and F1-score were calculated.

Recall (Sensitivity) is the ratio of correctly labeled hypoxic samples to all real hypoxic samples, it answers the following, i.e., Of the true hypoxic samples, how many were labeled as hypoxic? **Precision** is the ratio of correctly labeled hypoxic samples to all labeled hypoxic samples, it answers: how many of those that were labeled as hypoxic are actually hypoxic? **Specificity** is the ratio of correctly labeled not hypoxic samples to all real not hypoxic samples. Specificity addresses the question: of all the real not hypoxic samples, how many of those were correctly predicted? **Accuracy** is the ratio of correctly labeled samples to the whole pool of samples, it answers: how many samples were correctly labeled (hypoxic or not hypoxic) out of all the samples? Lastly, **F1-score** considers both precision and recall and is the harmonic mean of the precision and recall. F1 score is best is there is a balance between precision and recall in the system. If the cost of false positives and false negatives are both undesirable, F1 score is a great measure.

3. UNSUPERVISED MACHINE LEARNING ASSESSMENT: APPLICATION TO HYPOXIA DRIVING MECHANISMS IN THE NORTHWESTERN GULF OF MEXICO

The goal of this part of the study is to identify patterns of variability in hypoxia. Unsupervised ML is used for analysis between input variables (e.g. cluster analysis, anomaly detection, dimension-reduction, and multivariate analysis) to discover hidden patterns or model the distribution of data; therefore, unsupervised ML techniques should identify patterns of variability in hypoxia on the TXLA Shelf. We hypothesize (recall **H1** in Chapter 1) that the patterns will be associated with coastal and oceanographic processes that correlate with previously found significant drivers of hypoxic variability.

3.1 Principal Component Analysis (PCA)

Through PCA and interpretation with our current knowledge of DO and hypoxia in the northwestern GOM, we can understand the dominant processes controlling the variations. The spatial and temporal characteristics and the possible controlling mechanisms of the first four PCA modes will be described in the following paragraphs.

The first PCA mode is the dominant mode, accounting for 32.8% of the total variance. DO and temperature are positively correlated and are negatively correlated with salinity and the 7 nutrients (nitrate, nitrite, total nitrogen (nitrate + nitrite), phosphate, silicate, ammonia, urea). Relationships between variables in mode 1 can be seen in the coefficient plot (Figure 3.1). Elements falling within the same quadrant indicate positive correlation; in opposite quadrants indicates negative correlation. By multiplying the sign of the coefficients to the sign of the principal component (PC) mode the variables can be generally characterized as high or low. This can give insight into which mechanisms could be at play in the mode.

Positive values of mode 1 had relatively (in relation to the negative scores) low DO (mean = 2.5 mL/L), high salinity (mean = 33.36), low temperature (mean = 25.2 ºC), and high nutrients. Negative values of mode 1 had relatively high DO (mean = 4.6 mL/L), low salinity (mean = 31.3) high temperature (mean = 26.8 ºC), and low nutrients. The map of the first mode is characterized with a majority of positive values to the east of 94 ºW extending to the Mississippi River Delta and the negative values spanning across the shelf from the Mississippi River Delta to as far west as 96 ºW (Figure 3.2a). Positive values are located deeper in the water column than negative values (Figure 3.2b).

The first mode has significant seasonality ($p << 0.001$) with two peaks in May and July (Figure 3.2d). This is most likely a combined effect of seasonal variation in temperature, salinity, and nutrient concentrations. Also, there are significant differences in depth groups (depth / 10) indicating a depth effect ($p <<0.001$). This depth effect could be related to water-column stratification or different water masses.

Given this information, the positive values of mode 1 can describe the water that is below the pycnocline that is isolated from the surface. Stratification inhibits ventilation to subpycnocline water with large organic carbon biomass. The large biomass can fuel microbial decay processes that remineralize nutrients and deplete oxygen. These processes can explain the hydrographic characteristics of the positive values in mode 1. Furthermore, the relationship between mode 1 and DO indicates that lower values of DO are associated with higher score values (Figure 3.2e).

The negative values of mode 1 can be described as low salinity water stripped of its nutrients and as a zone where hypoxia is mostly driven by strong stratification caused by the freshwater and saltwater density difference. This type of water and process controlling hypoxia was described as zone 3 (blue zone) by Rowe & Chapman (2002). Rowe & Chapman (2002) indicated that this zone has lower nutrient concentrations due to the nutrient depletion processes in the adjacent zone 2 and that it can extend from the river mouth to as far west as beyond the Texas coast. Here, the negative values do extend from the river mouth to as far west as 96 ºW. Additionally, the significant difference in depth groups could be indicative of the importance of stratification. To summarize, mode 1 represents subpycnocline water that may be hypoxic and low salinity water stripped of its nutrients as described by Rowe & Chapman (2002).

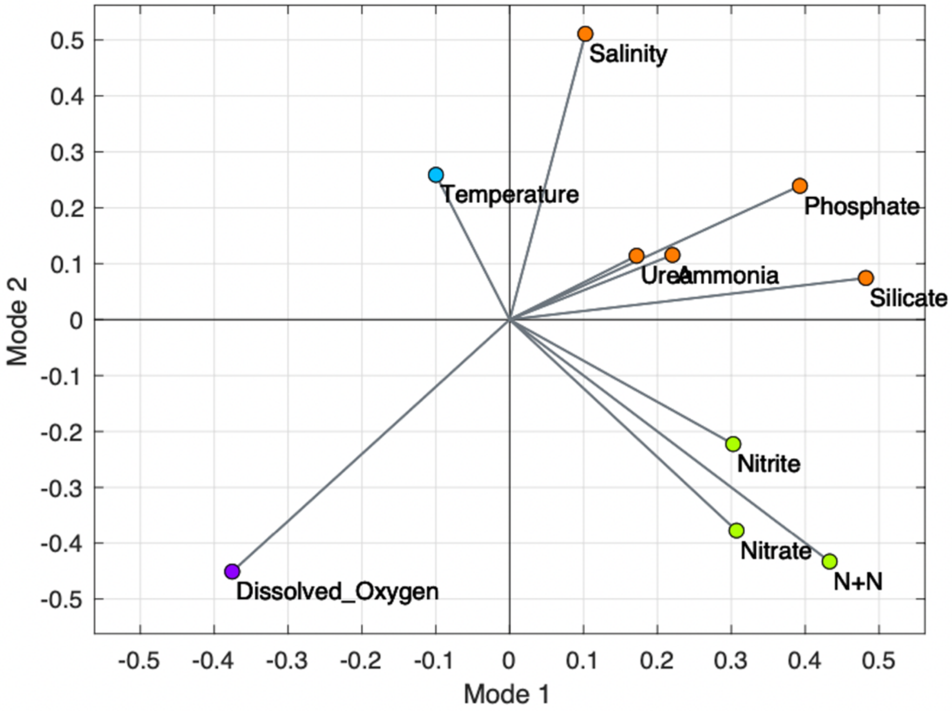


Figure 3.1 Principal Component Analysis Coefficients of Mode 1 and Mode 2. Variables in the same quadrant and colored the same are positively correlated and variables in opposite quadrants are negatively correlated.

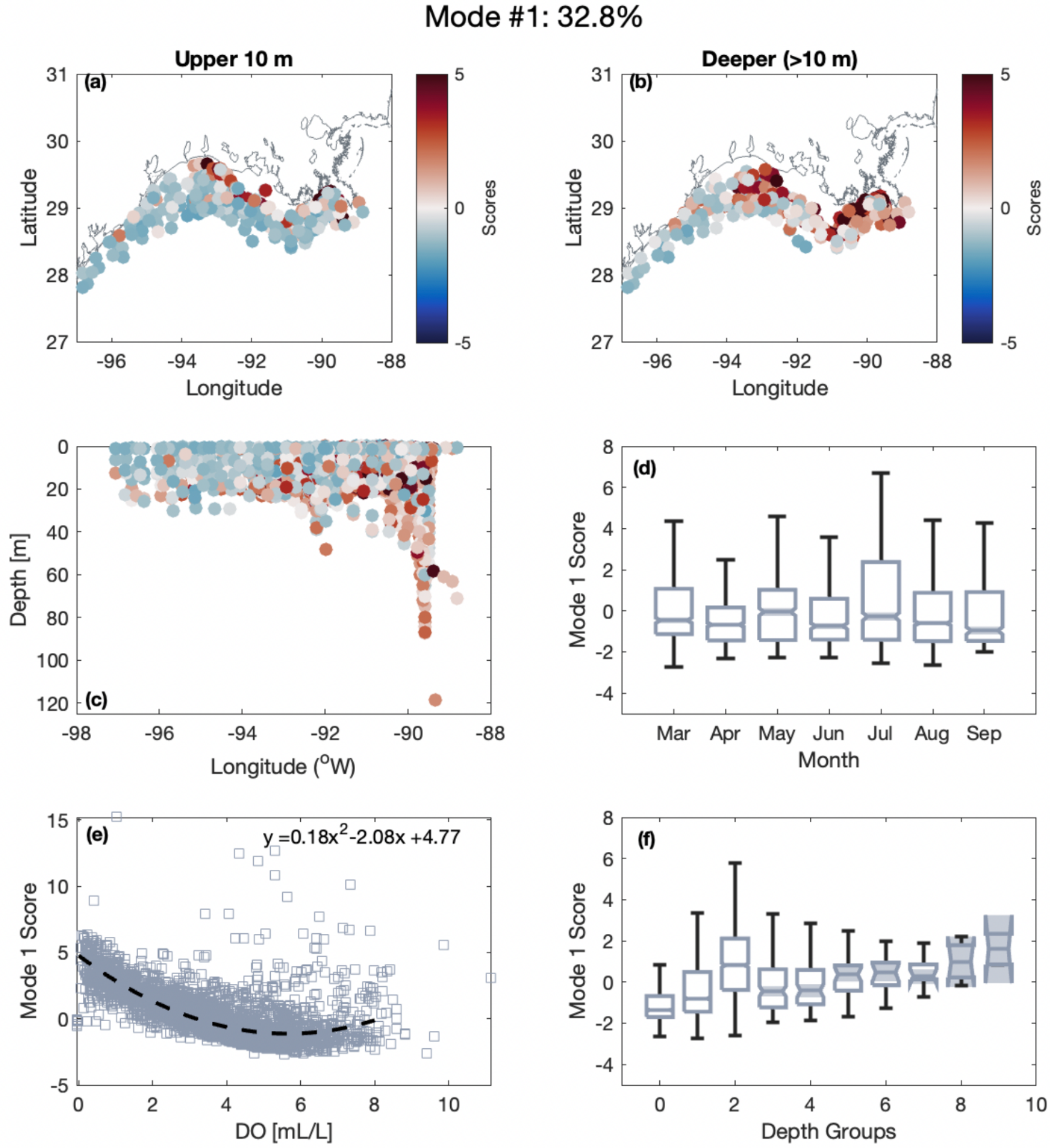


Figure 3.2 Principal Component Mode 1 (32.8% of total variance). (a) Spatial distribution of Mode 1 in the upper 10 m. (b) Spatial distribution of mode 1 deeper than 10 m. (c) Vertical distribution of mode 1. (d) Month means of mode 1. (e) Relationship between DO [mL/L] and mode 1. (f) Depth groups (sample depth / 10) in mode 1.

The second PCA mode accounts for 17.1% of the total variance in the data. Relationships between variables in mode 2 can be seen in the coefficient plot (Figure 3.1). DO, nitrate, nitrite, and total nitrate (nitrate + nitrite) are positively correlated with each other and are negatively correlated with temperature, salinity, urea, ammonia, silicate, and phosphate. Temperature, salinity, urea, ammonia, silicate, and phosphate are also positively correlated with each other. Positive values of mode 2 had relatively low DO (mean = 3.2 mL/L), higher salinity (mean = 33.8), high temperature (mean = 27.3 ºC), and low nutrients. Negative values of mode 2 had relatively high DO (mean = 5.7 mL/L), low salinity (mean = 29.6) low temperature (mean = 24.5 ºC), and high nutrients.

The map of the second PCA mode is characterized with opposite values between the area surrounding the mouth of the Mississippi River and the rest of the TXLA Shelf, suggesting this mode is controlled by the Mississippi-Atchafalaya River plume (Figure 3.3a&b). The vertical distribution in the water column also shows that the negative values are also mostly constrained to nearest the river mouth and near the surface (Figure 3.3c). This second mode also has significant seasonality ($p << 0.001$) with the lowest temporal value in March that increases to the highest temporal value in September (Figure 3.3d). There is also a significant difference between depth groups ($p << 0.001$).

Given the spatial distribution and hydrographic characteristics of the negative values, mode 2 can describe the Mississippi River plume (water that is high in DO, low in salinity, and high in nutrients). The MARS floods in the spring and with reversed winds and currents that happen in the summer, the freshwater can persist on the TXLA Shelf, allowing more time for oxygen-depleting processes to continue. The hydrographic characteristics of the positive values in mode 2 are characteristic of upwelled water, i.e., relatively low DO and nutrients that have been below the pycnocline, not ventilated with oxygen. Additionally, increasing values of mode 2 are associated with the increase in time, indicating seasonality (Figure 3.3d). This trend could be explained by the importance of the freshwater plume and it being prevented from moving downcoast from reversed winds and currents that also induce upwelling. It is also necessary to point out that the relationship between mode 2 and DO is a decreasing polynomial relationship. As mode 2 values become more negative, DO increases (Figure 3.3e). To sum up, mode 2 (17.1% of the variance in this dataset) is representative of the Mississippi River plume and is also representative of upwelled water induced by reversed winds and currents. The seasonality in mode 2 further supports these inferences since both the Mississippi River and upwelling on the TXLA Shelf have characteristic changes in time.

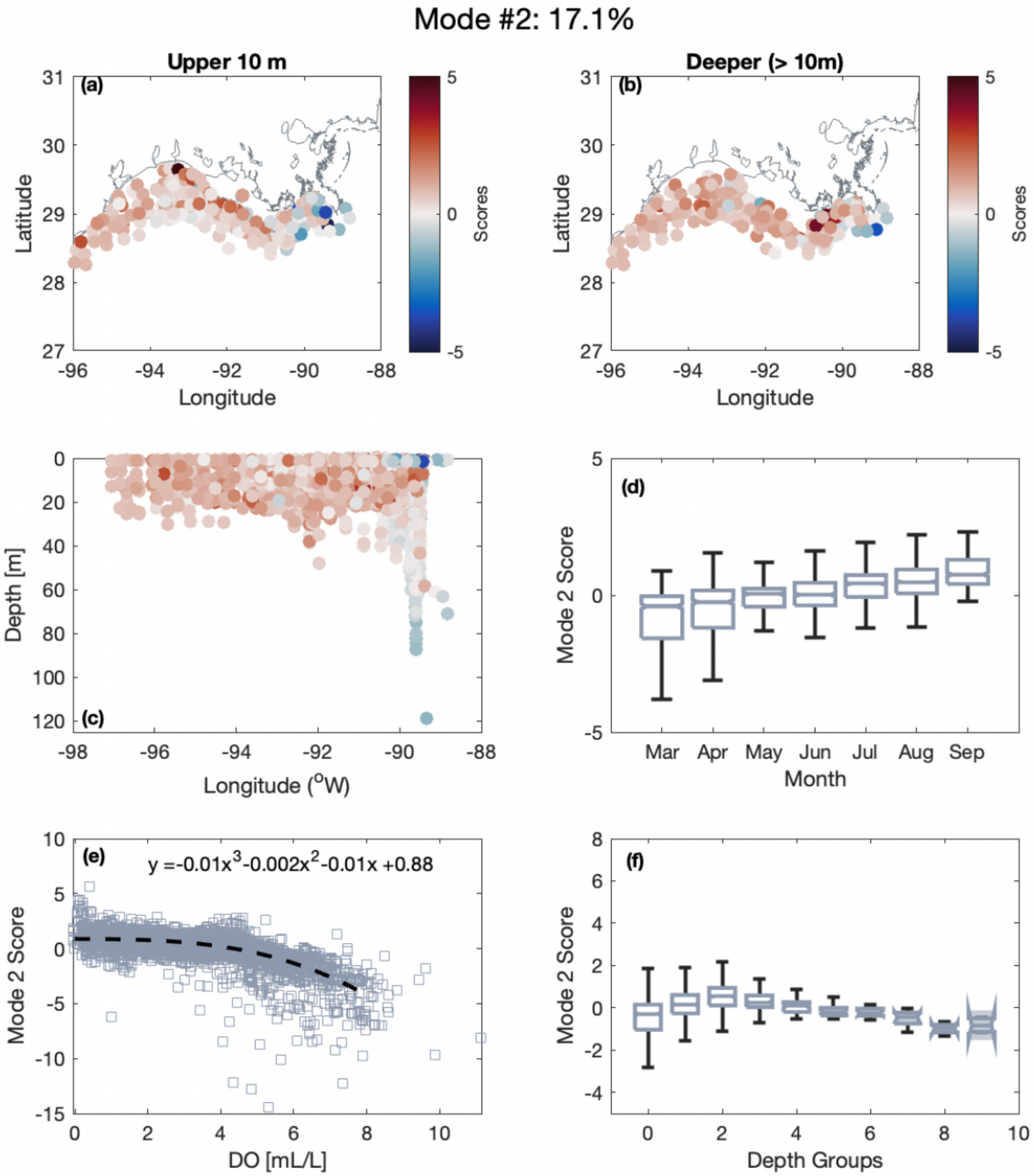


Figure 3.3 Principal Component mode 2 (17.1% of total variance). (a) Spatial distribution of mode 2 in the upper 10 m (b) Spatial distribution of mode 2 below 10 m (c) Vertical distribution of mode 2. (d) Month means of mode 2. (e) Relationship between DO [mL/L] and mode 2. (f) (Sample Depth / 10) means of mode 2.

Relationships between variables in the third and fourth PC modes can be seen in Figure 3.4. The third PC mode is characterized with opposite values between the surface and subsurface, suggesting this mode is controlled by differences between surface and subsurface water (Figure 3.5a&b). What is unique in mode 3 is that the relationship between the mode and DO has a slight positive relationship, meaning that higher values have higher DO (Figure 3.5e).

