UNDERSTANDING AND FORECASTING HYPOXIA USING MACHINE LEARNING ALGORITHMS

BY

EVAN JOSEPH COOPERSMITH

B.S.E., Princeton University, 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Environmental Engineering in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign 2008

Urbana, Illinois

Adviser:

Professor Barbara Minsker

Abstract

Among the various threats facing environmentally stressed bodies of water is hypoxia, a condition during which dissolved oxygen levels fall to low levels and endanger flora and fauna. To better understand hypoxia, researchers collect data from various locations during hypoxic events. However, these events can be sporadic and short-lived, making observation difficult. For this reason, this study's primary objective lies in forecasting where and when hypoxia may transpire in the hopes of observing its effects in real time, focusing on a case study in Corpus Christi Bay (Texas). Dissolved oxygen levels in this bay can be characterized by three temporal trends. A long-term decline in oxygen levels has been noted over the past several decades, an annual oscillation is a perennial feature, and daily periodicity is also apparent. To predict hypoxic events, these three mathematical trends are isolated and extracted to obtain unbiased forecasts using a sequential normalization approach. Once these data points are rendered time-independent (all three functions of time removed), machine learning algorithms are constructed employing the continuous, normalized values from a variety of sensor locations. By including latitude and longitudinal coordinates as additional variables, a spatial depiction of hypoxic conditions can be illustrated effectively, allowing for more efficient summer data collection and more accurate, real-time projections. Using k-nearest neighbor algorithms, approximate probabilities of observing hypoxia the following day were calculated, and estimates of dissolved oxygen levels were also computed. During periods in which hypoxia was observed, forecast probabilities of hypoxia had exceeded 80%. Conversely, during periods in which no hypoxia was observed, the model's estimate remained below 20% for days on end.

To Mom and Dad

Acknowledgements

Firstly, I would like to thank my advisor, professor Minsker for allowing a student with an undergraduate background which was quite distant from environmental engineering to expand his horizons geographically and intellectually. It has been a hugely beneficial, if sometimes harrowing experience.

Secondly, I would like to thank Andrew Collier, whose assistance in programming and more specifically, the spatial maps is seen throughout this thesis. He has been valuable as a colleague and irreplaceable as a friend.

Thirdly, I would like to that my parents for enduring the kvetching from my wanderlust-induced regret.

Finally, I would like to thank Rebecca – for after all, it is obvious why this past year has been so vastly different from the first.

Table of Contents

List of Figures

1. Introduction

Hypoxia is an estuarine condition in which aquatic dissolved oxygen levels drop to levels below 2 mg/l (Dauer et al., 1992). Generally, research has found hypoxia resulting from anthropogenic eutrophication (Campbell and J.G. Goodman, 2007), biological oxygen demand (Mallin et al, 2006), and more specifically, the leaching of nitrates from fertilizers into local groundwater (Booth and C. Campbell, 2007). Hypoxia has been observed in watersheds from the great lakes (Loewen et al, 2007), to European lakes (Galkovskaya and Minyanina, 2005), to the Corpus Christi (Texas) case study examined herein. Corpus Christi Bay's hypoxia, like that observed in other similar bodies of water, results from a seasonal periodicity of dissolved oxygen fluctuations (Ritter and Montagna, 1999), a diurnal cycle resulting from tides (Hagy and Murrell, 2007) or photosynthetic timing as nighttime oxygen consumption by fauna coincides with an absence of oxygen creation by flora (Goldshmid et al, 2004), and salinity induced stratification (Hodges and Furnans, 2007).

Hypoxic conditions were first observed in Corpus Christi Bay in 1988 (Montagna and Kalke, 1992) and each year thereafter during the past twenty summers. In particular, the bay's southeastern region has displayed incidents during which dissolved oxygen fell below the 2 mg/L threshold on numerous occasions. These events threaten benthic life, reducing diversity and biomass and often causing the relocation of sensitive benthic organisms to surface waters (Ritter and Montagna, 2006). Furthermore, the effects of hypoxic conditions in Corpus Christi Bay bear economic ramifications for proximal agriculture in the Gulf of Mexico's watershed (Donnelly and Scavia, 2007). With nitrate loading as a potential cause of hypoxia, such events can cause restrictions to be levied

upon nearby farmers. In fact, hypoxia events have been sufficiently severe as to warrant the Harmful Algal Bloom and Hypoxia Research and Control Act (HABHRCA) of 1998, a Federal mandate which stipulates the appropriation of funding for education, research, monitoring, reduction, prevention, and control of harmful hypoxic events. The act was subsequently reauthorized six years later.

Hypoxia's occurrence in Corpus Christi Bay and elsewhere is frequently limited to the lowest water layer, the hypolimnion (Galkovskaya and Minyanina, 2005). In Corpus Christi Bay, the onset of hypoxia is frequently coincident with the entrance of saline water from the Gulf of Mexico into the fresh water bay (Hodges and Furnans, 2007). This is in contrast to the traditional mechanistic hypoxia modeling that describes hypoxic events through the prism of nitrogen loading and fresh water fluxes from rivers (Hetland and DiMarco, 2008). Such stratification-driven hypoxic conditions have also been observed in lakes where flood water is diverted via spillways (Brammer et al, 2007).

This study seeks to better understand and predict these phenomena through two primary objectives. The first objective is to identify which factors are statistically correlated with the generation of hypoxic conditions. The second objective is to determine the feasibility of real-time hypoxia forecasting using data mining and machine learning approaches. Machine learning algorithms such as k-nearest neighbor have been proposed to solve environmental problems from soil water retention (Nemes et al, 2008), to arsenic in well-water (Meliker et al, 2008), to forest mapping (McRoberts et al, 2007). Data mining techniques and relevant machine learning algorithms are explored for creating twenty-four hour forecasts that would provide researchers with sufficient warning to launch hypoxia field campaigns on days that are likely to have hypoxia

conditions. This study will discuss the characteristics of the bay (Section 2), the methodology used to analyze the data (Section 3), the most important results of that analysis (Section 4), and finally, the conclusions drawn from those results (Section 5).

2. Corpus Christi Bay – Case Study

Corpus Christi Bay is located along the southeastern coast of Texas, just west of the Gulf of Mexico with a dividing barrier island, approximately 140 miles SSE of San Antonio as shown in Figure 2.1 (Google Earth).

Figure 2.1 – Corpus Christi Bay, Gulf of Mexico Adjacent

The bay's bathymetry is characterized by an extremely flat floor at an average depth of 3.6m (Ward, 1997). The bay itself represents an urban estuary, home to not only the city of Corpus Christi with a population of approximately 280,000 but also the Port of Corpus Christi, which serves as this country's seventh largest port. Additionally, adjacent to the bay are a number of petrochemical plants, augmenting the complexities with regards to water quality modeling. The relatively low tidal range within the Gulf of Mexico, combined with the restricted channel entrances, causes water circulation to be defined more by meteorological phenomena like wind, rather than the traditional mechanism of tides (Kulis and Hodges, 2006).

Several factors are hypothesized as plausible means by which dissolved oxygen levels in Corpus Christi may fluctuate. As suggested previously, saline water can cause stratification, leading to hypolimnetic hypoxia as the water column fails to become well-mixed by air and water currents. Hydrodynamic analysis suggests that water masses moving northward from Laguna Madre may provide the mechanism by which saline water enters the bay and generates hypoxic conditions (Hodges and Furnans 2007). Coincident with the salinity currents are the wastewater treatment plants on the Oso River (shown in purple in Figure 2.1), which discharge to Oso Bay and from there to Corpus Christi Bay, providing a potential anthropogenic source of outside contaminants and the aforementioned agricultural runoff that can lead to nitrate-based eutrophication.

Within Corpus Christi Bay lies a spatial grid of oxygen sensors that provides oxygen levels over the last five to ten years. These sensors have taken readings at over fifty locations (though not always concurrently), each affected by a diversity of hydrodynamic and water quality factors. Furthermore, at each of these sensors, readings may be gathered at depths ranging from floor to surface of this rather shallow body of water. As earlier research has demonstrated (and empirical evidence from Corpus Christi Bay has confirmed), the hypoxic events are observed predominantly in the benthic zone (Osterman, Poore, and Swarzenski, 2008). Moreover, in Corpus Christi Bay and test beds in the previously referenced paper, when hypoxia is observed at higher levels of the water column, the bottom layer of that water column will be hypoxic as well. For this reason, the analysis in this work will focus solely upon this benthic layer.