Additionally, there is significant seasonality ($p << 0.001$) in the temporal values that increases with time and has the highest value in July (Figure 3.5d). There is also a significant treatment effect of depth in this mode ($p << 0.001$). Positive values are characterized with high DO (mean = 4 mL/L), low salinity (mean = 30), high temperature (mean = 28.1 ºC) and variable means for nutrients. Negative values were characterized with low DO (mean = 3.6 mL/L), high salinity (mean= 34.2), low temperature (mean = 24.4 ºC), and variable nutrient concentration means. Nutrient concentration means are reported in the Appendix.

The hydrographic characteristics and the spatial distribution of opposite values of surface and subsurface can be associated with stratification, where positive values represent water above the pycnocline and negative values represent water below the pycnocline. This mode could also be representative of the oceanic shelf edge where negative values are below the shelf and positive values are located above the shelf edge. Although it is not clear which interpretation best describes this mode, mode 3 (13.2% of the total variance) does represent a difference in the vertical structure of the water column. To sum up, mode 3 can be associated with the vertical structure of the water column and describes a typical stratified continental shelf.

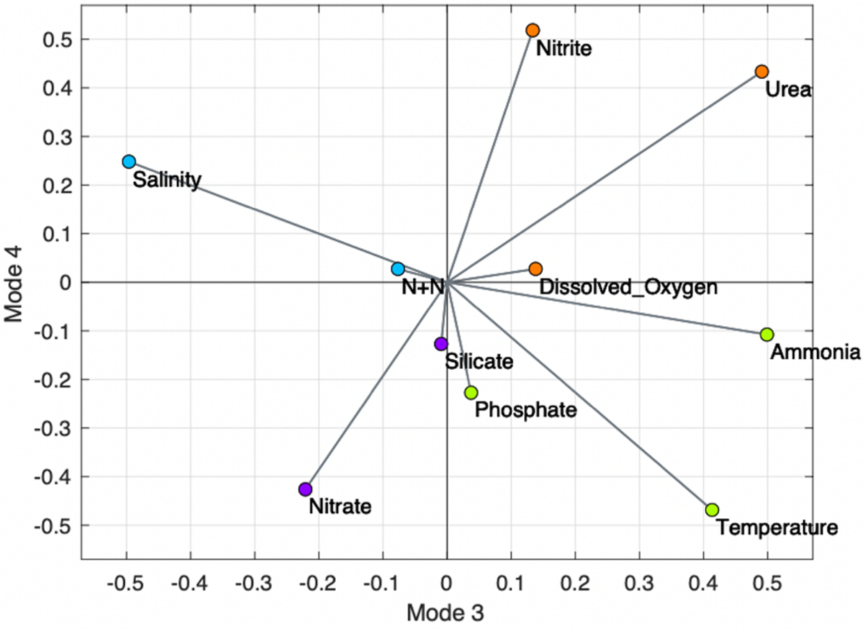


Figure 3.4 Principal Component Analysis Coefficients of Mode 3 and Mode 4. Variables in the same quadrant and colored the same are positively correlated and variables in opposite quadrants are negatively correlated.

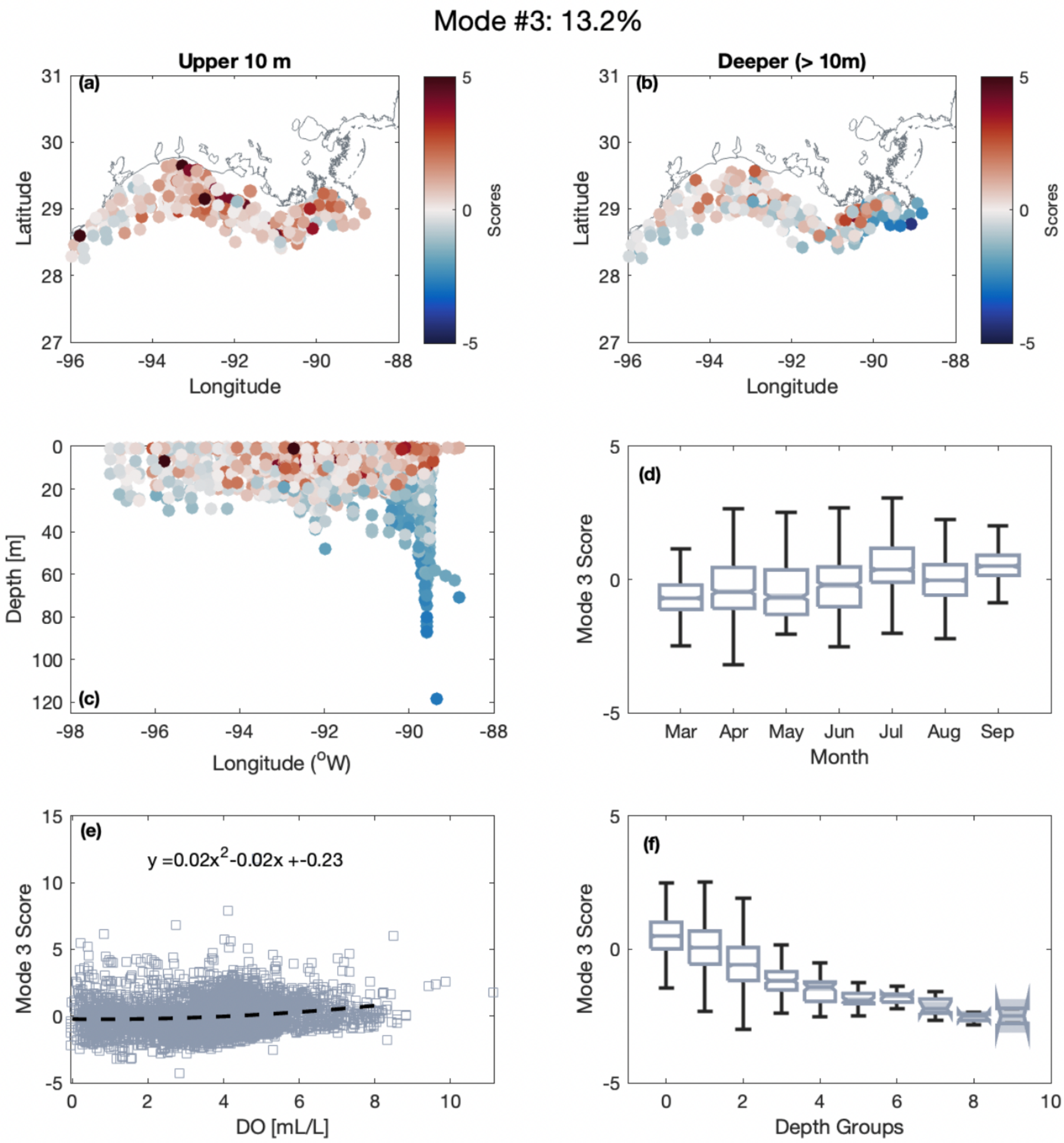


Figure 3.5 Principal Component Mode 3 (13.2% of total variance). (a) Spatial distribution of mode 3 in the upper 10 m. (b) Spatial distribution of mode 3 deeper than10 m. (c)Vertical distribution of mode 3. (d) Month means of mode 3. (e) Relationship between DO [mL/L] and mode 3. (f) Depth groups (sample depth/10) of mode 3.

The fourth PCA mode is characterized with mixed values covering the shelf (Figure 3.6a&b). There is no relationship between mode and DO (Figure 3.6e). There is significant seasonality ($p = 0$) and difference between depth groups ($p << 0.001$). What is unique about mode 4 is that there are two peaks in its temporal value in March and July (Figure 3.6d). It is not clear what mechanisms are responsible for the spatial and temporal pattern. Not all PCA modes can be easily explained with our current knowledge and sometimes multiple mechanisms are at play instead of one. Positive values of mode 4 had high DO (mean = 3.9 mL/L), high salinity (mean = 33.4), low temperature (mean = 23.5 ºC), with variable nutrient means. Negative values had low DO (mean = 3.7 mL/L), low salinity (mean = 31.1), high temperature (mean = 28.2 ºC), and varying nutrient means. Nutrient concentration means are reported in the Appendix of this thesis.

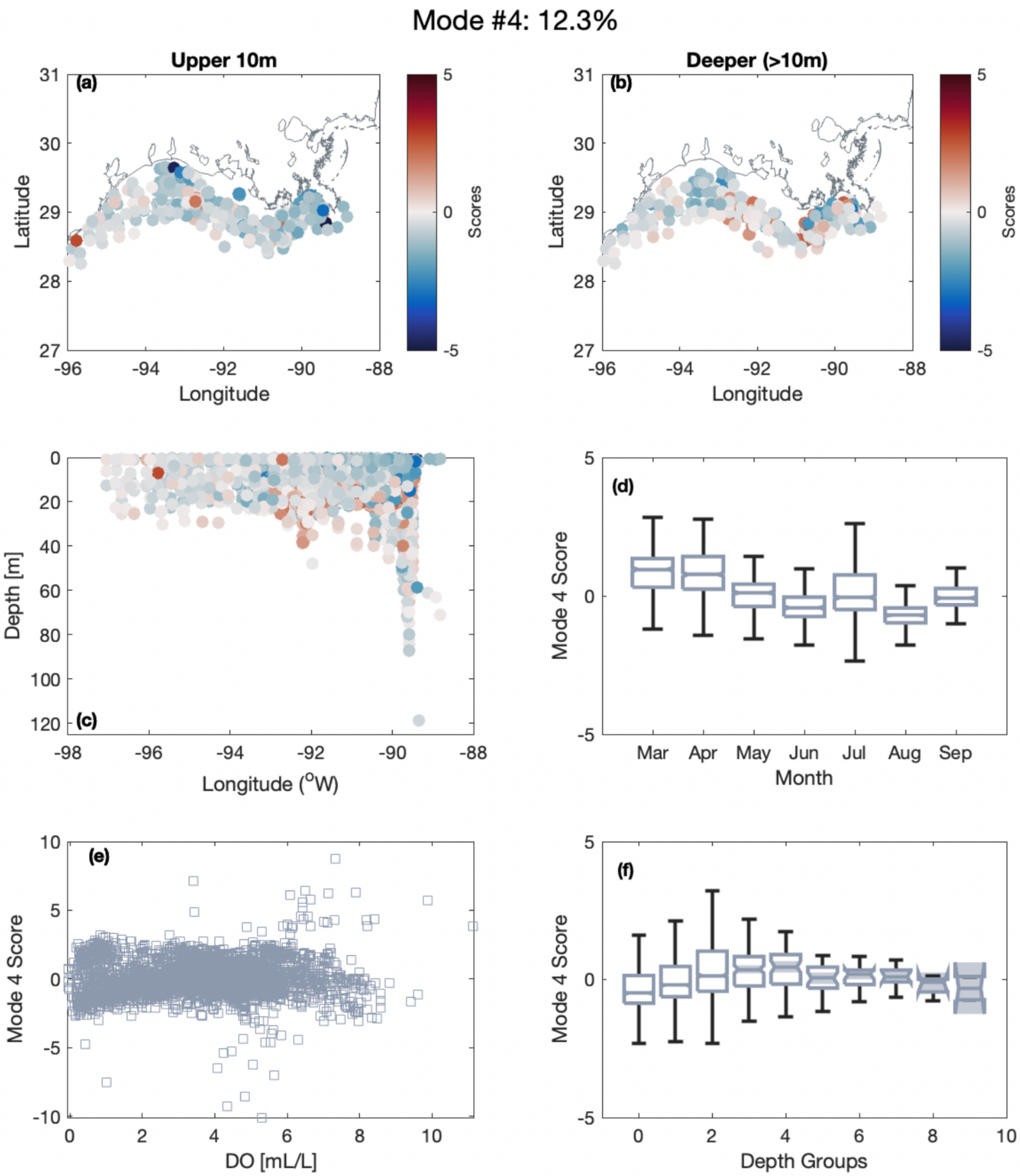


Figure 3.6 Principal Component Mode 4 (12.3% of total variance). (a) Spatial distribution of mode 4 in the upper 10 m. (b) Spatial distribution of mode 4 deeper than10 m. (c)Vertical distribution of mode 4. (d) Month means of mode 4. (e) Relationship between DO [mL/L] and Mode 4. (f) Depth groups (sample depth/10) of mode 4.

3.2 K-Means Clustering

K-means clusters can also identify patterns by grouping similar data together that aid in categorization based off hydrographic characteristics. Silhouette scoring consistently recommended two clusters as the optimal number. Since the two clusters were not discernably different from each other and were not insightful, eight clusters were chosen because it was the average of the third and fourth recommended cluster numbers. The clusters differed in sizes, and here the largest clusters are discussed. The largest were clusters 1, 5, and 7. Combined, 1, 5 and 7 account for ~75% of the data in the dataset.

Cluster #	# of Observations	% of Data
1	1695	35.6%
7	972	19.9%
5	936	19.2%
3	458	9.4%
4	313	6.4%
8	256	5.2%
2	154	3.2%
6	102	2.1%

Table 3.1 K-Means Clusters.

The spatial characteristics and the possible controlling mechanisms of the three largest clusters (clusters 1, 5 and 7) will be described in the following paragraphs. All clusters plotted together can be seen in Figure 3.7.

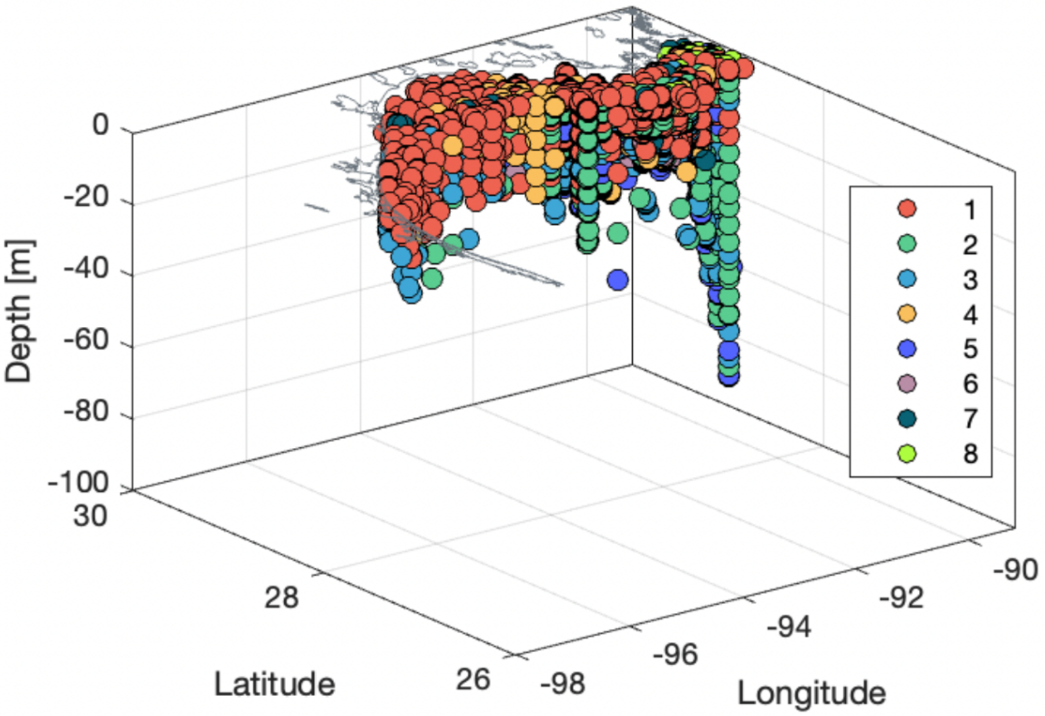


Figure 3.7 MCH data colored with 8 colors representative of the 8 K-Means Clusters on the TXLA Shelf.

The first K-means cluster is the dominant cluster containing 1,695 data points (35.6 % of the total dataset). Spatially, this cluster covers the entire east-west range of the TXLA Shelf and extends down from the surface to as far as 30 m depth (Red points in Figure 3.8). This cluster does not contain any hypoxic data points and has relatively high DO (mean = 4.5 mL/L), low salinity (mean = 30.2), high temperature (mean = 29.7 °C), and low nutrients (Table 3.2). Also, it is important to note that the majority of the values in this cluster were collected in the months of August.

Given this cluster's similar spatial distribution, hydrographic characteristics, and similar highest data collection moth in August, this cluster is analogous to the negative values of mode

1, which represents low salinity water stripped of its nutrients in the euphotic zone that spans from the Mississippi River mouth to as far west as the Texas coast as described by Rowe & Chapman (2002). A visual comparison of the spatial extent of the negative values of mode 1 and cluster 1 can be seen in Figure 3.8, and a table of the variables means are reported in Table 3.2. PC mode 1 accounted for 32.8% of the total variance in the dataset and cluster 1 accounted for 35.6% of the data. Since both PC mode 1 and cluster 1 make up ~30% of variability and data respectively, it is reasonable that both groups describe similar water on the TXLA Shelf that can be associated with RC02 zone 3. Specifically, 638 sample locations were shared between cluster 1 and the negative values of mode 1 and are shown in purple in Figure 3.8.

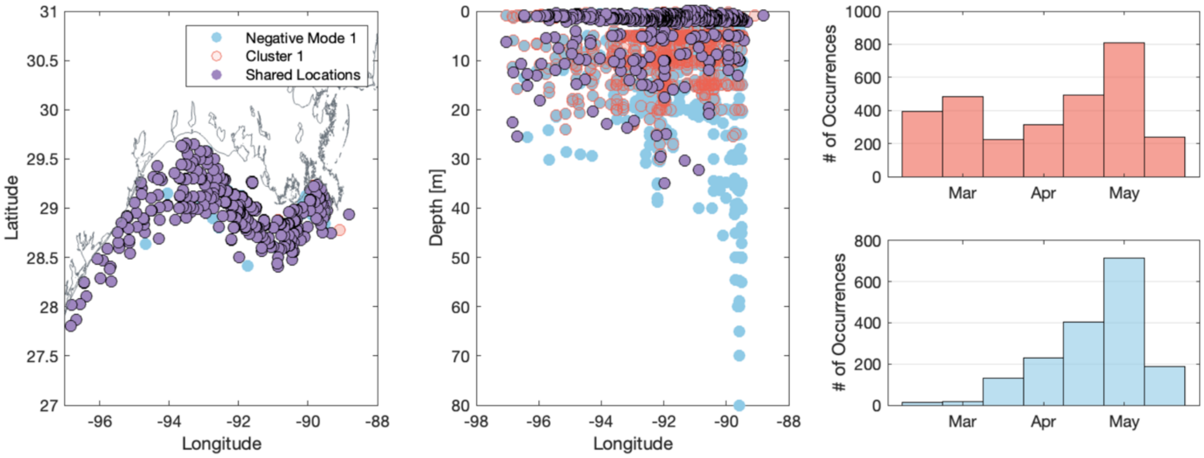


Figure 3.8 Comparison of Cluster 1 (red) and negative values of Mode 1 (blue). Shared values between cluster 1 and negative values of mode 1 are shown in purple. (a) Spatial distribution of cluster 1 and negative mode 1. (b) Vertical distribution of cluster 1 and negative mode 1. (c) Histogram of sample collection moths of cluster 1. (d) Histogram of sample collection months for negative values of mode 1.

Variable	Cluster 1	Negative Mode 1 Means
Salinity	30.2	31.3
Temperature (ºC)	29.7	26.8
DO [mL/L]	4.5	4.6
Nitrate [µmol/L]	0.5	0.8
Nitrite [µmol/L]	0.3	0.5
Total Nitrogen [µmol/L]	0.8	1.3
Phosphate [µmol/L]	0.3	0.3
Silicate [µmol/L]	5.9	5.8
Ammonia [µmol/L]	0.5	0.5
Urea [µmol/L]	0.4	0.6

Table 3.2 Variable means for cluster 1 and the negative values of mode 1.

The second largest cluster is cluster 7, which contains 972 data points. This cluster spans from the Mississippi River mouth to 94º W and 16 points are located further west than 94 ºW (Figure 3.9a) This cluster also contains points that extend deeper into the column closest to the mouth of the Mississippi River (Figure 3.9b). There are no hypoxic stations in this cluster, but there are a few high oxygen data points included near the mouth of the rivers that are likely Mississippi River plume water. Additionally, most of this cluster is from the months of March and April (Figure 3.9c). What stands out and is unique about this cluster is that it has cooler temperatures (mean= 21.4 ºC) in comparison to the whole MCH dataset. (Figure 3.9d & Table 3.3). K-means identifies homogenous subgroups and in this case this cluster is characterized with cold temperatures.

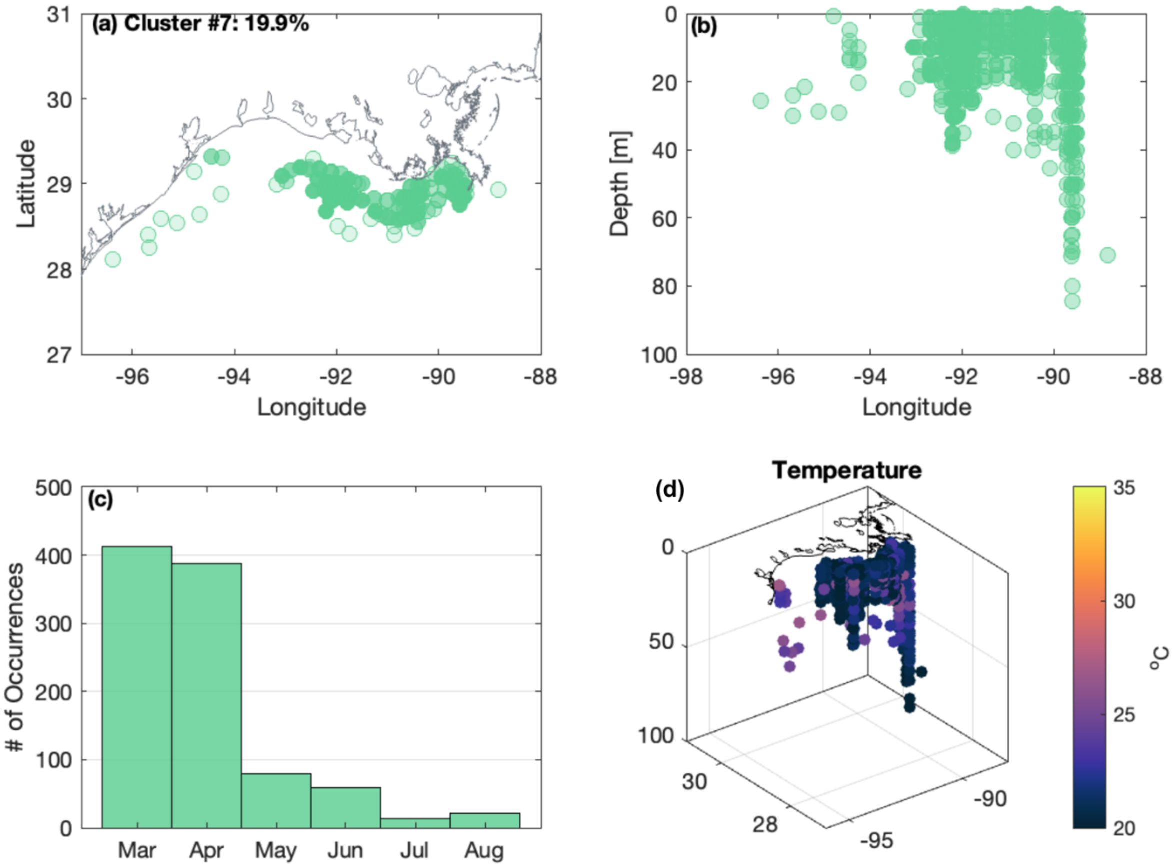


Figure 3.9 Cluster 7 (19.9%). (a) Spatial distribution of cluster 7. (b) Vertical distribution of cluster 7. (c) Histogram of sample collection months of cluster 7. (d) Cluster 7 temperature (°C)

Variable	Cluster 7 Means	Dataset Means
Salinity	32.9	32.1
Temperature (ºC)	21.4	26.2
DO [mL/L]	5.0	3.8
Nitrate [µmol/L]	1.5	2.2
Nitrite [µmol/L]	0.7	1.8
Total Nitrogen [µmol/L]	2.2	4.0
Phosphate [µmol/L]	0.3	0.5
Silicate [µmol/L]	4.9	13.3
Ammonia [µmol/L]	0.4	0.8
Urea [µmol/L]	0.5	0.7

Table 3.3 Variable means for cluster 7. Variables means of the MCH dataset are included for comparison.

The third largest cluster, cluster 5, contains 936 data points and spans across the TXLA Shelf from 89 ºW to 97 ºW and are located between 0 m and 87 m (Figure 3.10a&b). This cluster contains 75 hypoxic points. Examining the property/variable plots show that the samples in this cluster are high in silicate. (Figure 3.10d). There are two peaks in the month histogram indicating that the months of May and August were most important in this cluster (Figure 3.10c). This cluster is also characterized with anomalously high silicate values (mean = 19.5 µmol/L) (Table 3.4); therefore Cluster 5 can be described as the cluster associated with high silicate concentration.

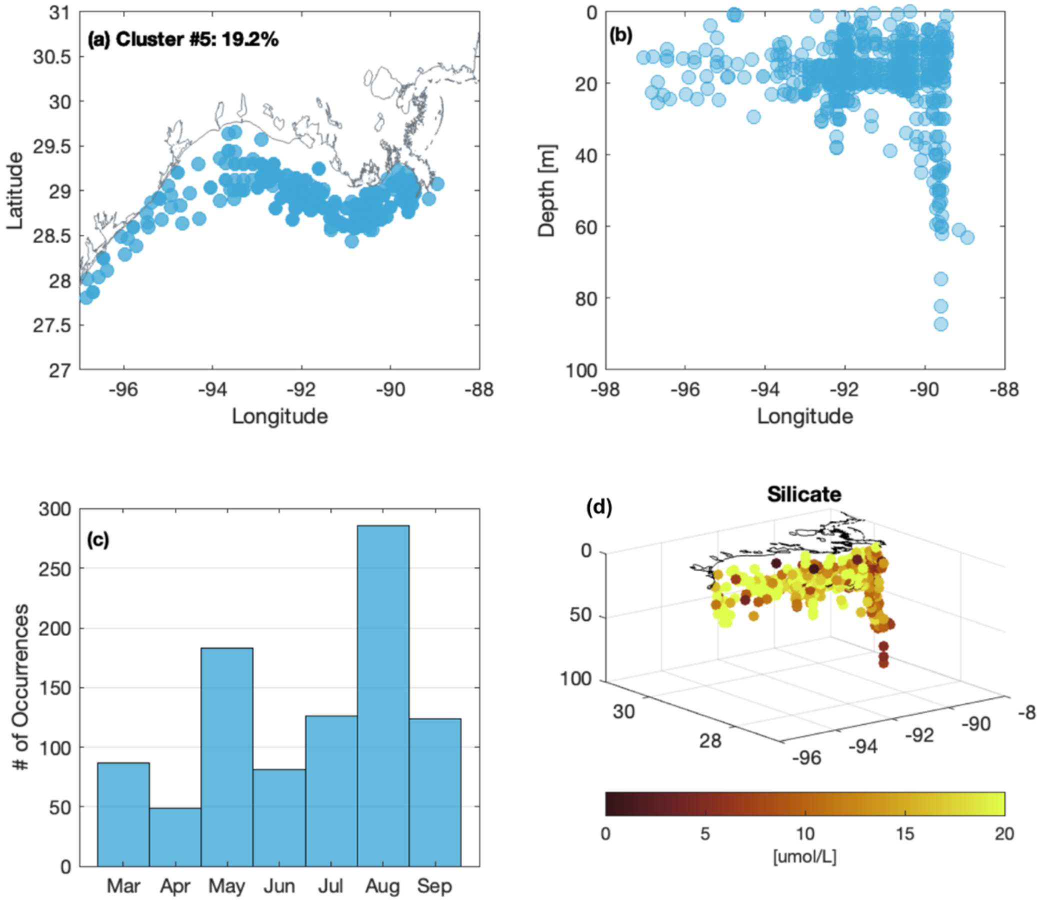


Figure 3.10 Cluster 5 (19.2%). (a) Spatial distribution of cluster 5. (b) Vertical distribution of cluster 7. (c) Histogram of sample collection months of cluster 7. (d) Cluster 7 silicate (µmol/L).

Variable	Cluster 5 Means	Dataset Means
Salinity	34.9	32.1
Temperature (ºC)	25.6	26.2
DO [mL/L]	2.6	3.8
Nitrate [µmol/L]	3.2	2.2
Nitrite [µmol/L]	1.8	1.8
Total Nitrogen [µmol/L]	4.9	4.0
Phosphate [µmol/L]	0.6	0.5
Silicate [µmol/L]	19.5	13.3
Ammonia [µmol/L]	0.8	0.8
Urea [µmol/L]	0.6	0.7

Table 3.4 Variable means for cluster 5. Variables means of the MCH dataset are included for comparison.

Most of the remaining clusters (3,4, 8, 2 and 6), which made up 25% of the data, had no clear process or mechanism to associate or describe them. Cluster 3 did not have a clear defining characteristic. Cluster 4 did not span the entirety of the TXLA Shelf but is characterized with higher nitrate values (mean = 9.1 µmol/L), high silicate (mean = 37.0 µmol/L), and hypoxic samples (mean = 1.3 mL/L). Cluster 8 is physically located deeper in the water column and is characterized with high nitrite (mean = 10.3 µmol/L), low oxygen (mean = 1.0 µmol/L), high silicate (mean = 26.5 µmol/L), and higher salinity water (mean = 35.3). Cluster 2 is characteristic of water that is high in ammonia (mean = 4.6 µmol/L), high in silicate (mean = 34.3 µmol/L) and has a mixture of hypoxia and high oxygen. Cluster 2 is located across the TXLA Shelf and spans the water column from the surface to 30 m. Lastly, cluster 6 is physically located near the Mississippi River mouth, with low salinity (mean = 19.9) and high nitrate (mean = 16.6 µmol/L) and high silicate (mean = 28.3 µmol/L) values. Based on the location and hydrographic characteristics cluster 6 is representative of the Mississippi River plume. These generalized characterizations of the clusters are based on variable plots and tables of means that can be seen in the Appendix.

To summarize, PCA showed that 75% of the variance in the dataset was accounted for in the first 4 PC modes. Mode 1 (32.8%) can represent water below the pycnocline that may be hypoxic and low salinity water stripped of its nutrients as described by Rowe & Chapman (2002). Mode 2 (17.1%) can be associated to the river plume and upwelling on the TXLA Shelf and Mode 3 represents the vertical structure of the water column. The mechanisms acting and producing the spatial and temporal variability in Mode 4 were not clear and could not easily be explained. K-means clusters 1, 7, and 5 made up 75% of the data in the MCH dataset. Cluster 1 (35.6 %) was analogous to PC mode 1 and was representative of water identified by RC02 in zone 3. Cluster 5 (19.9%) represented a group of data with cool temperatures and cluster 7 (19.2%) represented a group with high silicate. Cluster 6 (2.1%) represented the Mississippi River plume. These results indicate that PCA and K-means, both unsupervised machine learning techniques, can identify dominant drivers of hypoxia on the TXLA Shelf.

4. AN ASSESSMENT OF SUPERVISED MACHINE LEARNING ON CLASSIFYING HYPOXIA ON THE TXLA SHELF

To analyze the predictability of dissolved oxygen on the TXLA Shelf, eight supervised machine learning classification techniques were compared. These included classification algorithms such as: logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naïve bayes, k-nearest neighbors (KNN), support vector machine (SVM), decision tree, and random forest.

These classification techniques fall under the umbrella of supervised machine learning since the algorithm is trained with input forcings and desired outcomes (training dataset) to predict outcomes on a test dataset. In this case, the input forcings included salinity, temperature, nitrate, nitrite, silicate, phosphate, urea, and ammonia and the desired outcome was the labels 'hypoxic' or 'not hypoxic'. These input forcings were considered as to not introduce redundant relationships to dissolved oxygen.

4.1 Data Exploration

Prior to the application of the supervised machine learning methods to the dataset, the data were explored graphically. To illustrate the inherent complexity of this dataset, the population density histogram and pairwise relationships between variables were assembled and reviewed. The pairwise plots are visual representations of the dataset in two-dimensional sections, which are shown as an array of plots of paired features. Each plot of the array is a pairwise relation of features labeled by row and column number. The main diagonal of subplots of Figure 4.1 represents a density histogram of each variable. These pairwise relationships highlight the correlation or lack of correlation between quantities, but more importantly can establish linear and nonlinear relationships between features. In this work, the dataset involves a mixture of linear and nonlinear features.

Figure 4.2 shows the correlation matrix of Pearson correlation coefficients for all variable combinations. Correlation coefficients closer to -1, indicate a more negative linear relationship and are represented as blue; the closer to -1, the darker the blue. Coefficients closer to 1 indicate a more positive linear relationship and are represented in red; the closer to 1, the darker the red. Coefficients of 0 indicate no relationship and are represented as white in the matrix. These data show correlation coefficients that are closer to zero for most combinations indicating the lack of a linear relationship between those variables. The exceptions are salinity, phosphate, and silicate.

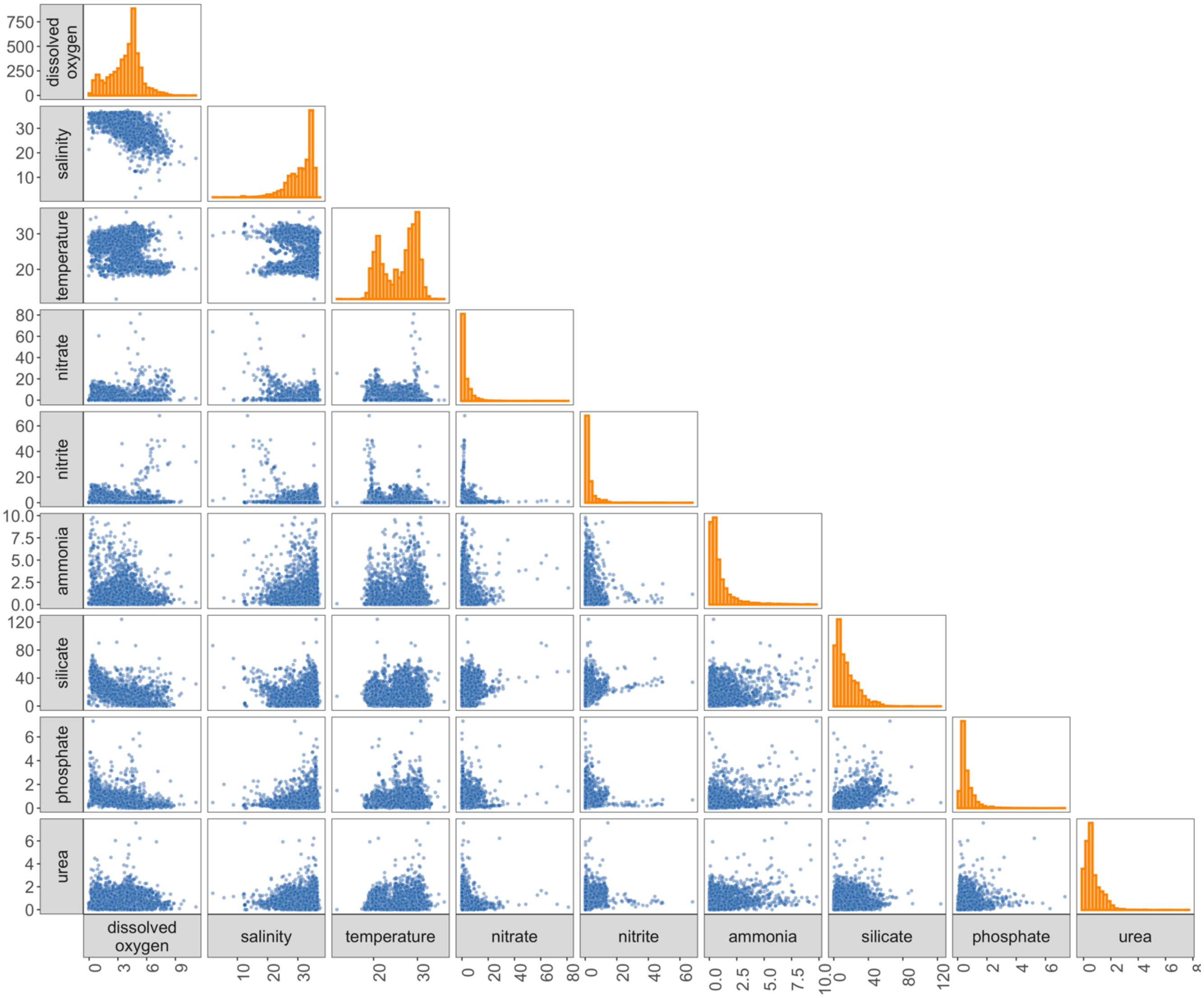


Figure 4.1 Pairwise relationships between all featured variables. Each plot shows a two-dimensional scatter plot of the variables labeled at that specific row and column. A histogram of value frequency for each quantity is shown in the diagonal of the plots grid. N=4886.