Historical data from the sensors in Corpus Christi Bay (Montagna & Ritter, 2006) are collected in two forms. The first is that of ‘grab’ samples, with the majority of the

records falling into this category. These grab samples are gathered by researchers transporting their sensors by boat from location to location throughout the bay and reporting back a suite of individual values at a great variety of locations, depths, and time intervals. In addition to these grab samples are continuous samples. In these cases, a sensor is deployed at a location and remains at that location and depth for a period of one to two weeks, during which it logs readings every fifteen minutes. Figure 2.2 shows the sampling locations in Corpus Christi Bay, with grab sample locations given in black and continuous sensor locations given in white.

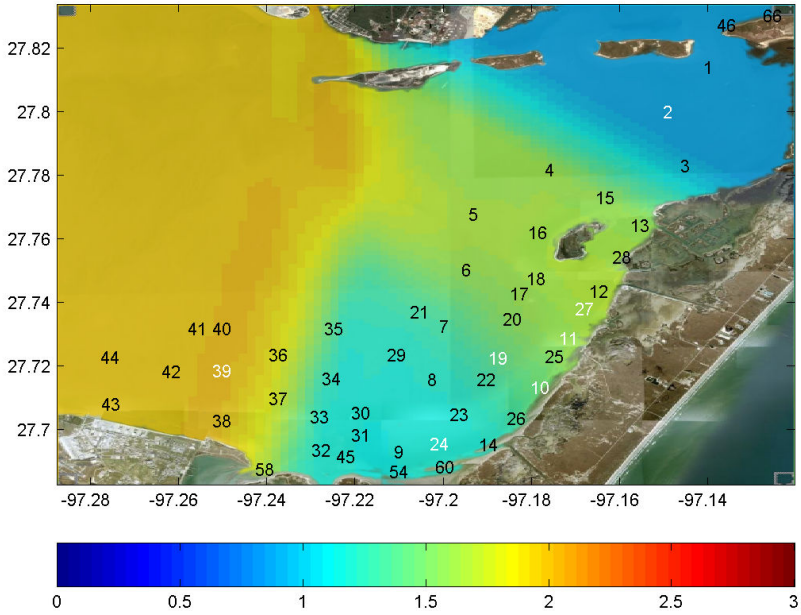


Figure 2.2 – Corpus Christi Bay, Sensor Placements

Although there are fewer locations for which continuous information is accessible, for the purposes of data mining and subsequent forecasting with machine learning, these data provide a critical record of the temporal dynamics needed for 24-hour ahead forecasting. The continuous readings are performed at either the water's surface or within inches of the bay floor. Among the locations for which bottom readings are taken,

one (sensor #2 in Figure 2.2) is used primarily as a means of numerical control since that location virtually never experiences a hypoxic event.

3. Methodology

Data mining is a process for extracting knowledge from data, with the steps illustrated in figure 3.1. First, data are gathered and the fraction of that information which applies to the time range in question is isolated. Next, those data are transformed into the necessary form to facilitate the model's interpretation. Iterative algorithms are performed on those data, and when this is finished, the output is evaluated. At this point, the insights from mathematical analysis become useful knowledge.

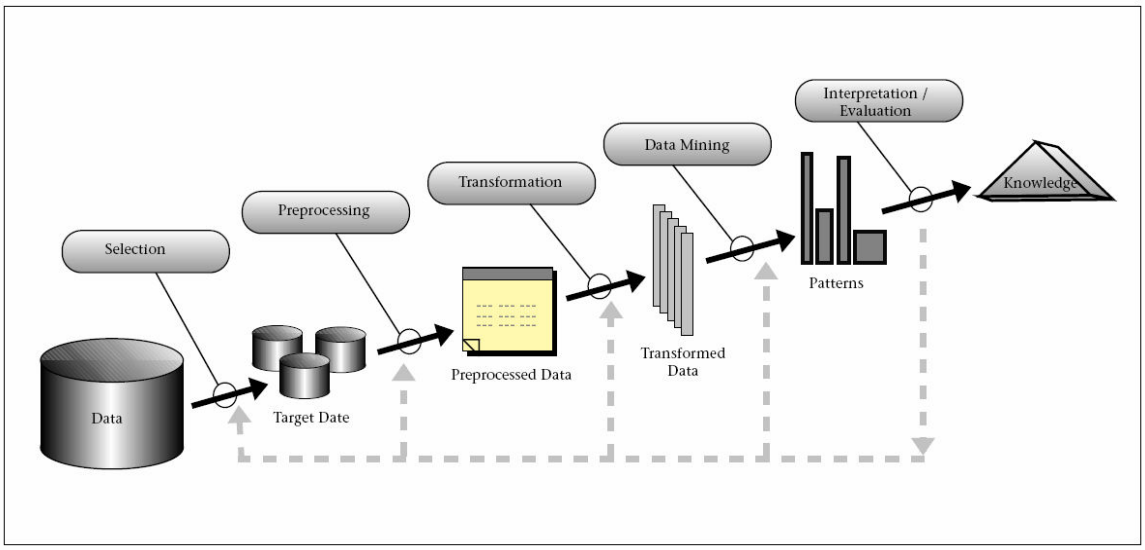


Figure 3.1 – The Process of Data Mining (Fayyad et al, 1996)

For the purposes of studying the causes of hypoxia, this data mining approach will require five steps. Before these five steps can take place, target data must be extracted (selection step in Figure 3.1). The dataset contains samples at various depths. The target data are bottom data, at continuous sensors only. For preprocessing and transformation, three separate temporal trends with regards to dissolved oxygen will need to be removed before meaningful comparisons can be made and defensible analysis can be performed. This leaves transformed data (Figure 3.1) with which to continue the analyses. Next, the data mining step is performed, having been assured that the remaining dissolved oxygen

patterns are *not* time-dependent. Continuing, additional variables that will be beneficial in generating the hypoxia forecasts are determined, the machine learning models which could potentially serve the desired objective of real-time predictions are assessed the pros and cons of each are weighed, and having settled upon a model with which to move forward, a non-parametric regression is performed. Fourth, once a model is chosen, its accuracy must be validated via hind-casting (a process by which previous events are predicted using only information known at that particular period in time) with a sliding window at a specific location. Fifth, given that the results are satisfactory, a means to extend the single-location forecasts into a spatial projection and address the spatial interpolation needs therein must be devised. This is the 'knowledge' segment of figure 3.1. The steps in this process are detailed in the following four sub-sections.

3.1 Removing Time-Dependent Trends

This section addresses the process of removing time dependent trends. First, over twenty-five years of historical dissolved oxygen readings taken in Corpus Christi Bay are examined, shown in Figure 3.2 A cursory examination reveals that not only is there a slow, yet persistent decline in dissolved oxygen levels, but also a periodic, annual oscillation superimposed upon that gradual degeneration. Further, it is accepted that oxygen data exhibit a diurnal periodicity, which is superimposed upon the aforementioned two cycles.

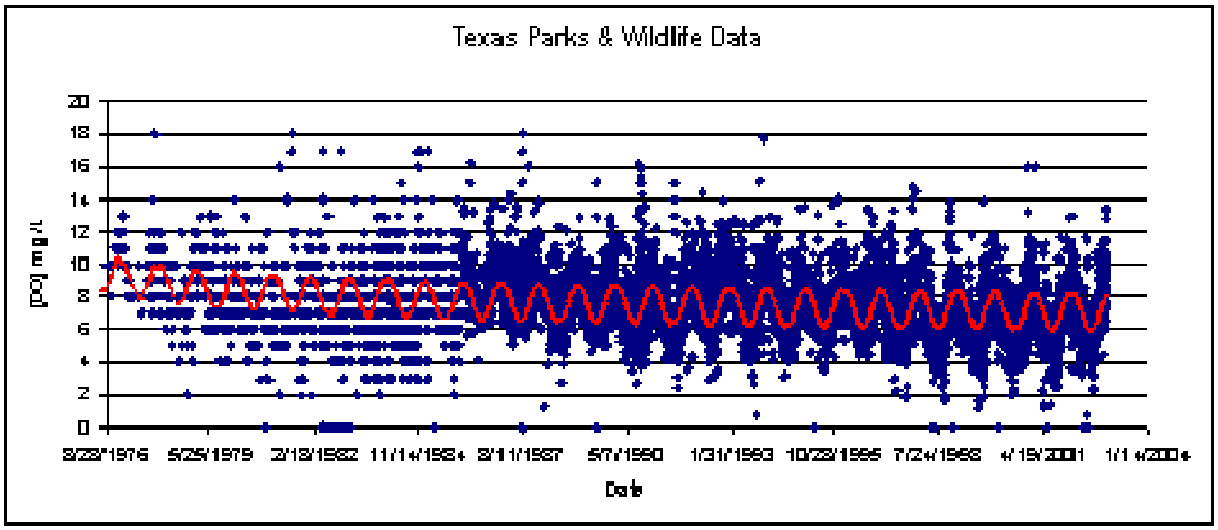


Figure 3.2 – Long Term Oxygen Data, 25 Years (Texas Parks & Wildlife)

Detrending such a dataset is accomplished via sequential normalization (Maidment and Parzen, 1984). This procedure calls for the elimination of the longest-periodicity trend first, and then to advance to smaller periodicities once the larger ones have been removed.

First, a basic linear description of the long-term decrease in dissolved oxygen levels is proposed:

$$DO_{long-tern}(t) = k_1 t + k_2 \qquad \text{(Equation 3.1.1)}$$

To remove this trend, that dataset is normalized. Essentially, this long-term decrease is smoothed while all other sources of variance remain. This will leave behind cyclical functions that are *not* obscured by the long-term trend. In achieving this, each point in the original dataset is normalized as follows:

$$DO_{norm,t} = \frac{DO(t)}{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1}\right)} = \frac{k_1 t + k_2}{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1}\right)} \qquad \text{(Equation 3.1.2)}$$

Having done this normalization, any array of records used may be adjusted. Adjusted DO' is calculated as follows:

$$DO_{t,adjusted} = \frac{DO_t}{DO_{norm,t}} = DO_t \frac{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1} \right)}{k_1 t + k_2} \qquad \text{(Equation 3.1.3)}$$

At this stage, the long-term trend is gone, and by constructing a discrete Fourier transform with two harmonics, capturing the seasonal periodicity within this adjusted data is now possible:

$$DO_{seasonal}(t) = k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_7 \qquad \text{(Equation 3.1.4)}$$

(t is measured in units of days)

One more time, an analogous normalization is performed to generate yet another adjusted dataset, which lacks both a long-term trend and an annual oscillation:

$$DO_{twice\,norm,t} = DO_{norm,t} \frac{DO_{seasonal}(t)}{\left(\frac{\sum_{i=0}^{n_2} DO_i}{n_2} \right)} = \frac{k_1 t + k_2}{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1} \right)} \frac{k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_7}{\left(\frac{\sum_{i=0}^{n_2} DO_i}{n_2} \right)}$$

(Equation 3.1.5)

Finally, to remove the diurnal undulation, the following transformation is performed on the twice-adjusted data (i.e., the data with the long-term and seasonal trends removed):

$$DO_{t,twice-adjusted} = DO_t \frac{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1} \right)}{k_1 t + k_2} \frac{\left(\frac{\sum_{i=0}^{n_2} DO_i}{n_2} \right)}{k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_7} \qquad \text{(Equation 3.1.6)}$$

From here, the diurnal relationship is calculated and a second discrete Fourier transform with two harmonics is performed:

$$DO_{diurnal}(t) = k_8 \sin\left(\frac{2\pi t}{24} - k_9\right) + k_{10} \sin\left(\frac{2\pi t}{12} - k_{11}\right) + k_{12} \qquad \text{(Equation 3.1.7)}$$

(t is now measured in units of hours)

Continuing:

$$DO_{thrice\,norm,t} = DO_{twice\,norm,t} \frac{DO_{diurnal}(t)}{\left(\frac{\sum_{i=0}^{n_3} DO_i}{n_3}\right)} = \frac{k_1 t + k_2}{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1}\right)} \frac{k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_7}{\left(\frac{\sum_{i=0}^{n_2} DO_i}{n_2}\right)} \frac{k_8 \sin\left(\frac{2\pi t}{24} - k_9\right) + k_{10} \sin\left(\frac{2\pi t}{12} - k_{11}\right) + k_{12}}{\left(\frac{\sum_{i=0}^{n_3} DO_i}{n_3}\right)}$$

(Equation 3.1.8)

At this point, normalization has removed three time- dependencies. Consequently, *any* value from *any* historical record can be compared to any other by adjusting thrice, as follows:

$$DO_{thrice-adjusted,t} = DO_t \frac{\left(\frac{\sum_{i=0}^{n_1} DO_i}{n_1}\right)}{[k_1 t + k_2]} \frac{\left(\frac{\sum_{i=0}^{n_2} DO_i}{n_2}\right)}{\left[k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_7\right]} \frac{\left(\frac{\sum_{i=0}^{n_3} DO_i}{n_3}\right)}{\left[k_8 \sin\left(\frac{2\pi t}{1} - k_9\right) + k_{10} \sin\left(\frac{2\pi t}{0.5} - k_{11}\right) + k_{12}\right]}$$

(Equation 3.1.9)

Alternatively, a function that possesses all three time cycles can be defined:

$$DO(t) = k_1 t + k_3 \sin\left(\frac{2\pi t}{365} - k_4\right) + k_5 \sin\left(\frac{2\pi t}{182.5} - k_6\right) + k_8 \sin\left(\frac{2\pi t}{1} - k_9\right) + k_{10} \sin\left(\frac{2\pi t}{365} - k_{11}\right) + c$$

(In this case, t is measured in days and $c = k_2 + k_7 + k_{12}$)

(Equation 3.1.10)

Please note that this normalization could theoretically be performed using arithmetic differences in lieu of the ratio-based approach by using the following format:

$$DO_{norm,t} = DO(t) - \frac{\sum_{i=0}^{n_1} DO_i}{n_1} \Rightarrow DO_{adjusted,t} = DO_t - DO_{norm,t} \qquad \text{(Equation 3.1.11)}$$

Admittedly, there is merit in this construction since it allows the amount of adjustment to be independent of the value of the function at any given point. However, the lack of flexibility when additional exogenous factors are introduced outweighs the benefit. The present mode can receive DO_t, and simply multiply by a series of conversions:

$DO(t) = DO_t$ * LongTermFactor * AnnualFactor * DiurnalFactor…

Any subsequent conversion could easily be introduced.

3.2 Isolating Parameters That Influence DO / Model Testing

The previous section has elucidated a method for normalizing temporal periodicity, allowing any record of dissolved oxygen to be compared to any other. Once the trends are removed, the resulting stationary data are then used to identify which parameters most influence dissolved oxygen levels and to fit machine learning algorithms to the data to forecast hypoxia based on those parameters. For this purpose, two distinct, nonparametric machine learning algorithms are examined. Nonparametric statistics, because they do not require assumptions regarding the distributions and correlations of the independent variables, offer us the opportunity to generate a rather robust model even in the absence of concrete proof of variable independence and distribution shapes.

The first model examined is the K-Nearest Neighbor algorithm, abbreviated 'KNN.' KNN remains among the most straightforward methods for classification, analysis, and forecasting. (Kumar et al, 2006, p394). Like many non-parametric techniques, the focus is on classification – division of a set of data into groups which

display or do not display specific properties. In this case, with regards to hypoxia, it would be appropriate to utilize KNN to classify data as a group of points which are likely to become hypoxic in twenty-four hours and a group of points which are not. KNN's approach aims to classify any record by locating the nearest records via some previously defined metric for distance. In this instance, a simple Euclidean distance function is applied (Kumar et al, 2006, p394):

$$d = \sqrt{\sum_i (x_i - y_i)^2}, \forall_i \qquad \text{Equation (3.2.1)}$$

Therefore, given an input vector consisting of a variety of variables relevant to the prediction of hypoxia, the following is applied to calculate the distance function and determine the similarity between one data point and another, historical value (Kumar et al, 2006, p395):

$$d\left(x_i, x_{i'}\right) = \sqrt{\sum_i \left(x_{ij} - x_{i'j}\right)^2}, \forall_{i,i'}, j \qquad \text{Equation (3.2.2)}$$

For the case of hypoxia, this equation might appear as below:

$$\min_i \left\{ \left[\frac{(DO_{norm} - DO_i)}{\sigma_{DO_{norm}}} \right]^2 + \left[\frac{(Sal - Sal_i)}{\sigma_{Sal}} \right]^2 + \left[\frac{(Temp - Temp_i)}{\sigma_{Temp}} \right]^2 \right\}, \forall_i \quad Given\ input\ .(DO_{norm}, Sal, Temp)$$

(Equation 3.2.3)

Then the *k* best matches are located (with *k* to be chosen prior to the model's execution) and exploit these 'similar' records as our the set from which to generate a plausible distribution of outcomes. For each past record, it is observed what happened to the 'similar set' in one day's time and those results are used as an empirical prediction. Therefore it is important for the training set to contain not only the historical value of

dissolved oxygen, but the value at that same location twenty-four hours later. This requires removing any data points within twenty-four hours of the end of a training set, when one would know the historical dissolved oxygen reading, but *not* the value twenty-four hours hence.