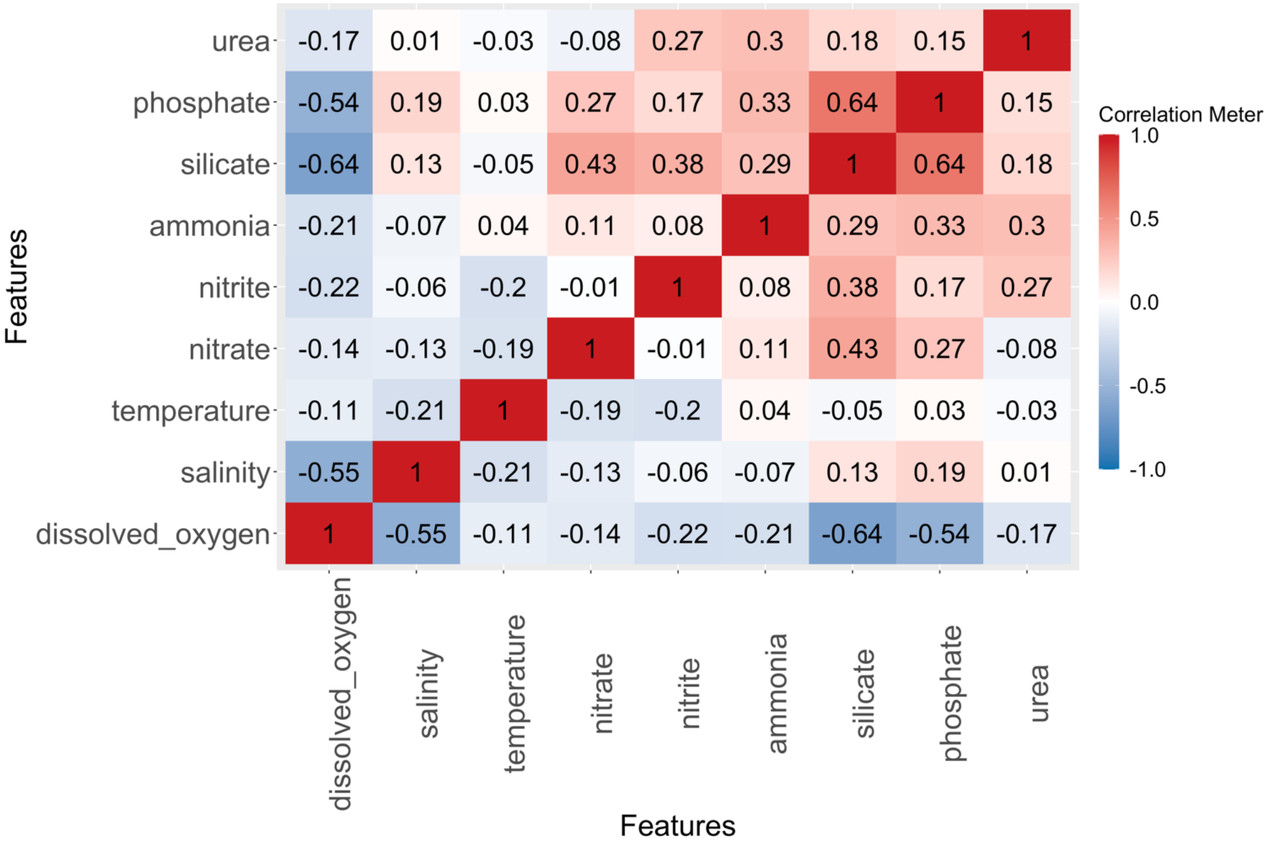
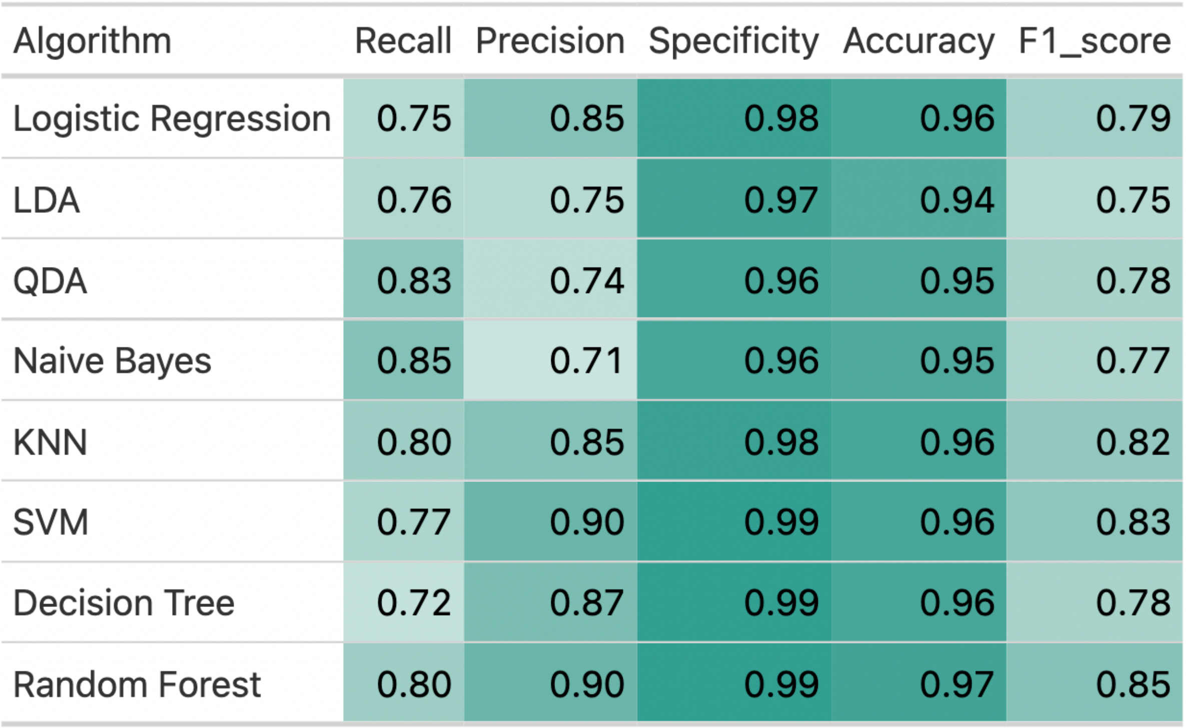


Figure 4.2 Correlation matrix of the dataset for each feature/variable combination. The Pearson correlation coefficient is given by the colorbar and it is annotated inside each grid of the matrix. Dark red indicates a positive correlation coefficient of 1, dark blue indicates a negative correlation coefficient of -1, and white indicates a correlation coefficient of 0.

4.2 Classification Comparison Results and Discussion

Based on the averaged five-fold cross-validation metrics, naive bayes had the best recall and SVM was the most precise. All eight algorithms had high specificity, and SVM, decision tree and random forest performed best in specificity out of all eight. All algorithms also scored high in accuracy with random forest being the most accurate. Lastly, KNN, SVM, and random forest had the highest F1-score (Figure 4.3).



Algorithm	Recall	Precision	Specificity	Accuracy	F1_score
Logistic Regression	0.75	0.85	0.98	0.96	0.79
LDA	0.76	0.75	0.97	0.94	0.75
QDA	0.83	0.74	0.96	0.95	0.78
Naive Bayes	0.85	0.71	0.96	0.95	0.77
KNN	0.80	0.85	0.98	0.96	0.82
SVM	0.77	0.90	0.99	0.96	0.83
Decision Tree	0.72	0.87	0.99	0.96	0.78
Random Forest	0.80	0.90	0.99	0.97	0.85

Figure 4.3 Five-fold cross-validation performance metrics for the eight classification algorithms. Darker values indicate better performance.

It is difficult to evaluate which algorithm is best to use when recall, precision, specificity, accuracy, and F1-score greatly vary across algorithms. We chose to focus on the receiving operating characteristic curve (ROC), a way to evaluate performance on classification models. A ROC curve plots the false positive rate (false hypoxic samples/ total not hypoxic samples) vs. the true positive rate (sensitivity) (true hypoxic samples/ total hypoxic labels). A perfect classifier would have 0 false positive rates and a true positive rate of 1. The ROC curves were created for all five folds of the five-fold cross-validation. Naive bayes had the most perfect curve for 4/5 folds with the highest recall (sensitivity) and low false positive rates. QDA performed best in classifying hypoxia in the fifth fold. Figure 4.4 shows the ROC curves for all eight algorithms in the first fold. QDA, KNN, and random forest also classified hypoxia well and their ROC curves were consistently just below the naive bayes ROC curve and are therefore justifiable options for classifying hypoxia in this dataset.

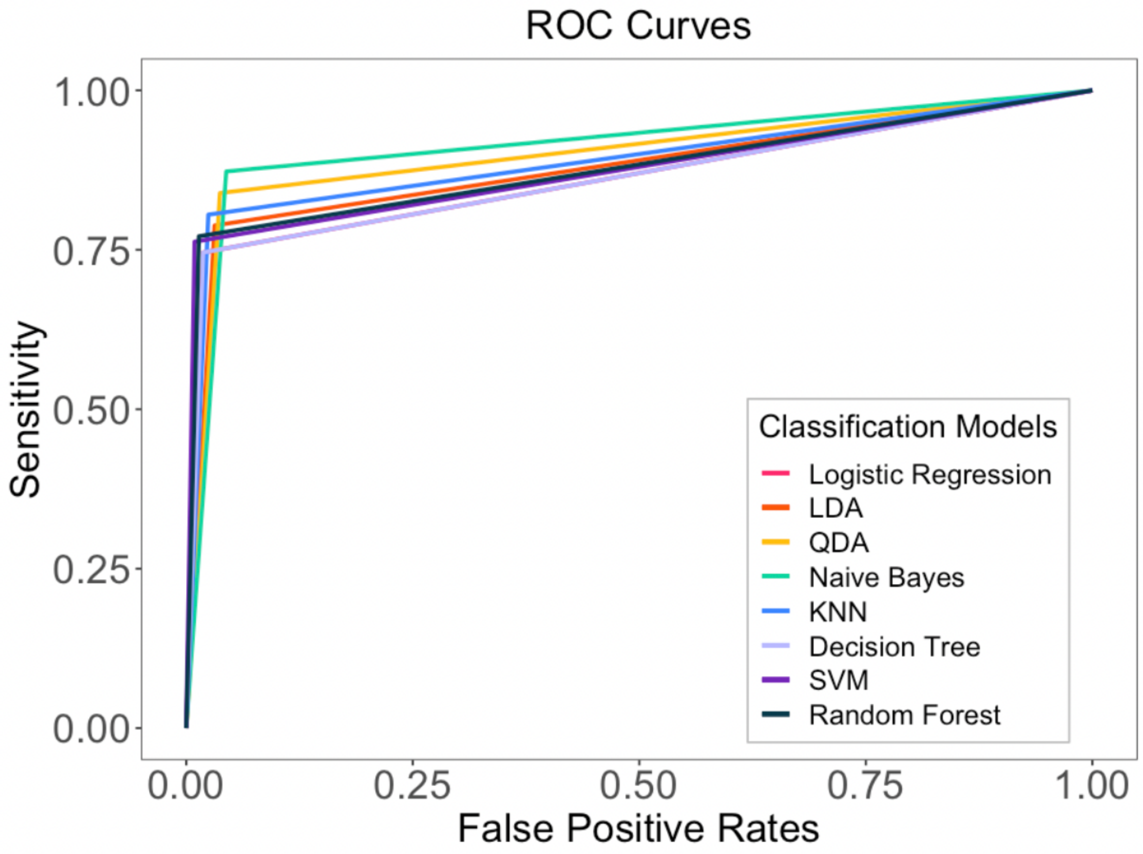
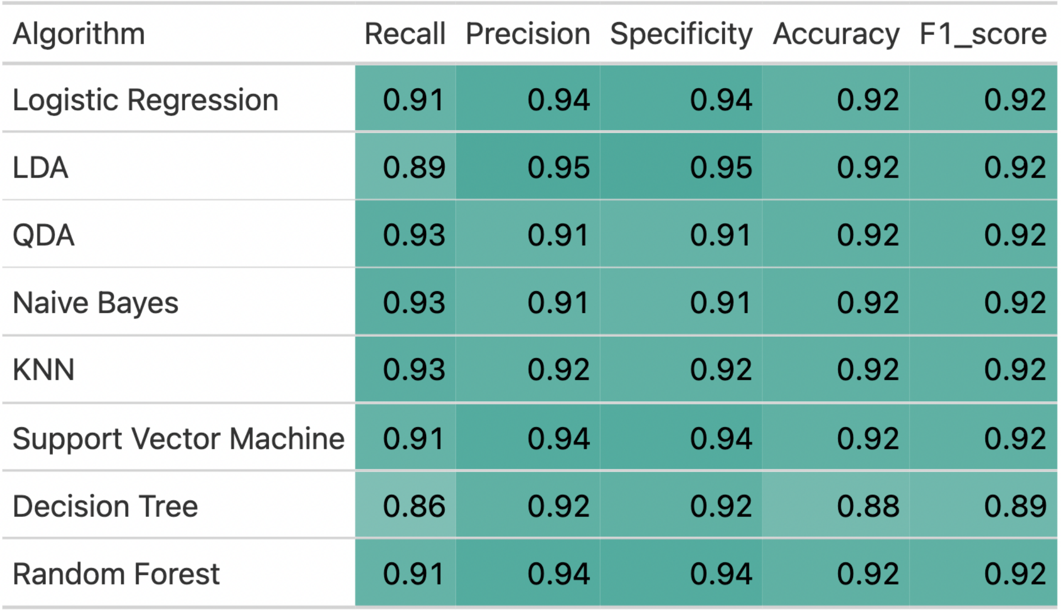


Figure 4.4 Receiving Operating Characteristic Curves for eight classification algorithms.

The MCH dataset has an imbalanced class distribution, meaning there is an uneven number of not hypoxic points to hypoxic points. Specifically, only 11% of the dataset is hypoxic (547 points). Due to the imbalanced class distribution, we also investigated how the eight algorithms performed with a balanced class distribution. The 547 hypoxic points in the dataset were randomly sampled and assigned to the training set and test set. Data points that are not hypoxic of the same sampling size as the hypoxic points (547) were randomly sampled and added to the training and test sets also. This means that the training and test sets had the same amount of data (547 points in each) and that there was an even class distribution. Training and testing the classifiers with a balanced class distribution equalized the metrics by increasing their performance in recall, precision and F1-score and moderately decreasing performance in specificity and accuracy as seen in Figure 4.5. The ROC curves were also equalized, meaning that classification performance across the eight algorithms were similar and differences between them were difficult to discern (Figure 4.6). Recall ($p = 5.5 \times 10^{-6}$), precision ($p = 0.0015$), accuracy ($p = 7 \times 10^{-6}$), and f1-score ($p = 1.6 \times 10^{-7}$) were all significantly different between the unbalanced and the balanced models. Specificity was not significantly different between unbalanced and balanced models ($p = 0.17$). The classifiers' ability to correctly predict not hypoxic samples was not affected by whether it was trained with balanced or unbalanced data. These results indicate that although the algorithm you choose is important, how the model is trained is significant and can greatly affect the performance of classification in recall, precision, accuracy, and f1-score, but not in specificity.



Algorithm	Recall	Precision	Specificity	Accuracy	F1_score
Logistic Regression	0.91	0.94	0.94	0.92	0.92
LDA	0.89	0.95	0.95	0.92	0.92
QDA	0.93	0.91	0.91	0.92	0.92
Naive Bayes	0.93	0.91	0.91	0.92	0.92
KNN	0.93	0.92	0.92	0.92	0.92
Support Vector Machine	0.91	0.94	0.94	0.92	0.92
Decision Tree	0.86	0.92	0.92	0.88	0.89
Random Forest	0.91	0.94	0.94	0.92	0.92

Figure 4.5 K-Fold cross validation performance metrics for the eight classification algorithms trained with a balanced dataset. Darker values indicate better performance.