In addition to previous normalized values of dissolved oxygen, salinity and temperature may also play a role in determining tomorrow's DO. As the body of knowledge expands, it is trivial to add additional terms to this minimization and thereby account for numerous other factors. An example can be found in equation 3.2.3. Please note, each difference $x_{ij} - x_i'_j$ is divided by a scale factor, σ_i. This is due to the fact that these variables are almost certainly not on identical numerical scales. Consequently, with the scale factor in place, this distance minimizing function weights variables on larger and smaller numerical scales appropriately.

The strength of the KNN approach lies in its simplicity, its computational ease with regards to adaptation, and its ability to incorporate new variables. As shown, a variable's importance is quite easily determined by either allowing that one variable to move while holding all others constant and observing sensitivity or by hind-casting with and without that variable and comparing results. If these changes are insignificant, then perhaps that variable need not be included. However, with these strengths comes the weakness associated with many nonparametric models – the curse of dimensionality. As the number of variables grows, the challenge of locating a 'similar' record increases exponentially. The two competing limitations are either failing to include a sufficient number of variables needed to characterize fully the phenomenon in question or alternatively, using a similar set which may not fully share the same statistical properties

as the event it represents. Another nonparametric tool which was explored is that of regression trees (Breiman et al. 1984). Though these bear a resemblance to classification trees which are more commonly used and better known, their advantage lies in their capacity to handle continuous variables. This algorithm splits the dataset into branches by maximizing the ratio of information gain. At the end of each path of branches is a node, within which is a set of historical records which, like KNN, should be similar to the current data point whose future is unknown. Each split is performed by minimizing the sum of variances within the two new nodes – this is to say, the similarity within one node is measured:

$$Similarity = \sum_{i=1}^{r} \left(x_i - \bar{x} \right)^2 \qquad \text{(Equation 3.2.4)}$$

In this case, x_i signifies the value of the input, x-bar represents the mean of the node's population, and r is the total number of elements within the node (Kumar et al, 2006, p392). Each new node can then split itself, or be left as a potential similar set.

This technique relies upon a procedure by which a large data set is divided iteratively, until, like KNN, small subsets are obtained. However, rather than define a distance metric, which incorporates all relevant variables, regression trees, in each division (branch) select the variable that maximizes the information gain (i.e., minimizes the value from equation 3.2.4) within the set of dependent variable records. This is attained by minimizing the variance in the two new subsets of the whole. Repeating this process recursively over these divided subsets yields final 'leaf nodes' whose statistical distributions can be used to generate forecasts. For example, knowing the mean and standard deviation of a node which is similar to a current data point facilitates an estimate

of the probability that current point's dissolved oxygen level will be above or below a given threshold in twenty-four hours.

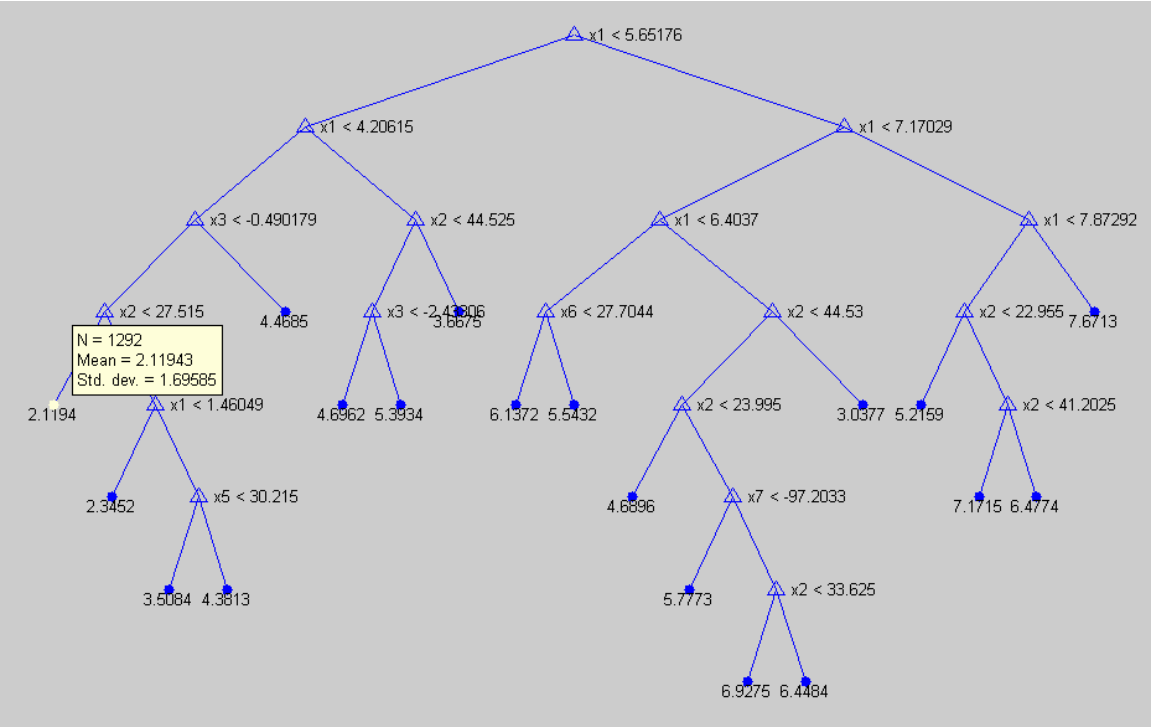


Figure 3.3 – A Sample Regression Tree (the statistics of one node is shown) At each node, the average dissolved oxygen value is shown, as well as which independent variable (numbered 1-7) will be the splitting criterion and the threshold at which the data are cut.

The challenges presented by regression trees lie in determining when to stop splitting the data set. This can be determined via minimum node size, i.e. "stop when the leaves contain fewer than *k* elements," or via maximum node number, i.e. "continue splitting until the total number of nodes reaches *n*." Finally, once the 'tree' is produced, one can 'prune' the leaves, removing those that seem irrelevant or insufficiently dissimilar and allow the parent node to remain. Note, the Matlab classical regression tree library used in this study contains a built-in algorithm to execute this process (Breiman et al, 1993).

Regression trees possess some notable inherent strengths. First, unlike KNN, once the tree is created, its structure can be stored. Therefore, any subsequent event can be funneled through this tree *very* quickly. Having isolated those variables that necessitate a split under any condition and the parameters that define the split, i.e. "if variable x > 2, then choose branch a, else choose branch b," this architecture can remain for an infinite number of predictions. Though the inputs will change as conditions change, the stored structure of the tree need not be updated until it is found that there are so many additional records that a new tree might contain insights lacked in the previous incarnation. KNN on the other hand requires a search and sort, which runs in at best, O[n(log(n))] time for each individual prediction. Additionally, regression trees allow for variables to play a meaningful role in certain cases while being ignored in others where they are less important. For instance, it might be possible that given a certain range of values, dissolved oxygen levels and salinity statistics gain the most information while at other times, DO and temperature are the key variables. By choosing splitting criteria one at a time, different regions of the tree could easily address both circumstances. Furthermore, pruning a tree once it is created is straightforward, allowing alteration if the tree becomes unbalanced.

However, regression trees do possess some notable weaknesses. Because the structure is fixed, it is entirely plausible that two inputs which are similar (but not identical) will produce equivalent results. That is, if a small change in present conditions is not substantial enough to relocate the forecast to a different node, the prediction will not change. Essentially, time series data arrive continuously, but the nodes of a regression tree are disjoint – at any time, conditions force us into one node or another

with no overlap. With KNN, of the *k* similar matches, after a slight change in input conditions, at least a few historical records would no longer fall into that similar set, to be replaced with other matches, thereby altering the implied distribution of outcomes. Avoiding this choppiness of estimation requires that the number of nodes grow very large. However, in minimizing variances for the purpose of maximum information gain in splitting criterion selection (equation 3.2.4), normality (or at least a comparable distribution) is implicitly assumed in the data set that is split. Moreover, to estimate probabilities of future outcomes (i.e. what is the probability of observing hypoxia tomorrow), a distribution of hypoxia forecasts from the set of elements within a node must be assumed. If the nodes are extremely numerous, then the average node size will be quite small, and this assumed distribution may have less support.