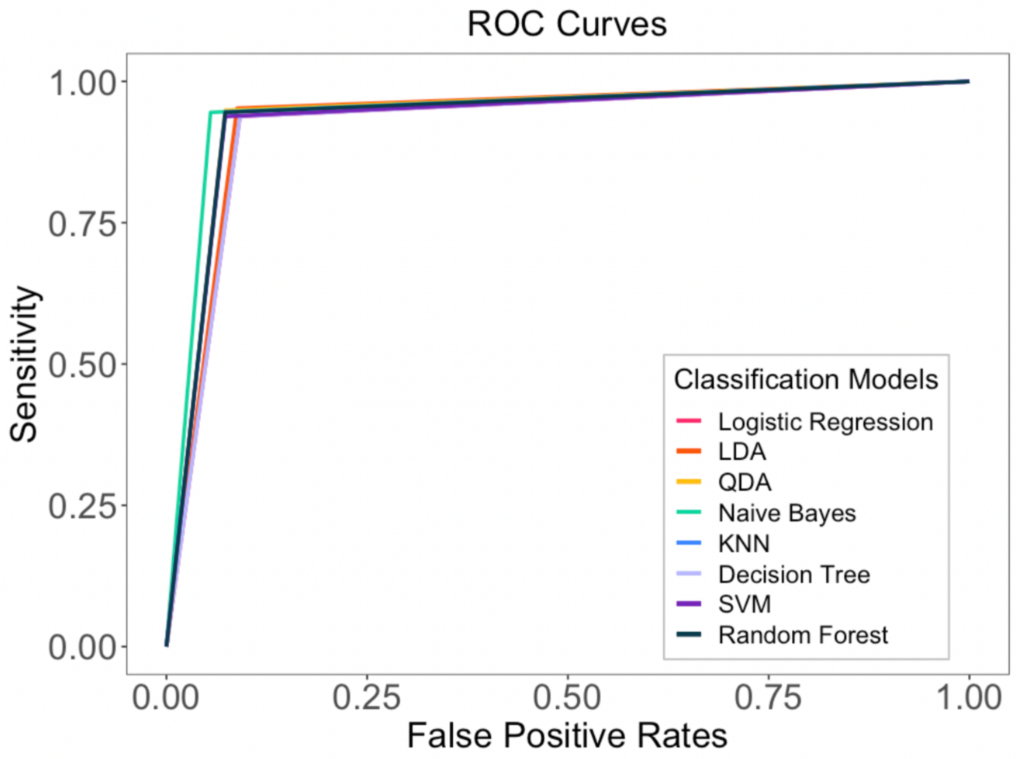
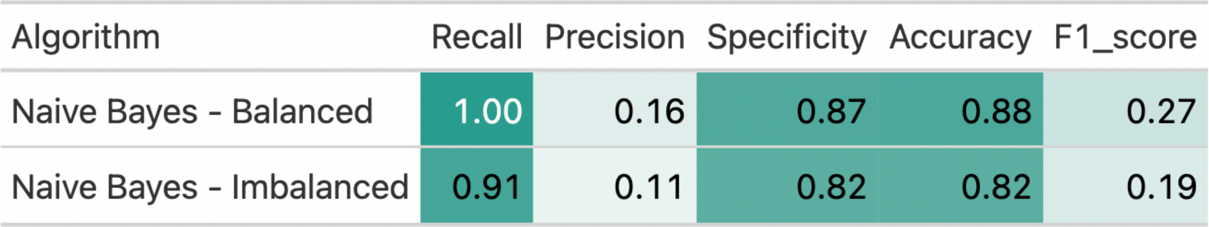


Figure 4.6 Receiving Operating Characteristic Curves for eight classification algorithms trained and tested with even class distribution.

4.3 Ancillary Data Validation Results and Discussion

Since naive bayes performed best on the imbalanced dataset, we tested the model with the supporting data. The training set included the entire MCH dataset and the test set was the combined REU research cruises of 2016, 2017, and 2018 and XR01 and XR02. The combined ancillary data included 11 hypoxic samples out of 457 samples. This model had a high sensitivity with a recall of 0.91, meaning it correctly classified most of the hypoxic samples. However, the precision was poor, meaning it had many false positives or Type 1 errors (i.e., the model mislabeled 83 not hypoxic stations as hypoxic). To further investigate naïve bayes performance on classification in practice this test was repeated for a balanced training set. The metrics slightly improved and the ROC curve indicates that training with balanced data increased performance but differences were not significant (p = 0.8) (Figure 4.7 & 4.8).

To sum up, the naïve bayes classifier model was highly sensitive and correctly labeled all hypoxic samples in the ancillary data but lacked precision. This could be because the test data only included samples from the months of June, April and August and samples were collected further offshore than the MCH training dataset. Performance of the model on predicting hypoxia in the ancillary data was not affected by differences in unbalanced and balanced training.



Algorithm	Recall	Precision	Specificity	Accuracy	F1_score
Naive Bayes - Balanced	1.00	0.16	0.87	0.88	0.27
Naive Bayes - Imbalanced	0.91	0.11	0.82	0.82	0.19

Figure 4.7 Performance metrics on the Naïve Bayes classifier trained with balanced data and imbalanced data.

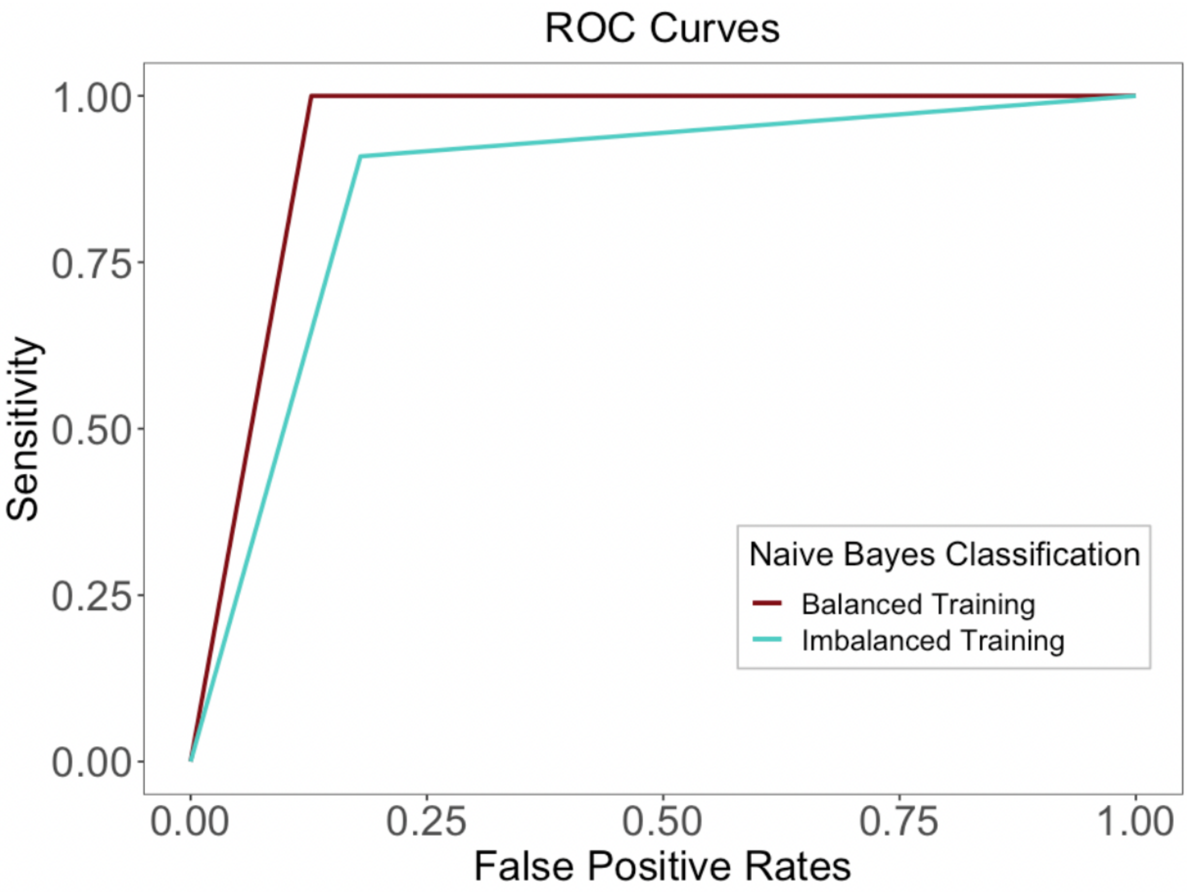


Figure 4.8 ROC curves for Naïve Bayes classifier trained with balanced and imbalanced data.

5. SUMMARY AND CONCLUSIONS

Observations of dissolved oxygen, salinity, temperature, and six dissolved nutrient concentrations of the waters on the TXLA Shelf in the months of March – September in 2004 – 2014 were used in unsupervised and supervised machine learning techniques to identify patterns of variability and to evaluate the performance of classification algorithms for classifying hypoxia on the TXLA Shelf. Unsupervised machine learning techniques, principal component analysis and K-means clustering, were used to identify variability patterns that were ascribed to previously known drivers and processes of hypoxia in the region. Eight classification algorithms such as logistic regression, LDA, QDA, naïve bayes, KNN, SVM, decision tree, and random forest were trained with the observations on TXLA Shelf and their performances were compared. Naïve Bayes performed best on classifying hypoxia and when tested with supporting data it had high recall and low false positive rates.

Results from the PCA analysis showed that 75% of the variance in the dataset was accounted for in the first PC modes. Mode 1 (32.8%) was representative of water below the pycnocline that may be hypoxic and low salinity water stripped of its nutrients as described by Rowe & Chapman (2002). Mode 2 (17.1%) can be attributed to the river plume and upwelling on the TXLA Shelf and Mode 3 represents the vertical structure of the water column. The mechanisms acting and producing the spatial and temporal variability in Mode 4 were not clear and could not easily be explained.

Results from K-means clustering showed that 75% of the data was represented in Cluster 1, 7 and 5. Cluster 1 represented the same water that the negative scores of PC Mode 1 represented (low salinity water stripped of its nutrients – Rowe & Chapman 2002, zone 3 water. Cluster 7 represented a homogenous group of cool water on the eastern TXLA Shelf. Cluster 5 represented water that was high in silicate and cluster 6 represented the Mississippi River plume. Remaining clusters were characterized by their homogeneous hydrographic characteristics that were not clearly representative of a single mechanism.

These results indicate that unsupervised machine learning methods such as PCA and K-Means can successfully aid in identification of drivers of hypoxia. Our results also indicate and validate that the vertical structure of the water column and the river plume play a significant role in controlling the variability and spatial distribution of hypoxia on the TXLA Shelf. Tools that can identify patterns of variability can and should be used as a diagnostic tool for hypoxia. The modes identified could be further used to form and train a machine learning model for more accurate prediction that is based on multivariate data and dominant controlling mechanisms to be applied to datasets where DO is missing. This has implications for the ability to detect and indirectly determine if hypoxia is present for coastal hypoxia events and has important implications for management options.

Results from the classification comparison of the eight algorithms showed that naïve bayes performed the best at classifying hypoxia on the MCH dataset. When the training set was balanced, performance of the eight classifiers were equalized and performance was strongly dependent on training data. The naïve bayes classifier was further explored with the ancillary data and lacked precision and excelled in identifying hypoxic samples. When trained with a balanced dataset, performance was not significantly different from the model performance that was trained with imbalanced data. These results went against the assumption that the algorithm would excel at classifying what the majority of what the training data was. This was surprising that precision was lacking since 90% of the data was not hypoxic. Sensitivity of the model was high and is valuable for extreme event predictions such as hypoxia.

The second part of this study showed that supervised machine learning techniques can use hydrographic characteristics that are associated with hypoxia as a predictive tool for determining if hypoxia is present. Although ML methods can classify hypoxia, performance is reliant on input training data and should be carefully considered when applied for prediction. Our metric to determine classifier performance, the ROC curve, is satisfactory for determining performance since it takes the balance between true positive and false positive rates into account. A classifier used for predictive application should have high accuracy in identifying true positives and minimize false positives. Future users of machine learning should consider this relationship when applying machine learning for prediction.

This study establishes that straightforward machine learning techniques can aid in identifying known main drivers of hypoxia and that hydrographic characteristics of those processes can be used to predict hypoxia on the TXLA Shelf. These techniques have the potential to be applied to other hydrographic data where DO is missing to evaluate if hypoxia is present, a powerful tool that could be used in water quality and resource management in the region. Also, with the continued increase in ocean observing and growing datasets available, data driven methods such as machine learning should be utilized and implemented into the formation of policy and management since ML can easily replicate the significant drivers. While the approaches presented in this study were specifically for the TXLA Shelf, the methodology should be readily applicable to other coastal systems and locations with similar datasets.

REFERENCES

Ammerman, J., & Sylvan, J. (2004). Phosphorus Limitation of Phytoplankton Growth in the Mississippi River Plume: A Case for Dual Nutrient Control? *AGU Fall Meeting Abstracts, -1*, 07.

Bayes, T. 1764. "An Essay Toward Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

Belabbassi, L. (2006). Examination of the relationship of river water to occurrences of bottom water with reduced oxygen concentrations in the northern Gulf of Mexico. Doctoral dissertation, Texas A&M University. Texas A&M University. Available electronically from https : / /hdl .handle .net /1969 .1 /5018.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bianchi, T. S., DiMarco, S. F., Cowan, J. H., Hetland, R. D., Chapman, P., Day, J. W., & Allison, M. A. (2010). The science of hypoxia in the Northern Gulf of Mexico: A review. *Science of the Total Environment, 408*(7), 1471-1484. <Go to ISI>://WOS:000275970800001

Boesch, D., Brinsfield, R.B., Magnien, R.E. (2001). Chesapeake Bay eutrophication. J. Environ. Qual. 30, 303–320.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. *A training algorithm for optimal margin classifiers.* Paper presented at the Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory.1992

Chang, N.-B., Xuan, Z., & Yang, Y. J. (2013). Exploring spatiotemporal patterns of phosphorus concentrations in a coastal bay with MODIS images and machine learning models. *Remote Sensing of Environment, 134*, 100-110. https://www.sciencedirect.com/science/article/pii/S0034425713000746

Cho, K. W., Reid, R. O., & Nowlin, W. D. (1998). Objectively mapped stream function fields on the Texas-Louisiana shelf based on 32 months of moored current meter data. *Journal of Geophysical Research-Oceans, 103*(C5), 10377-10390. <Go to ISI>://WOS:000073634000011

Cochrane, J. D., & Kelly, F. J. (1986). Low-Frequency Circulation on the Texas-Louisiana Continental Shelf. *Journal of Geophysical Research-Oceans, 91*(C9), 645-659. <Go to ISI>://WOS:A1986E074200014

Coopersmith, E. J., Minsker, B., & Montagna, P. (2011). Understanding and forecasting hypoxia using machine learning algorithms. *Journal of Hydroinformatics, 13*(1), 64-80. <Go to ISI>://WOS:000287301900006

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Dale, V. H., Kling, C. L., Meyer, J. L., Sanders, J., Stallworth, H., Armitage, T., et al. (2010). Characterization of Hypoxia. In *Hypoxia in the Northern Gulf of Mexico* (pp. 9-50).

Deng, T., Chau, K.-W., & Duan, H.-F. (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management, 284*, 112051. https://www.sciencedirect.com/science/article/pii/S0301479721001134

Diaz R. J. (2001). Overview of hypoxia around the world. Journal of environmental quality, 30(2), 375-281. https://doi.org/10.2134/jeq2001.302275x

Diaz, R. J., & Rosenberg, R. (2006). Spreading dead zones and consequences for marine ecosystems. Science 321(5891(, 926-929. <Go to ISI>://WOS:000258436700028

DiMarco, S. F., Chapman, P., Walker, N., & Hetland, R. D. (2010). Does local topography control hypoxia on the eastern Texas–Louisiana shelf? *Journal of Marine Systems, 80*(1), 25-35. https://www.sciencedirect.com/science/article/pii/S0924796309002620

DiMarco, S. F., Strauss, J., May, N., Mullins-Perry, R. L., Grossman, E. L., & Shormann, D. (2012). Texas Coastal Hypoxia Linked to Brazos River Discharge as Revealed by Oxygen Isotopes. *Aquatic Geochemistry, 18*(2), 159-181. <Go to ISI>://WOS:000300281200005

DiMarco, S.F. & Zimmerle, H.M. (2017). MCH Atlas: Oceanographic Observations of the Mechanism Controlling Hypoxia Project. Texas A&M University, Texas Sea Grant Publication TAMU-SG-17-601. Pp. 350. ISBN: 978-0-692-87961-0. http://mchatlas.tamu.edu

Dortch, Q., & Whitledge, T. E. (1992). Does nitrogen or silicon limit phytoplankton production in the Mississippi River plume and nearby regions? *Continental Shelf Research, 12*(11), 1293-1309. https://www.sciencedirect.com/science/article/pii/027843439290065R

Etter, P. C., Howard, M. K., & Cochrane, J. D. (2004). Heat and freshwater budgets of the Texas-Louisiana shelf. *Journal of Geophysical Research-Oceans, 109*(C2). <Go to ISI>://WOS:000220217200004

Feng, Y., DiMarco, S., & Jackson, G. (2012). Relative role of wind forcing and riverine nutrient input on the extent of hypoxia in the northern Gulf of Mexico. *Geophysical Research Letters, 39*(9).

Feng, Y., Fennel, K., Jackson, G. A., DiMarco, S. F., & Hetland, R. D. (2014). A model study of the response of hypoxia to upwelling-favorable wind on the northern Gulf of Mexico shelf. *Journal of Marine Systems, 131*, 63-73. <Go to ISI>://WOS:000333501700006

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique, 57*(3), 238-247.

Forrest, D. R., Hetland, R. D., & DiMarco, S. F. (2011). Multivariable statistical regression models of the areal extent of hypoxia over the Texas-Louisiana continental shelf. *Environmental Research Letters, 6*(4). <Go to ISI>://WOS:000298674700036

Guo, H., Huang, J. J., Zhu, X., Wang, B., Tian, S., Xu, W., & Mai, Y. (2021). A generalized machine learning approach for dissolved oxygen estimation at multiple spatiotemporal scales using remote sensing. *Environmental Pollution, 288*, 117734. https://www.sciencedirect.com/science/article/pii/S0269749121013166

Hetland, R. D., & DiMarco, S. F. (2008). How does the character of oxygen demand control the structure of hypoxia on the Texas–Louisiana continental shelf? *Journal of Marine Systems, 70*(1), 49-62. https://www.sciencedirect.com/science/article/pii/S0924796307000711

Kim, J., Chapman, P., Rowe, G., & DiMarco, S. F. (2020). Categorizing zonal productivity on the continental shelf with nutrient-salinity ratios. *Journal of Marine Systems, 206*. <Go to ISI>://WOS:000525321600005

Kim, J., Chapman, P., Rowe, G. T., DiMarco, S. F., & Thornton, D. C. O. (2020). Implications of different nitrogen input sources for potential production and carbon flux estimates in the coastal Gulf of Mexico (GOM) and Korean Peninsula coastal waters. *Ocean Science, 16*(1), 45-63. <Go to ISI>://WOS:000506335500001

Lahiry, S. (2007). Relationships between nutrients and dissolved oxygen concentrations on the Texas-Louisiana shelf during summer of 2004. Master's thesis, Texas A&M University. Available electronically from https : / /hdl .handle .net /1969 .1 /ETD - TAMU -1950.