3.3 Calibration and Validation Using Historical Data at a Single Location

To calibrate the models described in the previous section at a single location, a sliding window approach is used. A segment of data (the most recent segment) is first reserved for validation. In the Corpus Christi Bay case, the longest-spanning readings from any continuous source are from sensor twenty-four during the summer of 2005. This will be the location used for hind-casting and model validation.

Note: The continuous sample is twenty-three days long. Therefore, the walk forward test can only last twenty-two days. During the sensor's first day of operation, the readings from twenty-four hours prior which would be necessary in a validation hind-cast are missing. Consequently, readings from the first twenty-hours at that particular location cannot be forecast.

3.4 Spatial Interpolation

Having validated forecasts at individual locations with hind-casting, the next step is to interpolate spatially from these results to multiple locations using the following three-step procedure.

I. Establishing Baselines

Before employing the machine learning methods to generate spatial forecasts, a baseline expectation is established for every variable of interest (in this case oxygen) at every location within the area for which predictions will be made. The prediction area is a rectangular space between the user-defined ordered pairs with spatial resolution defined by two user-input parameters. At each of the location coordinates in this grid, the historical database is used to create the best estimate of the mean and standard deviation for each independent variable.

Now to calculate the baseline estimates:

$$\hat{\mu}_{x,y}^{v} = \sum_{i=1}^{n} \left[\frac{\frac{1}{\left(d_{x,y}^{i}\right)^{2}}}{\sum_{i=1}^{n} \frac{1}{\left(d_{x,y}^{i}\right)^{2}}} * \hat{\mu}^{v,i} \right], \ \forall \left\{ \left[x_{\min} + k_1 \frac{\left(x_{\max} - x_{\min}\right)}{x_{precision}} \right], \left[y_{\min} + k_2 \frac{\left(y_{\max} - y_{\min}\right)}{y_{precision}} \right], v \right\}$$

(Equation 3.4.1)

And similarly:

$$\hat{\sigma}_{x,y}^{v} = \sum_{i=1}^{n} \left[\frac{\frac{1}{\left(d_{x,y}^{i}\right)^{2}}}{\sum_{i=1}^{n} \frac{1}{\left(d_{x,y}^{i}\right)^{2}}} * \hat{\sigma}^{v,i} \right], \ \forall \left\{ \left[x_{\min} + k_1 \frac{\left(x_{\max} - x_{\min}\right)}{x_{precision}} \right], \left[y_{\min} + k_2 \frac{\left(y_{\max} - y_{\min}\right)}{y_{precision}} \right], v \right\}$$

(Equation 3.4.2)

(x_{min}, y_{min}) and (x_{max}, y_{max}) – *The boundaries of the user-defined rectangular grid*

$x_{precision}$ and $y_{precision}$ – *Latitudinal and longitudinal resolution over grid*

$\hat{\mu}^{v}_{x,y}$ = *Estimated stationary mean for variable v at point x,y*

$\hat{\sigma}^{v}_{x,y}$ = *Estimated stationary standard deviation for variable v at point x,y*

The historical database consists of discrete stations, each with a contingent of historical data points. For each station, the following is defined:

$\hat{\mu}^{v,i}$ = *Stationary mean for variable v at station i*

$\hat{\sigma}^{v,i}$ = *Stationary standard deviation for variable v at station i,*

And for the spatial baseline estimates:

$d^{i}_{x,y}$ = *Euclidian distance from point x,y to station i.*

For each variable and each input location, the information content is weighted using the inverse of the square of the distance from any point (x,y) to each station.

In these formulations *n* is the total number of measurement stations and k_1 and k_2 are integers between 0 and $x_{precision}$ and 0 and $y_{precision}$ (inclusive) respectively. Inspection of the above formulae reveals that all combinations of k_1 and k_2 will create coordinate pairs for every point within the determined grid. Note that when a baseline estimate is

generated at the exact same location as one of the stations, $\frac{1}{(d_{xy})^2}$ is set to a very large number to ensure that the historical data values at that station are used.

II. Integrating Today's Input

Given the chosen resolution, Step I gives baseline estimates at each location based upon all realizations up to and including (theoretically) yesterday's data. The second step is then to use today's data to generate tomorrow's forecast. Initially, the procedure recreates the same steps as shown above. In this case, instead of *n* stations with their own data distributions, *m* current data points, each containing a vector of relevant independent variable readings are used to compute the following:

$$v_{x,y} = \sum_{j=1}^{n} \left[\frac{\frac{1}{\left(d_{x,y}^{j}\right)^2}}{\sum_{j=1}^{n} \frac{1}{\left(d_{x,y}^{j}\right)^2}} * v^{j} \right], \forall \left\{ \left[x_{\min} + k_1 \frac{(x_{\max} - x_{\min})}{x_{precision}} \right], \left[y_{\min} + k_2 \frac{(y_{\max} - y_{\min})}{y_{precision}} \right], v \right\}$$

(Equation 3.4.3)

Next, the elements of the matrix $v_{x,y}$ in Equation 3.4.3 are normalized by comparing these values to the baseline distributions at the same locations, as follows:

$$N_{x,y}^{v} = \frac{\left(v_{x,y} - \hat{\mu}_{x,y}^{v}\right)}{\hat{\sigma}_{x,y}^{v}}, \forall \left\{ \left[x_{\min} + k_1 \frac{(x_{\max} - x_{\min})}{x_{precision}} \right], \left[y_{\min} + k_2 \frac{(y_{\max} - y_{\min})}{y_{precision}} \right], v \right\}$$

(Equation 3.4.4)

Where

$N_{x,y}^{v}$ = *Normalized value of variable v at location x,y,*

v^{j} = *The value of variable v at measurement location j,*

And for every other non-input location:

$v_{x,y} =$ *The value of variable v at point x,y.*

Again, a distance variable is defined:

$d^{j}_{x,y} =$ *Euclidian distance from point measurement location j.*

This process normalizes the variable by the number of standard deviations of today's data away from the historical estimates at each coordinate pair.

III. Forecasting Tomorrow's Values

Given the spatially normalized values computed in Steps I and II for each of the independent variables, the machine learning algorithms are then implemented as described in section 3.2 using the k-nearest neighbors approach for all relevant variables. One such example construction is given in equation 3.4.5..

$$\min\left\{\begin{array}{l}\left[\frac{(DO_{norm}-DO_i)}{\sigma_{DO_{norm}}}\right]^2+\left[\frac{(Sal-Sal_i)}{\sigma_{Sal}}\right]^2+\left[\frac{(Temp-Temp_i)}{\sigma_{Temp}}\right]^2+\left[\frac{(Wind_{NS}-Wind_{NS_i})}{\sigma_{WIND_{NS}}}\right]^2+ \\ \left[\frac{(Wind_{EW}-Wind_{EW_i})}{\sigma_{WIND_{EW}}}\right]^2+\left[\frac{(Lat-Lat_i)}{\sigma_{Lat}}\right]^2+\left[\frac{(Long-Long_i)}{\sigma_{Long}}\right]^2\end{array}\right\}$$

(Equation 3.4.5)

All variables with the subscript 'i' are historical realizations stored in the database.

4. Results

4.1 Detrending

This section mirrors the progression described in the previous discussion of methodology, showing the findings at each milestone. First, as before, begin with the detrending process. Figure 4.1 represents the same data as in Figure 3.1.1 with the long term decline removed. R^2 statistics reveal that 7% of all variance in dissolved oxygen is addressed by the long term trend, and 24.4% of what remains is described in the annual cycle illustrated in Figures 4.1 and 4.2, which is then removed from the data.