McDougall, T.J. & P.M. Barker (2011). Getting started with TEOS-10 and the Gibbs Seawater (GSW) Oceanographic Toolbox, 28pp., SCOR/IAPSO WG127, ISBN 978-0-646-55621-5.

Milliman, J. D., & Meade, R. H. (1983). World-wide Delivery of River Sediment to the Oceans. *Journal of Geology, 91*(1), 1-21. <Go to ISI>://WOS:A1983PY94200001

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association, 58*, 415-434.,

Murray, S. P. (ED.). 1998. An observational study of the Mississippi-Atchafalaya coastal plume: final report. OCS Study M.MS 98-0040. U.S. Department of the Interior, .Minerals Management Service, Gulf of Mexico OCS Region, New Orleans, LA.
---------, E. Jarosz, and E. T. Weeks. 1998. Physical oceanographic observations of the coastal plume, p. 5-105. In: S. P. Murray (eel.). An observational study of the Mississippi-Atchafalaya coastal plume: final report. OCS Study MMS 98-0040. U.S. Department of the Interior, Minerals Management Service, Gulf of Mexico OCS Region, New Orleans, LA.

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review, 110*(7), 699-706.

Nowlin, W. D., Jochens, A. E., DiMarco, S. F., Reid, R. O., & Howard, M. K. (2005). Low-frequency circulation over the Texas-Louisiana continental shelf. *Geophysical Monograph-American Geophysical Union, 161*, 219.

Nowlin WD Jr, Jochens AE, DiMarco SF, Reid RO, Howard MK (2001). Deepwater physical oceanography reanalysis and synthesis of historical data: synthesis report. OCS study MMS 2001-064, U.S. Dept. of the Interior, Minerals Management Service, Gulf of Mexico OCS Region, New Orleans, LA; 2001. 530 pp.

Nowlin Jr WD, Jochens AE, Reid RO, DiMarco SF (1998). Texas–Louisiana shelf circulation and transport processes study: synthesis report. Volume I: technical report. OCS Study MMS 98 0035. U.S. Dept. of the Interior, Minerals Management Service, Gulf of Mexico OCS Region, New Orleans, LA; 1998. 502 pp.

Osterman, L. E., Poore, R. Z., & Swarzenski, P. W. (2008). The last 1000 years of natural and anthropogenic low-oxygen bottom-water on the Louisiana shelf, Gulf of Mexico. *Marine Micropaleontology, 66*(3-4), 291-303. <Go to ISI>://WOS:000254134000010

Osterman, L. E., Poore, R. Z., Swarzenski, P. W., Senn, D. B., & DiMarco, S. F. (2009). The 20th-century development and expansion of Louisiana shelf hypoxia, Gulf of Mexico. *Geo-Marine Letters, 29*(6), 405-414. <Go to ISI>://WOS:000271949800006

Rabalais, N. N., Turner, R. E., Dortch, Q., Justic, D., Bierman, V. J., & Wiseman, W. J. (2002). Nutrient-enhanced productivity in the northern Gulf of Mexico: past, present and future. *Hydrobiologia, 475*(1), 39-63. <Go to ISI>://WOS:000178252700004

Rabalais, N. N., Turner, R. E., Justic, D., Dortch, Q., Wiseman, W. J., & SenGupta, B. K. (1996). Nutrient changes in the Mississippi River and system responses on the adjacent continental shelf. *Estuaries, 19*(2B), 386-407. <Go to ISI>://WOS:A1996UX17700005

Rabalais, N. N., Turner, R. E., & Scavia, D. (2002). Beyond science into policy: Gulf of Mexico hypoxia and the Mississippi River. *Bioscience, 52*(2), 129-142. <Go to ISI>://WOS:000173839600004

Rabalais, N. N., Turner, R. E., Sen Gupta, B. K., Boesch, D. F., Chapman, P., & Murrell, M. C. (2007). Hypoxia in the northern Gulf of Mexico: Does the science support the plan to reduce, mitigate, and control hypoxia? *Estuaries and Coasts, 30*(5), 753-772. <Go to ISI>://WOS:000252201300002

Rabalais, N. N., Wiseman, W. J., & Turner, R. E. (1994). Comparison of Continuous Recrods of Near-Bottom Dissolved-Oxygen From the Hypoxia Zone Along the Louisiana Coast. *Estuaries, 17*(4), 850-861. <Go to ISI>://WOS:A1994QJ25600012

Rabalais, N.N., Díaz, R.J., Levin, L.A., Turner, R.E., Gilbert, D., Zhang, J., (2010). Dynamics and distribution of natural and human-caused hypoxia. Biogeosciences 7 (2), 585–619.

Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 159–203. http://www.jstor.org/stable/2983775

Ross, A. C., & Stock, C. A. (2019). An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science, 221*, 53-65. https://www.sciencedirect.com/science/article/pii/S0272771418310217

Rowe, G., & Chapman, P. (2002). Continental Shelf Hypoxia: Some Nagging Questions. *Gulf Mexico Sci, 20*, 153-160.

Sharples, J., Middelburg, J. J., Fennel, K., & Jickells, T. D. (2017). What proportion of riverine nutrients reaches the open ocean? *Global Biogeochemical Cycles, 31*(1), 39-58. https://doi.org/10.1002/2016GB005483. https://doi.org/10.1002/2016GB005483

Sylvan, J. B., Dortch, Q., Nelson, D. M., Brown, A. F. M., Morrison, W., & Ammerman, J. W. (2006). Phosphorus limits phytoplankton growth on the Louisiana shelf during the period of hypoxia formation. *Environmental Science & Technology, 40*(24), 7548-7553. <Go to ISI>://WOS:000242793400019

Sylvan, J. B., Quigg, A., Tozzi, S., & Ammerman, J. W. (2007). Eutrophication-induced phosphorus limitation in the Mississippi River plume: Evidence from fast repetition rate fluorometry. *Limnology and Oceanography, 52*(6), 2679-2685. <Go to ISI>://WOS:000251129700031

Turner, R. E., Rabalais, N. N., Swenson, E. M., Kasprzak, M., & Romaire, T. (2005). Summer hypoxia in the northern Gulf of Mexico and its prediction from 1978 to 1995. *Marine Environmental Research, 59*(1), 65-77. <Go to ISI>://WOS:000224011400004

Valera, M., Walter, R. K., Bailey, B. A., & Castillo, J. E. (2020). Machine Learning Based Predictions of Dissolved Oxygen in a Small Coastal Embayment. *Journal of Marine Science and Engineering, 8*(12). <Go to ISI>://WOS:000602008800001

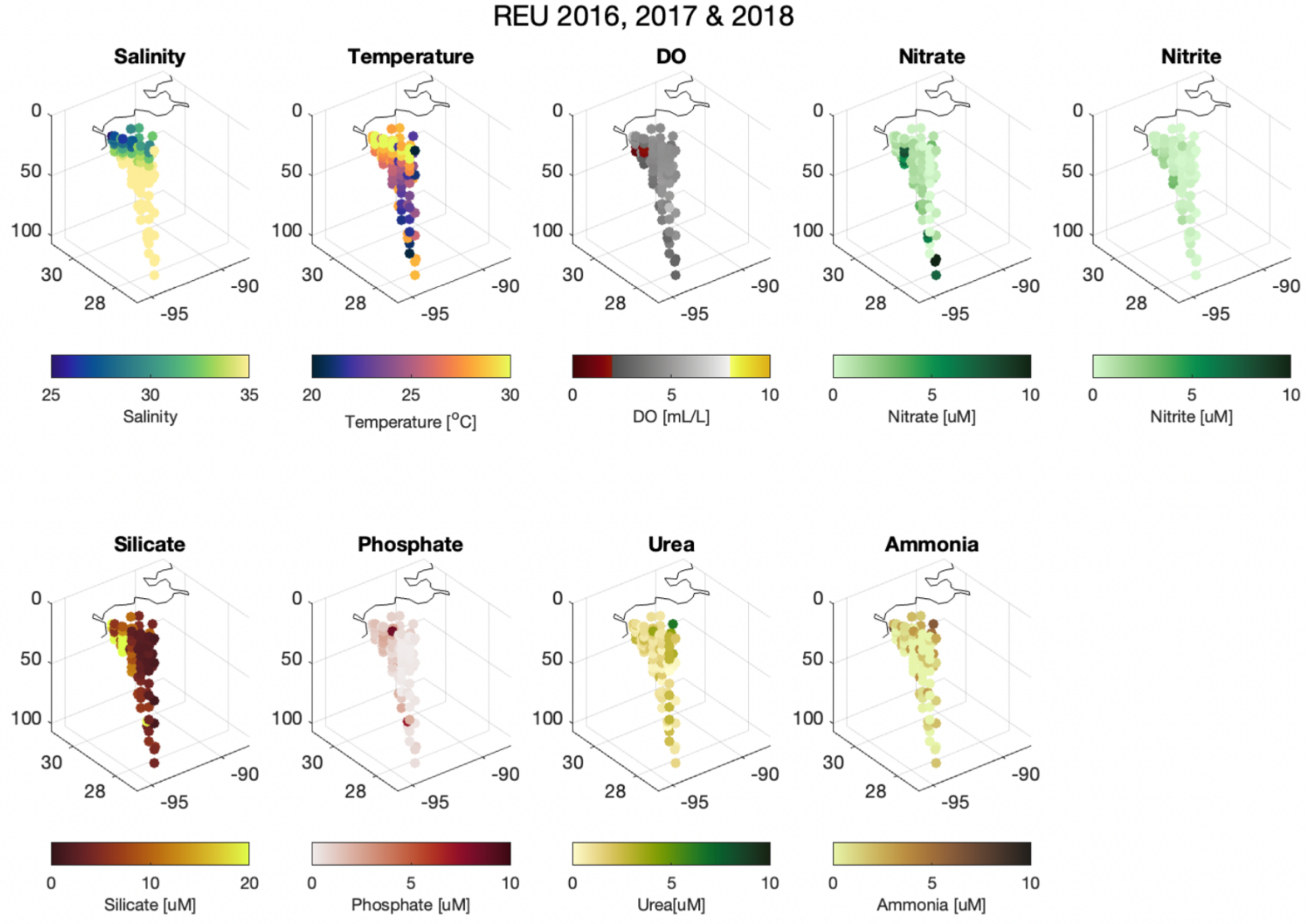
Wiseman, W. J., Rabalais, N. N., Turner, R. E., Dinnel, S. P., & MacNaughton, A. (1997). Seasonal and interannual variability within the Louisiana coastal current: stratification and hypoxia. *Journal of Marine Systems, 12*(1-4), 237-248. <Go to ISI>://WOS:A1997YB09500017

Yu, J., Tian, Y., Wang, X., & Zheng, C. (2021). Using machine learning to reveal spatiotemporal complexity and driving forces of water quality changes in Hong Kong marine water. *Journal of Hydrology, 603*, 126841. https://www.sciencedirect.com/science/article/pii/S002216942100891X

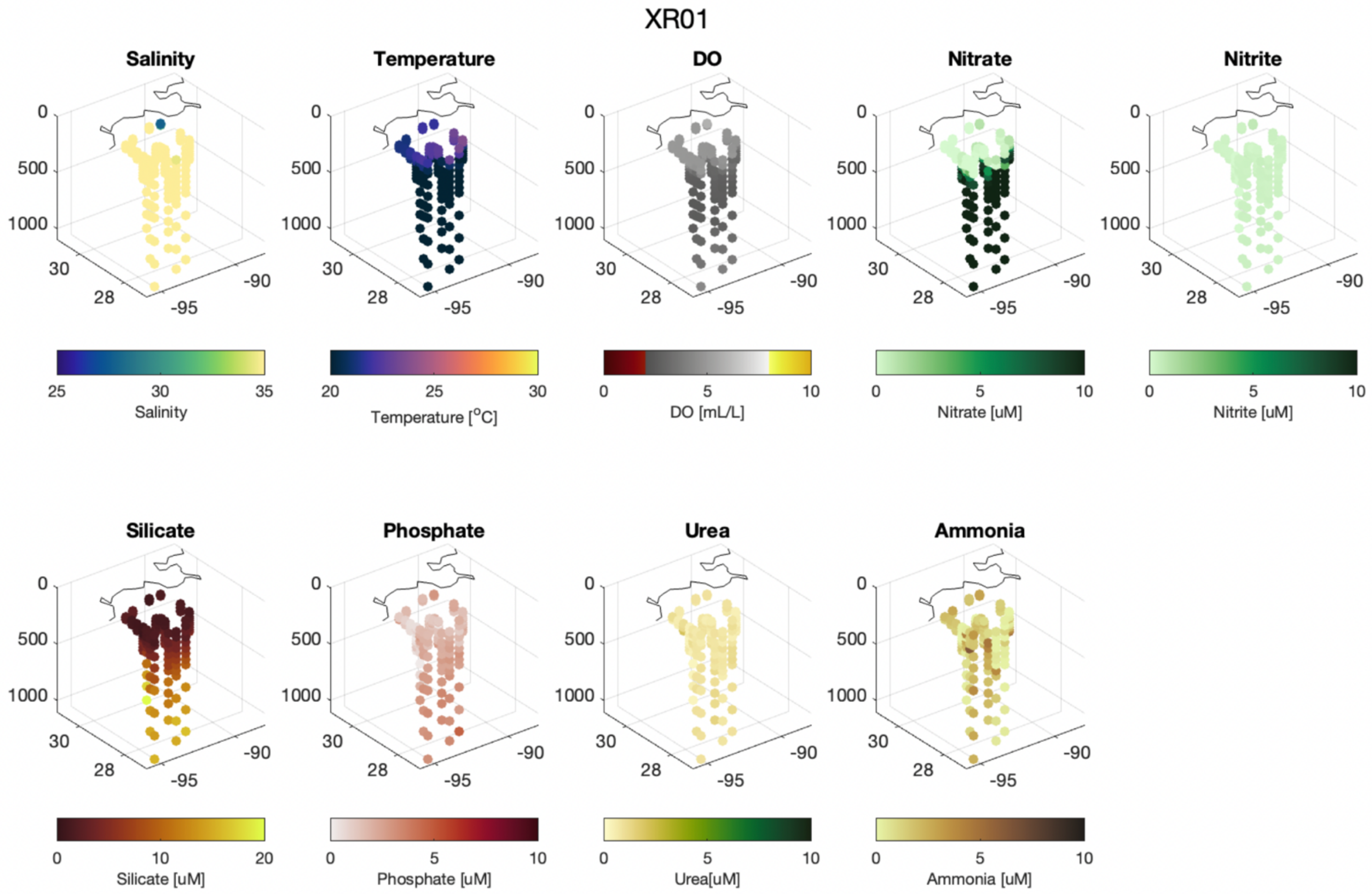
Yu, X., Shen, J., & Du, J. B. (2020). A Machine-Learning-Based Model for Water Quality in Coastal Waters, Taking Dissolved Oxygen and Hypoxia in Chesapeake Bay as an Example. *Water Resources Research, 56*(9). <Go to ISI>://WOS:000578452200058

Zhang, W. X., Hetland, R. D., DiMarco, S. F., & Fennel, K. (2015). Processes controlling mid-water column oxygen minima over the Texas-Louisiana shelf. *Journal of Geophysical Research-Oceans, 120*(4), 2800-2812. <Go to ISI>://WOS:000354417200024

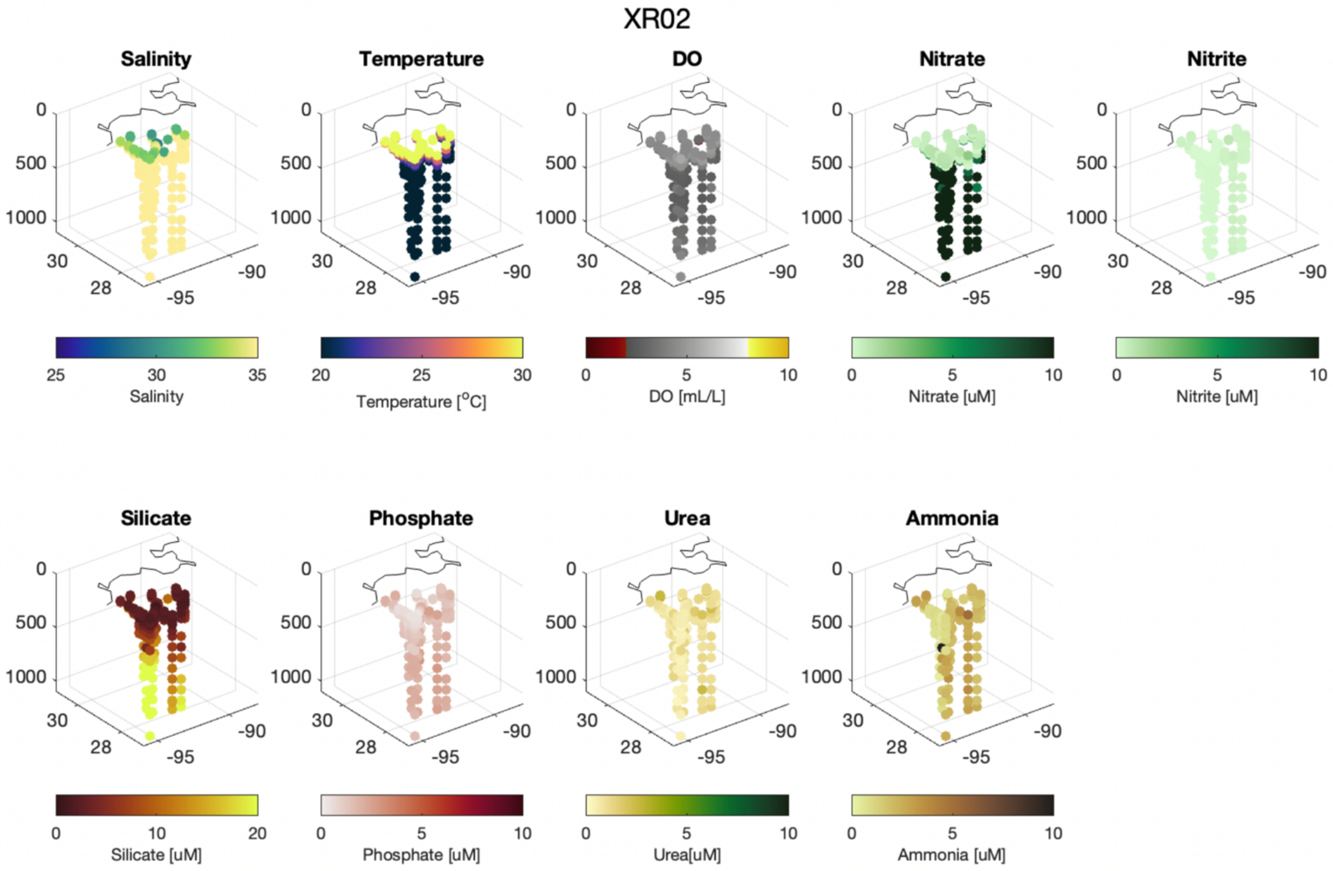
APPENDIX



A.1. Environmental Variable plots of the Research Experience for Undergraduates Cruises (2016, 2017 & 2018).