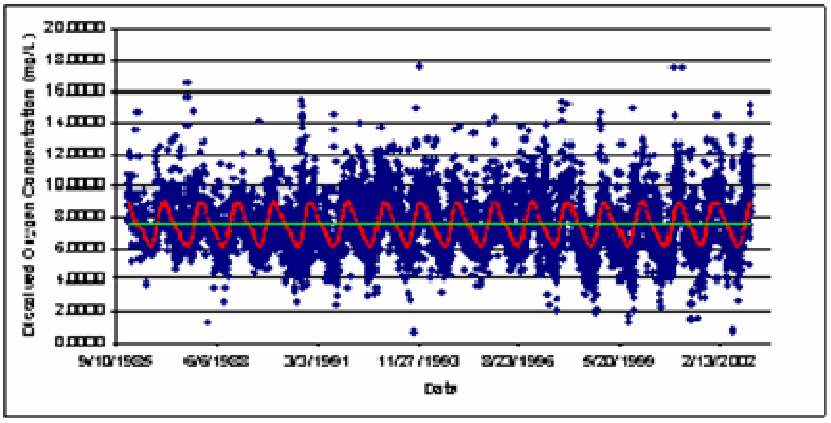


Figure 4.1 – Long Term Data, Long-Term Cycle Removed

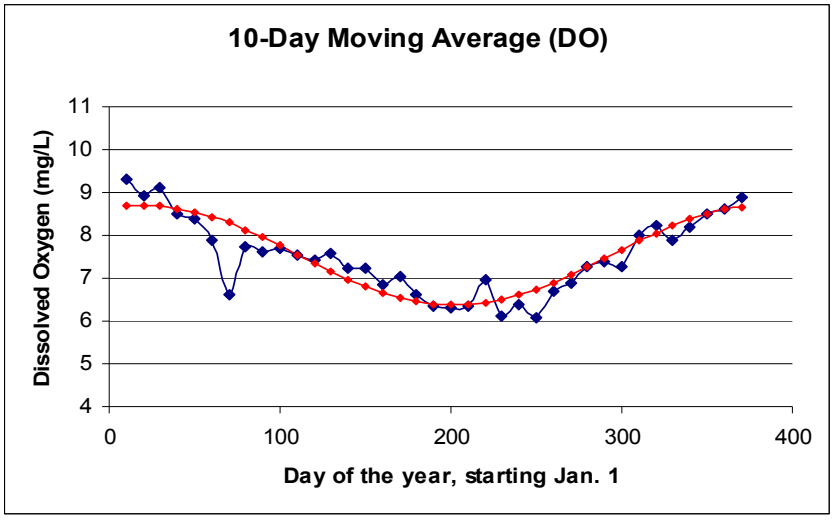


Figure 4.2 The Annual Cycle, Normalized Data (no long-term cycle)
Actual readings given in blue, fit sinusoid given in red.

At this point, having extracted the annual periodicity, the only remaining time cycle contained within the data is the diurnal oscillation. By superimposing all points at each specific time of day (i.e. all points from 12AM, all points from 12:15 AM, etc), the daily sinusoid emerges as shown in Figures 4.3 and 4.4. Yet another Fourier transform with two harmonics can be employed to specify this particular equation, facilitating the normalization described in equation 3.1.8.

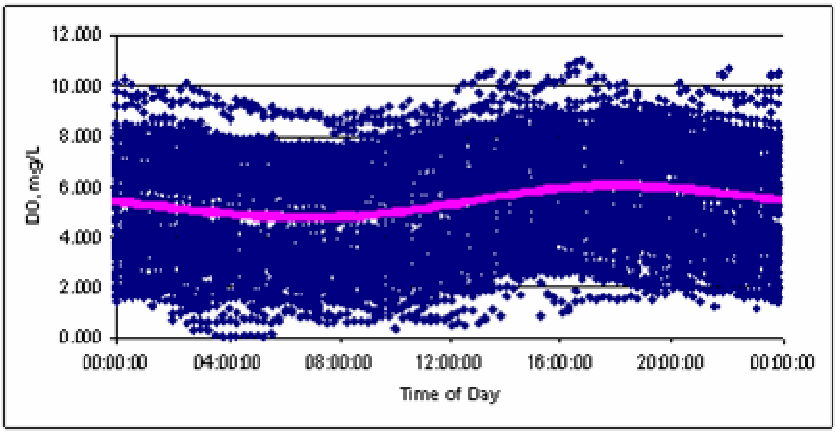


Figure 4.3 – Long Term Data, Long-Term and Annual Cycle Removed

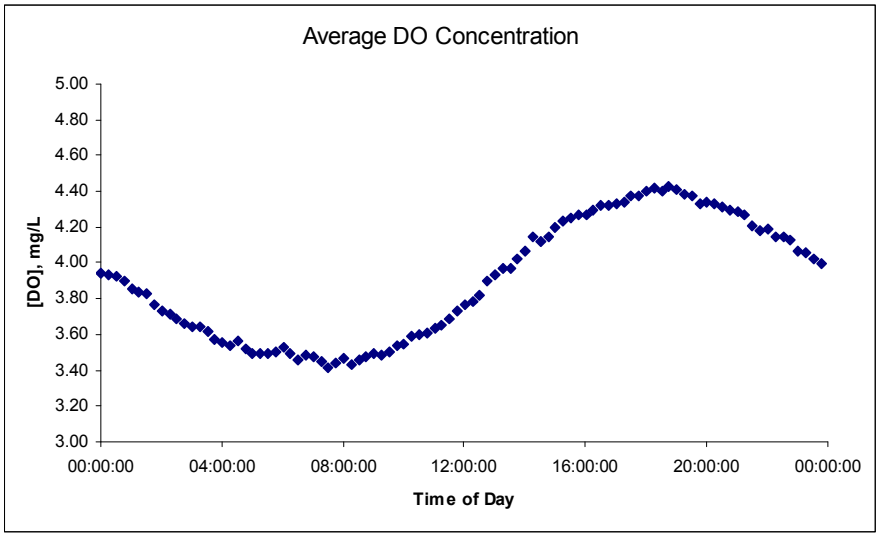


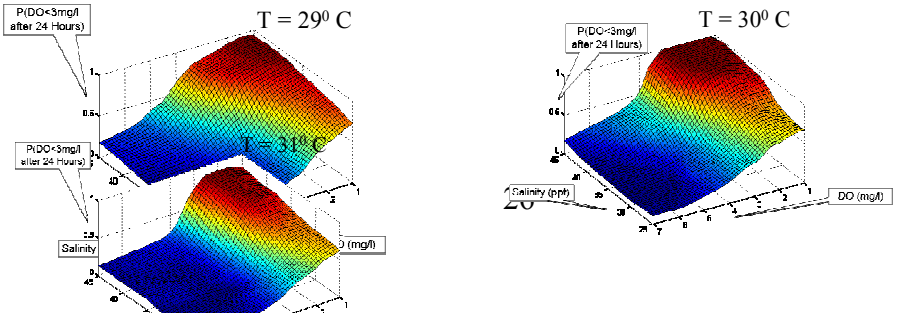
Figure 4.4 – The Diurnal Cycle, Station 24, Summer Data

Each time-dependent cycle is now explicitly specified. With regards to the seasonal periodicity, observe that on average, dissolved oxygen levels are at their lowest

during summer months (Figure 4.2). Moreover, dissolved oxygen levels reach their low points during the early morning hours (Figure 4.4). This is understandable given the photosynthetic mechanism for hypoxia described in the introduction. Hours of viable sunlight allow for oxygen to be produced by flora, which is subsequently consumed by faunae in the nighttime hours leaving a minimum just after sunrise when the rate of photosynthetic oxygen production surpasses the rate of consumption. Figure 4.4 is consistent with this mechanism, as the minimum point is found around 6:00 AM with a peak coming in early evening hours as the sun begins to set.

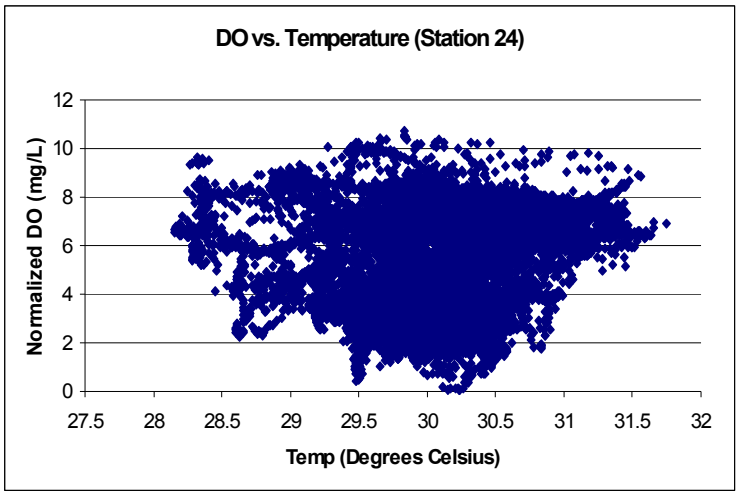
4.2 Determining Relevant Variables

Using the fully-detrended dataset, the next step is to observe which variables (apart from time) play the greatest role in predicting dissolved oxygen twenty-four hours hence. As described previously, saline water masses entering Corpus Christi Bay from the Gulf of Mexico are hypothesized as potential drivers of hypoxia. Additionally, temperature plays a role in determining oxygen's solubility in water. For the purposes of this inquiry, variables were tested by gathering data from one location (station 24) for which over 12,000 records were available. On this subset of the database, the k-nearest neighbor algorithm was employed to forecast dissolved oxygen levels in twenty-four hours time using only two independent variables are used, one of which is dissolved oxygen itself. The gradient in oxygen predictions for different values of each variable are then used to determine the sensitivity of forecasts to that variable and consequently, its utility for forecasting, as shown in Figure 4.2 a.