A.2. Environmental Variable plots of XR01.



A.3. Environmental Variable plots of XR02.

Variable	Positive Mode 1 Means	Negative Mode 1 Means
Salinity	33.4	31.3
Temperature (ºC)	25.2	26.8
DO [mL/L]	2.5	4.6
Nitrate [µmol/L]	4.5	0.8
Nitrite [µmol/L]	3.8	0.5
Total Nitrogen [µmol/L]	8.3	1.3
Phosphate [µmol/L]	24.7	5.8
Silicate [µmol/L]	0.9	0.3
Ammonia [µmol/L]	0.8	0.5
Urea [µmol/L]	1.2	0.5

Table A.4. Principal Component Positive and Negative Mode 1 Mean values.

Variable	Positive Mode 2 Means	Negative Mode 2 Means
Salinity	33.7	29.6
Temperature (ºC)	27.3	24.5
DO [mL/L]	3.2	4.7
Nitrate [µmol/L]	1.5	3.4
Nitrite [µmol/L]	1.7	2.0
Total Nitrogen [µmol/L]	3.2	5.3
Phosphate [µmol/L]	15.3	10.2
Silicate [µmol/L]	0.6	0.4
Ammonia [µmol/L]	0.7	0.6
Urea [µmol/L]	0.9	0.6

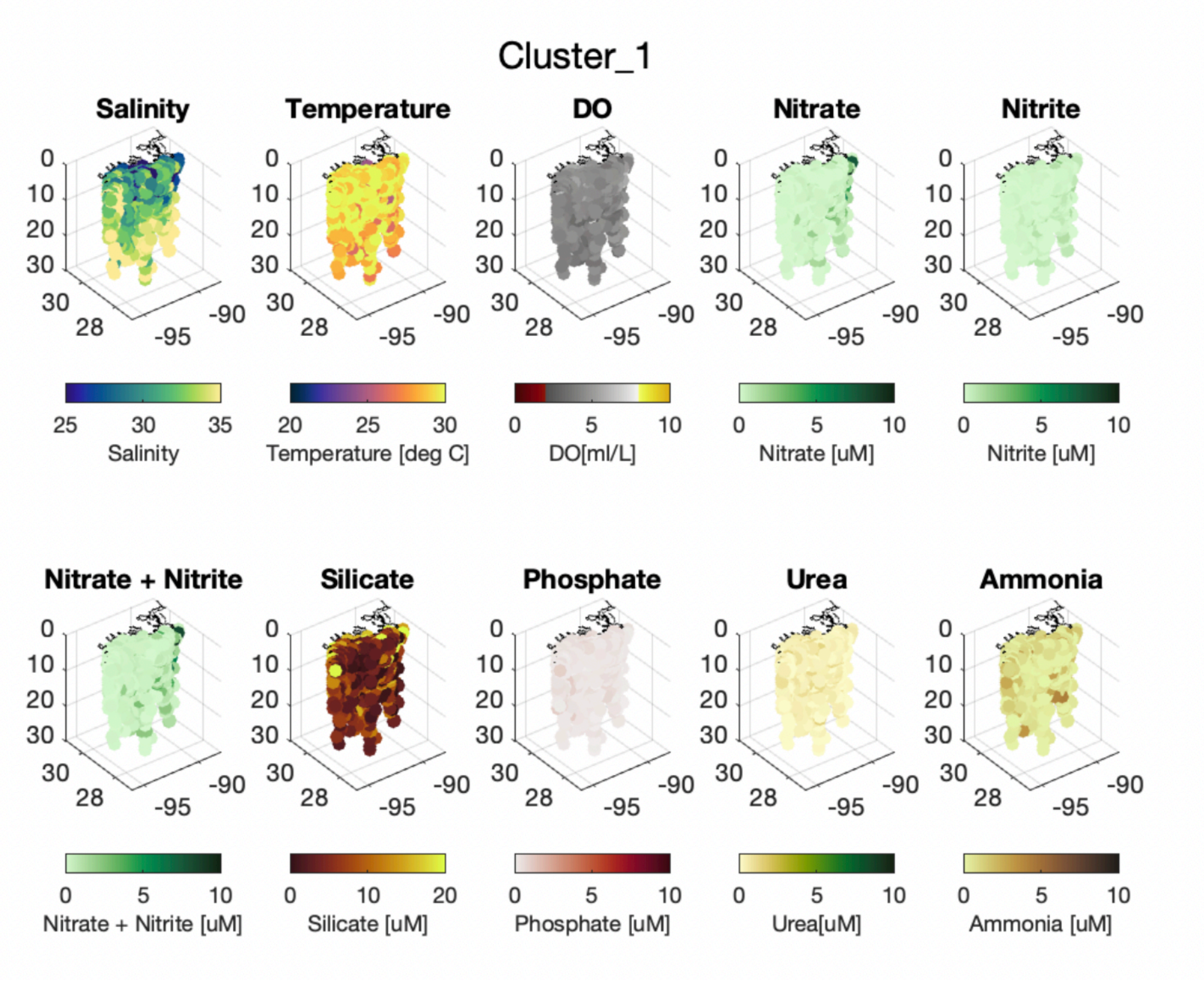
Table A.5. Principal Component Positive and Negative Mode 2 Mean values.

Variable	Positive Mode 3 Means	Negative Mode 3 Means
Salinity	30.0	34.2
Temperature (ºC)	28.1	24.4
DO [mL/L]	4.0	3.6
Nitrate [µmol/L]	1.3	3.1
Nitrite [µmol/L]	2.1	1.5
Total Nitrogen [µmol/L]	3.4	4.6
Phosphate [µmol/L]	0.5	0.5
Silicate [µmol/L]	12.6	14.0
Ammonia [µmol/L]	1.2	0.4
Urea [µmol/L]	0.9	0.4

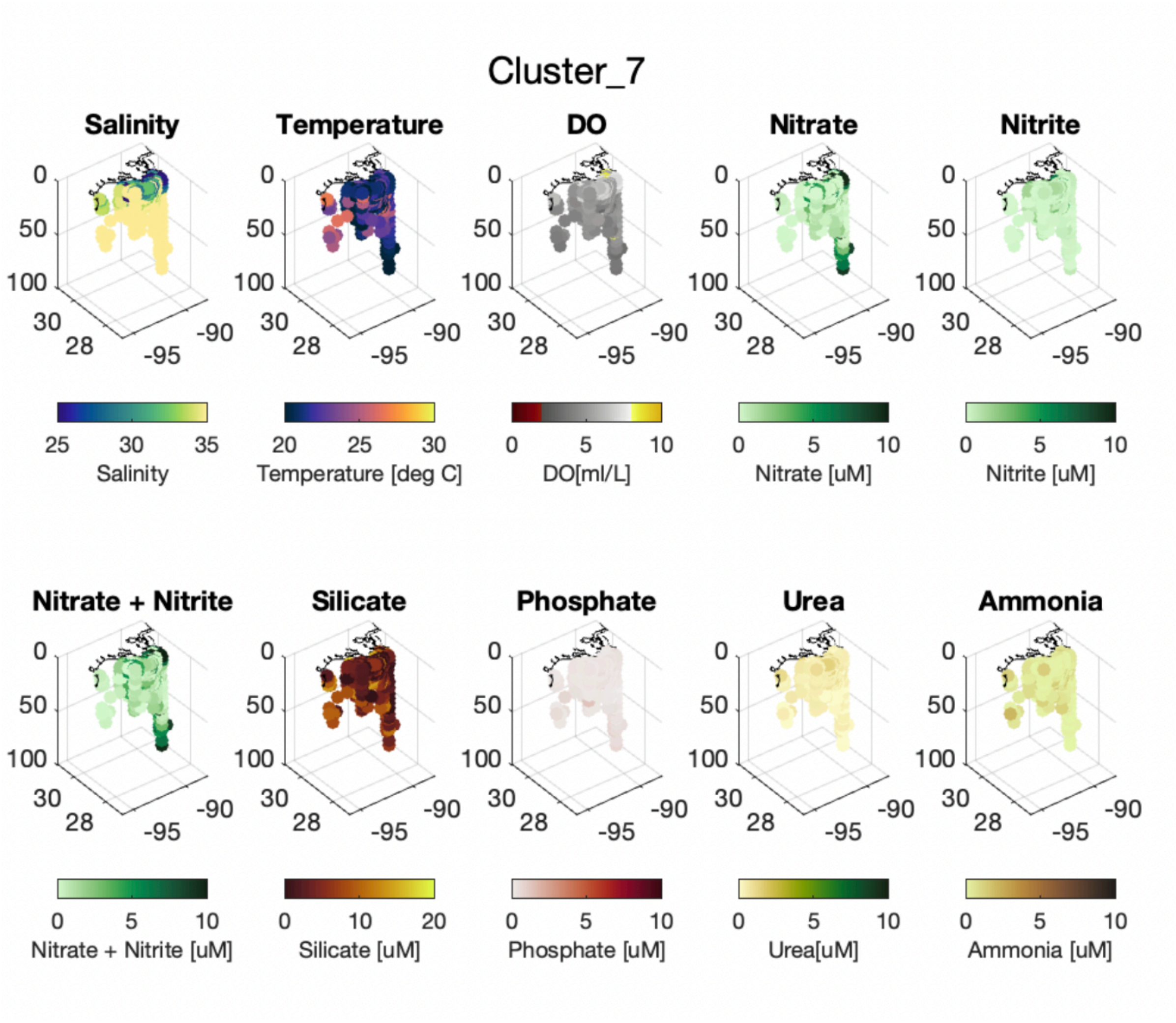
Table A.6. Principal Component Positive and Negative Mode 3 Mean values.

Variable	Positive Mode 4 Means	Negative Mode 4 Means
Salinity	33.4	31.1
Temperature (ºC)	23.5	28.3
DO [mL/L]	3.9	3.7
Nitrate [µmol/L]	1.3	3.0
Nitrite [µmol/L]	3.1	0.8
Total Nitrogen [µmol/L]	4.3	3.8
Phosphate [µmol/L]	0.4	0.6
Silicate [µmol/L]	11.8	14.4
Ammonia [µmol/L]	1.0	0.4
Urea [µmol/L]	0.7	0.9

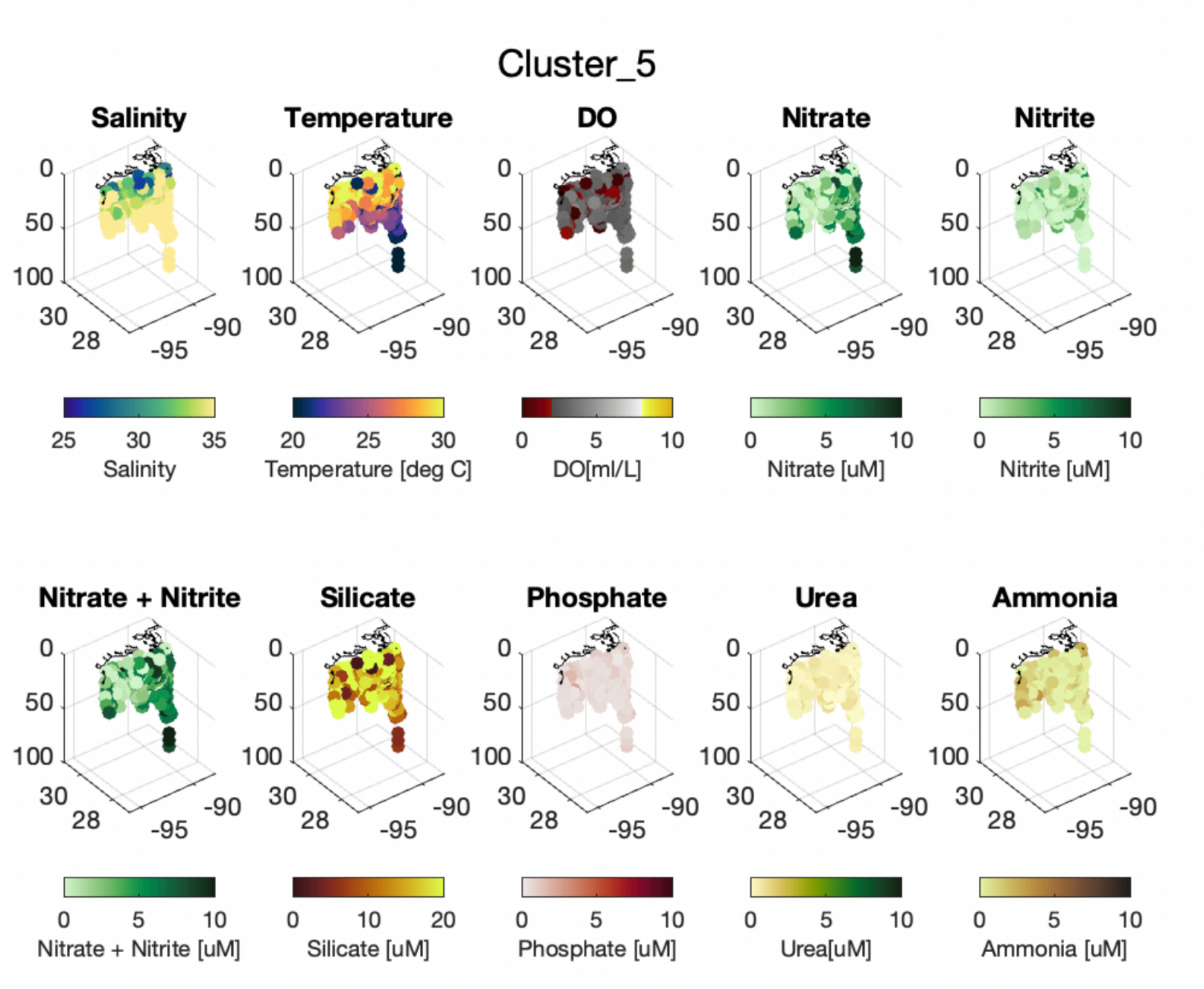
Table A.7. Principal Component Positive and Negative Mode 4 Means.



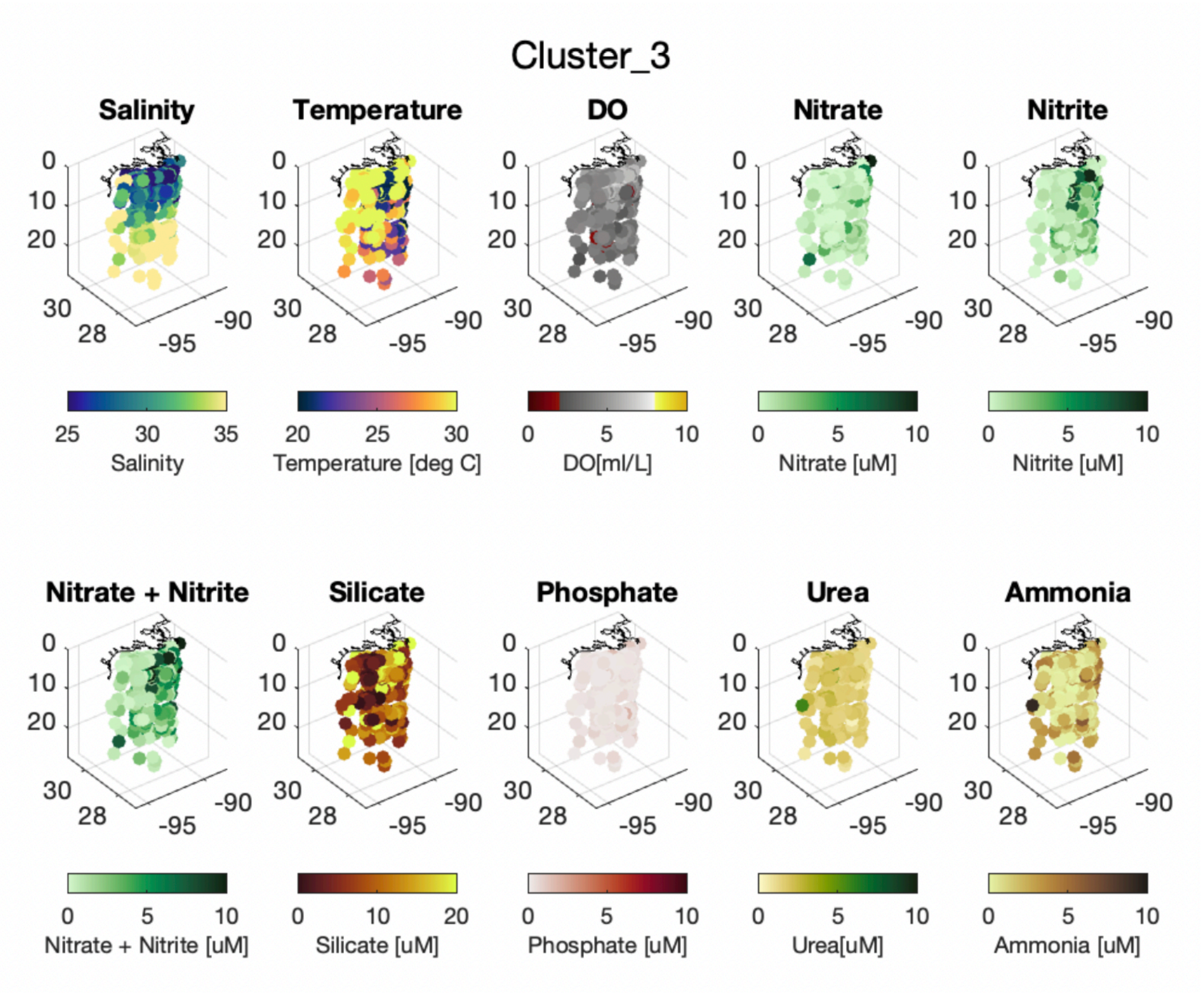
A.8 Environmental Variable plots of Cluster 1.