Figures 4.2. a through c Probability of Hypoxia vs. Salinity, Temperature Held Constant

Figure 4.5 c. Figures 4.5. a through c show that the most important variable in forecasting dissolved oxygen tomorrow is dissolved oxygen today, but salinity also plays a substantial role. The gradient along the salinity axis in Figure 4.5 b. is considerable even when juxtaposed with dissolved oxygen. However, taking note of the fact that the general shape of each figure is slightly different at varying levels of temperature (even if only separated by a degree or two Celsius), temperature also appears to be a relevant predictive variable.



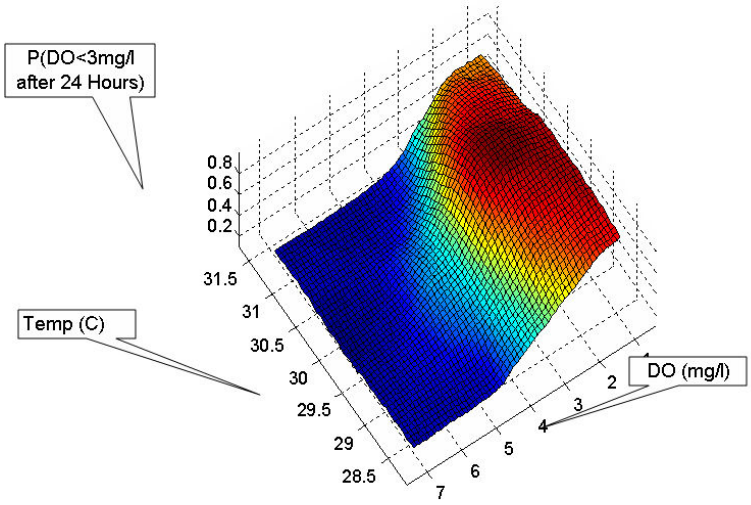


Figure 4.7, Dissolved Oxygen vs. Temperature

As shown in both figures 4.6 and 4.7, the middle range of summer temperatures seems to lend itself to incidents of lower dissolved oxygen levels, and therefore, hypoxic conditions. This is an interesting development in that, as is seen in Figure 4.6, a regression over the scatterplot would yield an insignificant coefficient of correlation, though Figure 4.7 confirms that the role of temperature is anything but statistically insignificant. It is possible that such a pattern is a relic of the diurnal cycle, since the hypoxic events often occur in the early morning, which would *not* represent the highest temperature of the day. Theory dictates that higher temperatures should cause lower dissolved oxygen levels due to oxygen lost to the atmosphere and decreased solubility (Lopes et al 2008). However, analyses performed to verify that this odd shape is an echo of the diurnal oscillation did not confirm this hypothesis. The same plot shown in figure 4.2.4 was recreated using only data from specific time periods of the day (i.e. only records gathered between 12 and 2 AM, 2 and 4 AM, etc). These subsets of the data still displayed the same triangular-shaped scatter-plot, leading to the conclusion that the diurnal cycle is not a sufficient explanation for this feature of the data. As a result, temperature will be included in the independent variables used for forecasting.

4.3 Single & Multiple Location Forecasts

Having determined that dissolved oxygen, salinity, temperature, and potentially, wind are good candidates as independent variables for generating predictions of hypoxic events after twenty-four hours, dissolved oxygen levels are forecast at one particular location. The expected value for dissolved oxygen is inferred using the mean of the K-

nearest neighbors, while the variance is used to generate confidence intervals around the calculated expectation. Finally, by observing which proportion of the similar set becomes hypoxic after twenty-four hours, a non-parametric estimate of the probability of observing hypoxia tomorrow is calculated.

Expanding this single location forecast to yield an encompassing spatial forecast is complicated by the fact that only stations generating outputs at regular intervals (in this case fifteen minutes) can be employed. Furthermore, these sensors must be located at the bay floor to pick out the hypoxic precursors and the subsequent hypoxic incidents they cause. Estimates will be most robust within the convex hull of the sensors providing regular-interval output. However, by normalizing the distribution at each location, incorporating latitude and longitude into the k-nearest neighbor algorithm as independent variables, it is possible to iterate over a grid of latitudinal and longitudinal coordinates and generate a spatial map (see, e.g., Figures 4.8 through 4.10). These figures were generated using dissolved oxygen, salinity, temperature, and wind (speed and direction) over the latitudinal and longitudinal ranges illustrated.

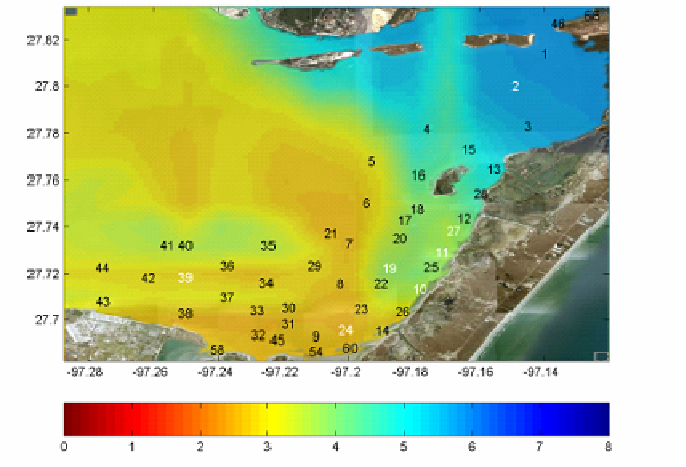


Figure 4.8 - Expected Dissolved Oxygen Levels (mg/l)

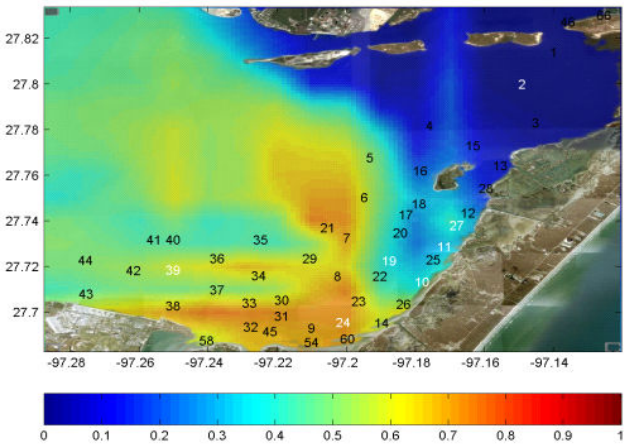


Figure 4.9 – Probability of Hypoxia

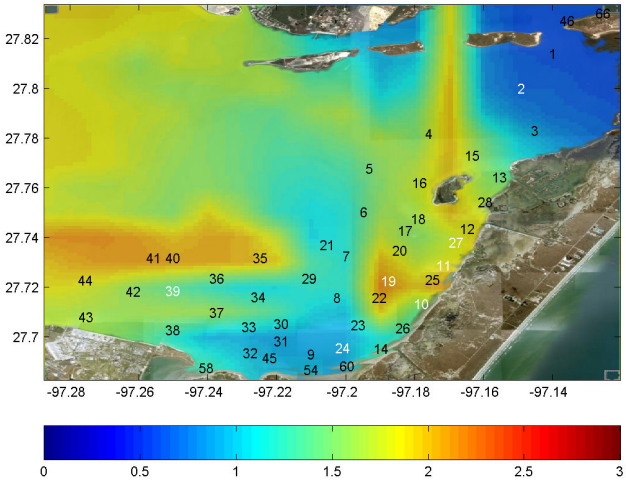


Figure 4.10 – Standard Deviation (mg/l)

In addition to aiding understanding of the spatial aspects of hypoxia, these images provide insight for further sampling locales. Regions characterized by higher standard deviations (Figure 4.10), reflecting more uncertainty, as well as lower expected dissolved oxygen levels (Figure 4.8) are prime candidates for an additional sensor. Based on the results given in Figures 4.8 through 4.10, researchers introduced new sensors (#8, #34, #199, and #202 in Figure 4.11) in summer 2007, from which they learned that hypoxic risk is far wider than initially believed (Figure 4.11). Once again, this figure was

constructed using dissolved oxygen, salinity, temperature, wind speed and wind direction over the range illustrated – this time new station data were included.