A.9 Environmental Variable plots of Cluster 7.



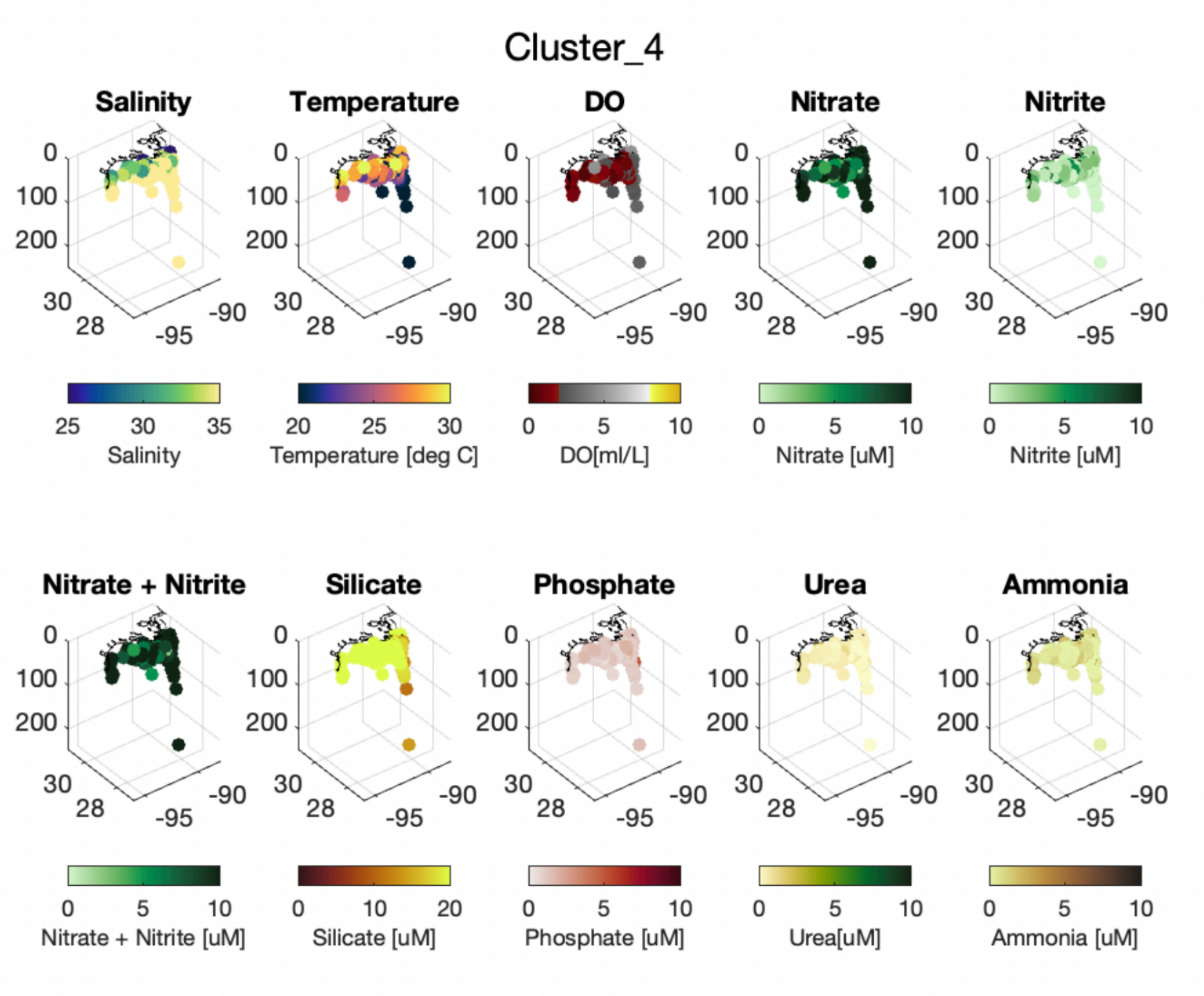
A.10. Environmental Variable plots of Cluster 5.



A.10. Environmental Variable plots of Cluster 3.

Variable	Cluster 3 Means
Salinity	30.9
Temperature (ºC)	25.6
DO [mL/L]	4.2
Nitrate [µmol/L]	0.9
Nitrite [µmol/L]	2.4
Total Nitrogen [µmol/L]	3.2
Phosphate [µmol/L]	11.8
Silicate [µmol/L]	0.3
Ammonia [µmol/L]	1.7
Urea [µmol/L]	1.7

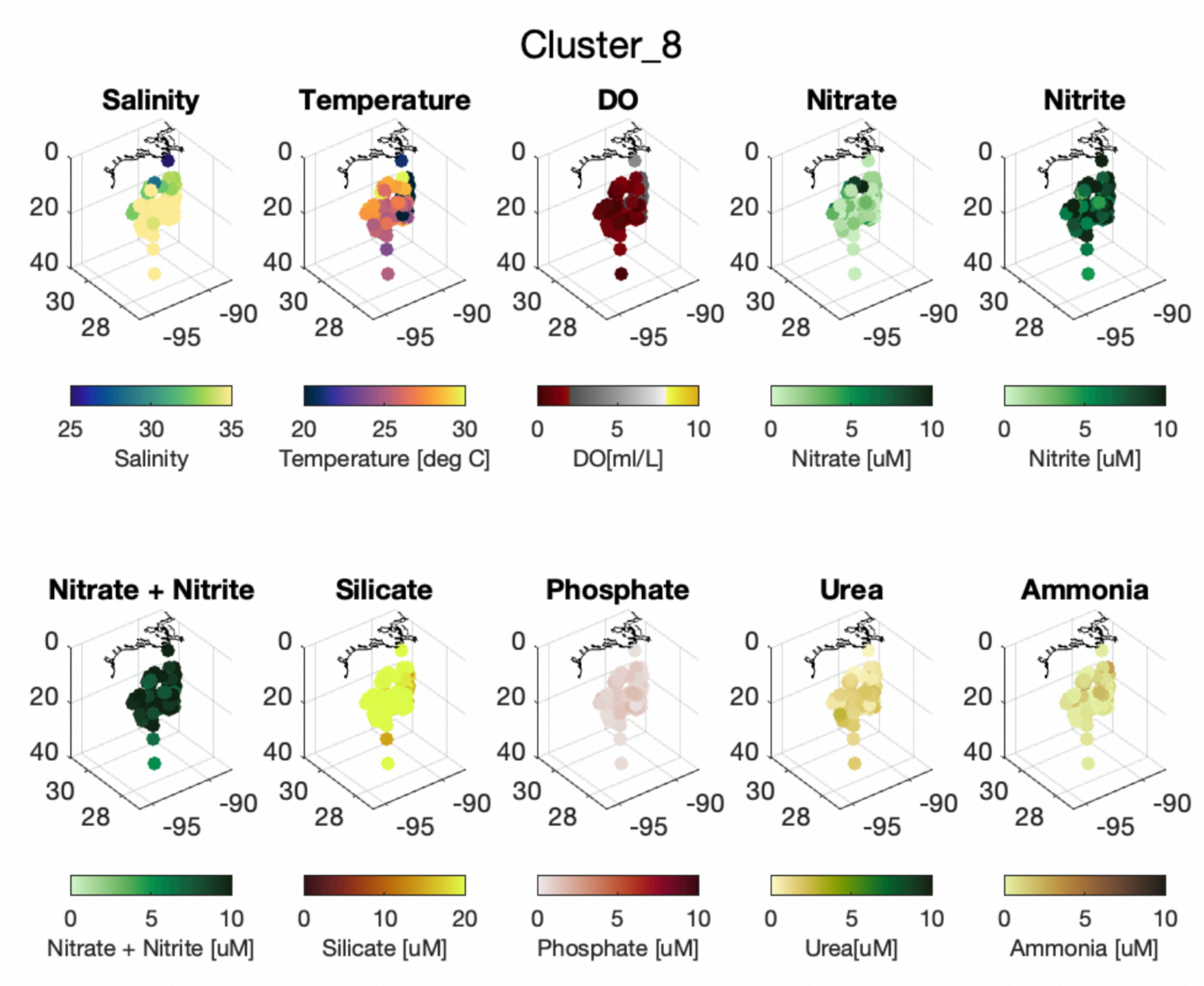
Table A.11. Environmental Variable Means of Cluster 3.



A.12. Environmental Variable plots of Cluster 4.

Variable	Cluster 4 Means
Salinity	34.7
Temperature (ºC)	25.1
DO [mL/L]	1.3
Nitrate [µmol/L]	9.1
Nitrite [µmol/L]	1.9
Total Nitrogen [µmol/L]	10.9
Phosphate [µmol/L]	37.0
Silicate [µmol/L]	1.3
Ammonia [µmol/L]	0.5
Urea [µmol/L]	0.6

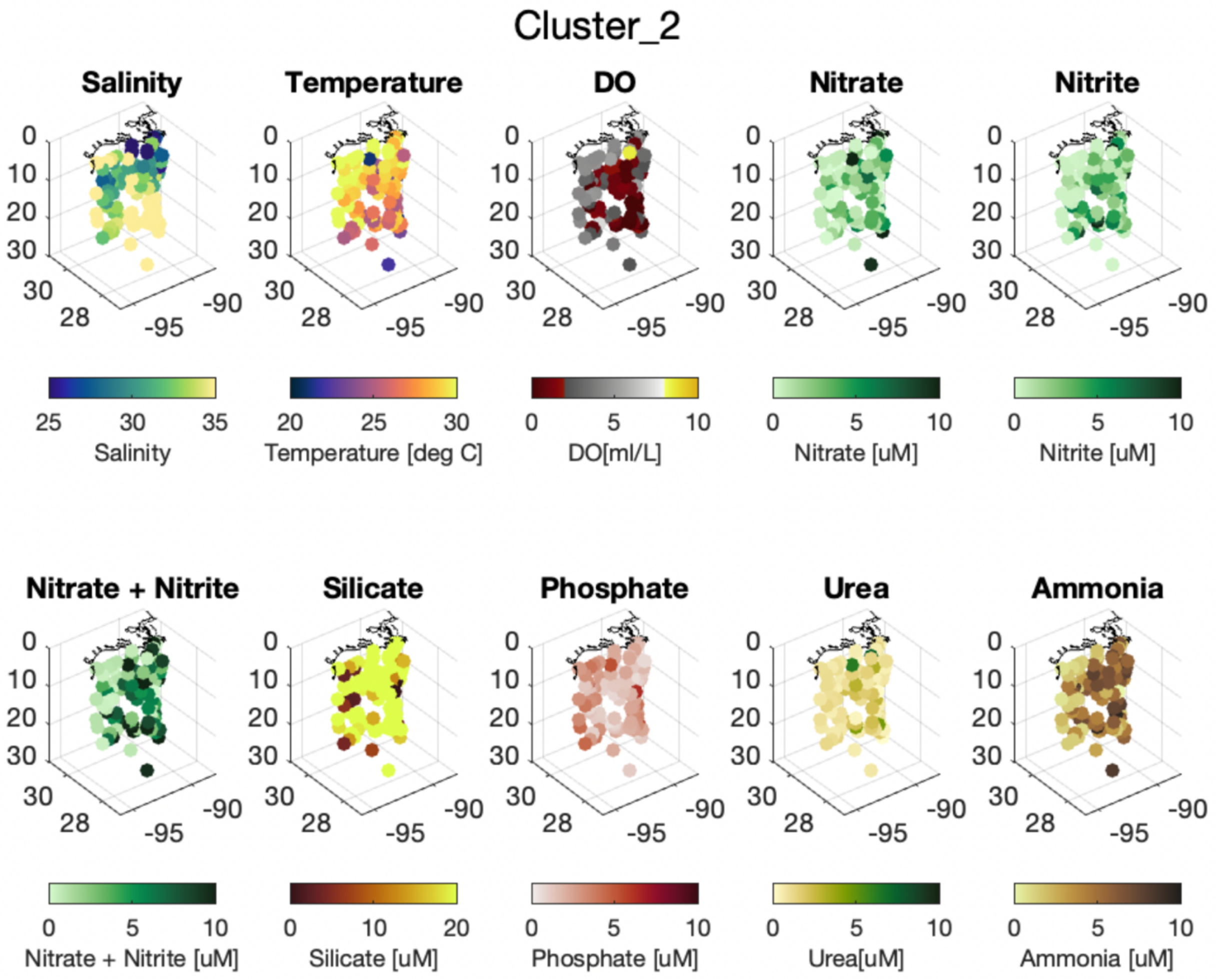
Table A.13. Environmental Variable Means of Cluster 4.



A.14. Environmental Variable plots of Cluster 8.

Variable	Cluster 8 Means
Salinity	35.3
Temperature (ºC)	25.4
DO [mL/L]	1.0
Nitrate [µmol/L]	1.3
Nitrite [µmol/L]	10.3
Total Nitrogen [µmol/L]	11.5
Phosphate [µmol/L]	26.5
Silicate [µmol/L]	0.9
Ammonia [µmol/L]	1.3
Urea [µmol/L]	0.7

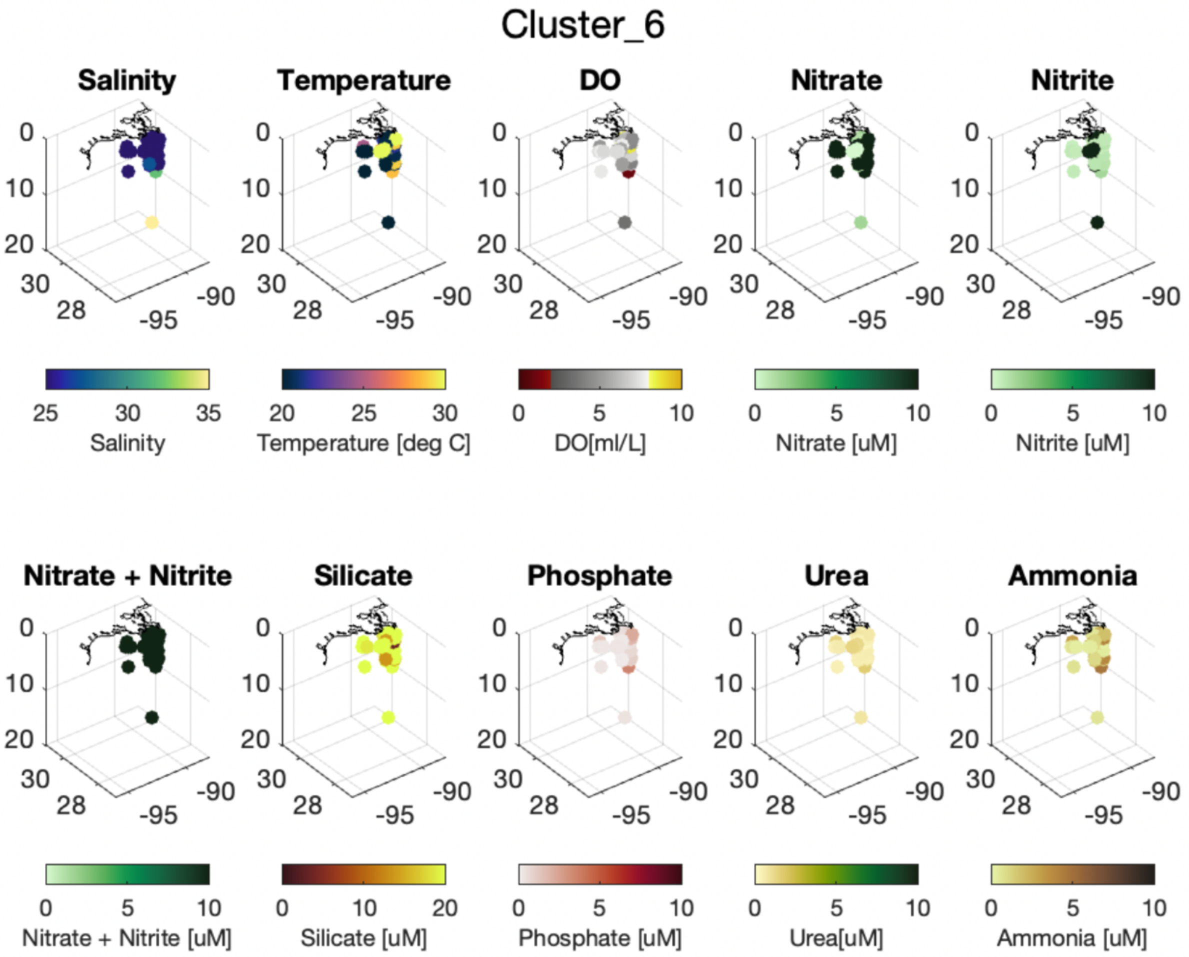
Table A.15. Environmental Variable Means of Cluster 8.



A.16. Environmental Variable plots of Cluster 2.

Variable	Cluster 2 Means
Salinity	32.1
Temperature (ºC)	27.9
DO [mL/L]	2.0
Nitrate [µmol/L]	2.7
Nitrite [µmol/L]	2.5
Total Nitrogen [µmol/L]	5.2
Phosphate [µmol/L]	34.2
Silicate [µmol/L]	2.1
Ammonia [µmol/L]	1.1
Urea [µmol/L]	4.6

Table A.17. Environmental Variable Means of Cluster 2.



A.18. Environmental Variable plots of Cluster 6.

Variable	Cluster 6 Means
Salinity	19.9
Temperature (ºC)	22.8
DO [mL/L]	6.5
Nitrate [µmol/L]	16.6
Nitrite [µmol/L]	11.7
Total Nitrogen [µmol/L]	28.3
Phosphate [µmol/L]	28.3
Silicate [µmol/L]	0.4
Ammonia [µmol/L]	0.6
Urea [µmol/L]	1.5

Table A.19. Environmental Variable Means of Cluster 6.