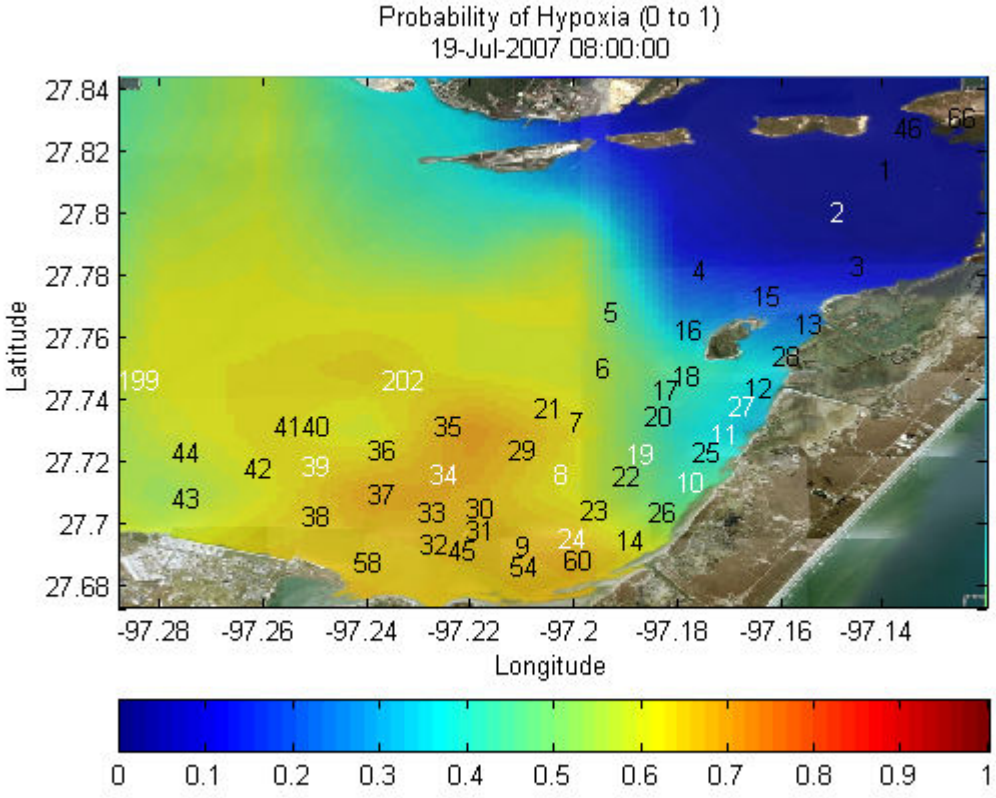
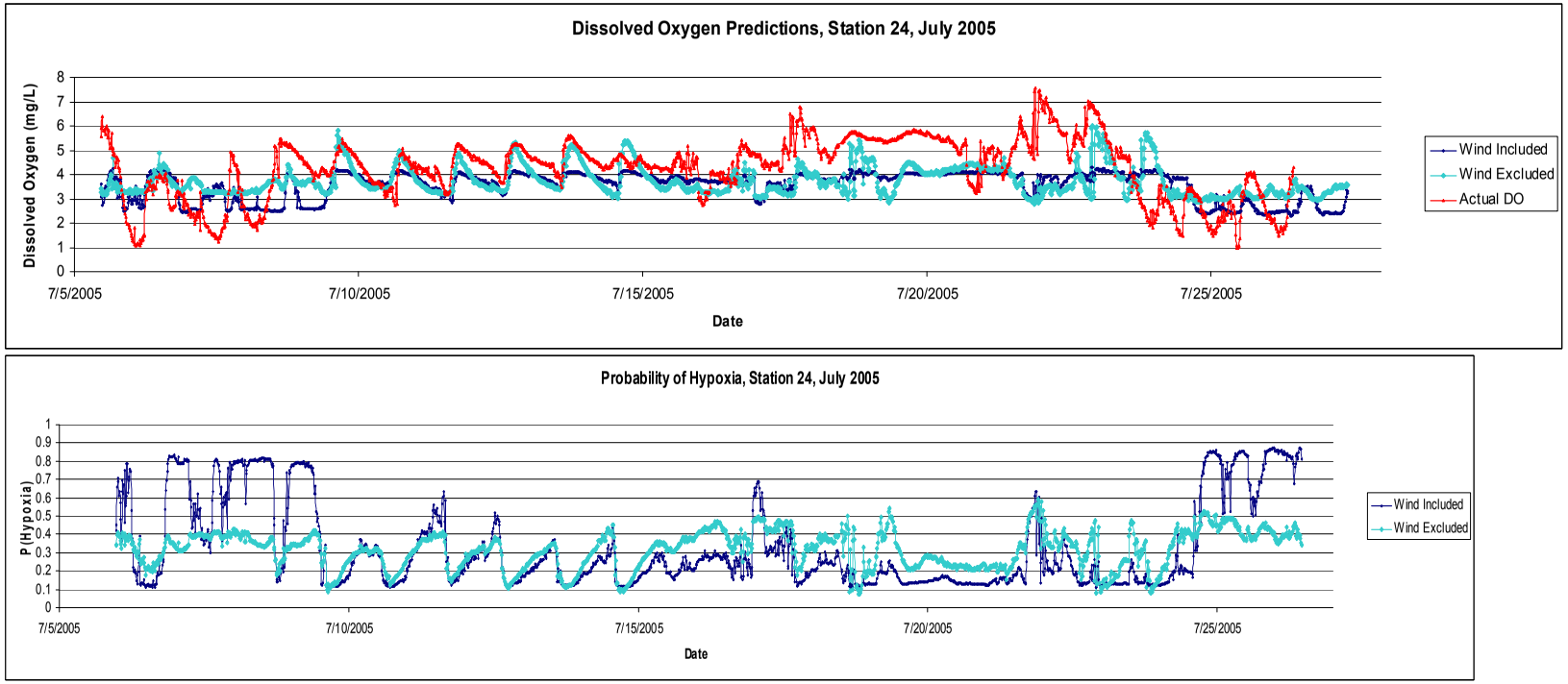


Figure 4.11 – Hypoxic Risk with New Sensors Included During Summer 2007

4.4 Verification of Hydrodynamic Hypothesis: Wind's Impact

To this point, the meteorological events that adjust the trajectory of water masses and consequently, affect the probability of hypoxia are ignored. As mentioned in the introduction and the methodological description of spatial mapping in section three, wind is a variable whose impact must be explored and verified. Since wind is a two-dimensioned variable (a wind vector contains speed and direction), it requires two additional dimensions within the k-nearest neighbor algorithm, and separate results.

Once again, station twenty-four is used, for a period within the summer of 2005 during which continuous readings were taken every fifteen minutes for approximately twenty days. Figure 4.12 shows the projections of dissolved oxygen levels with and without the inclusion of wind data (dark and light blue lines, respectively) as well as the actual observed result (red). Figure 4.13 shows the model's estimation of the probability of hypoxia with and without the addition of wind data. In the absence of the wind data, the estimated probability of hypoxia fails to exceed 50%. Once wind data are added, the probability of hypoxia hovers at or above 80% at the beginning and the conclusion of this recorded interval. These periods are coincident with the observations of hypoxic events at those times. Furthermore, during the days near the 20th of July, dissolved oxygen levels remain stable and well above hypoxic levels. The wind-aided model estimates the probability of hypoxia as barely exceeding 10%. Once the algorithm is bolstered through the inclusion of wind data, its accuracy, as well as its sensitivity, is improved noticeably as small changes in the wind conditions cause substantial changes in the estimated probability of hypoxia.



Figures 4.12 and 4.13, Top: Dissolved Oxygen Levels
Bottom: Estimated Probability of Hypoxia

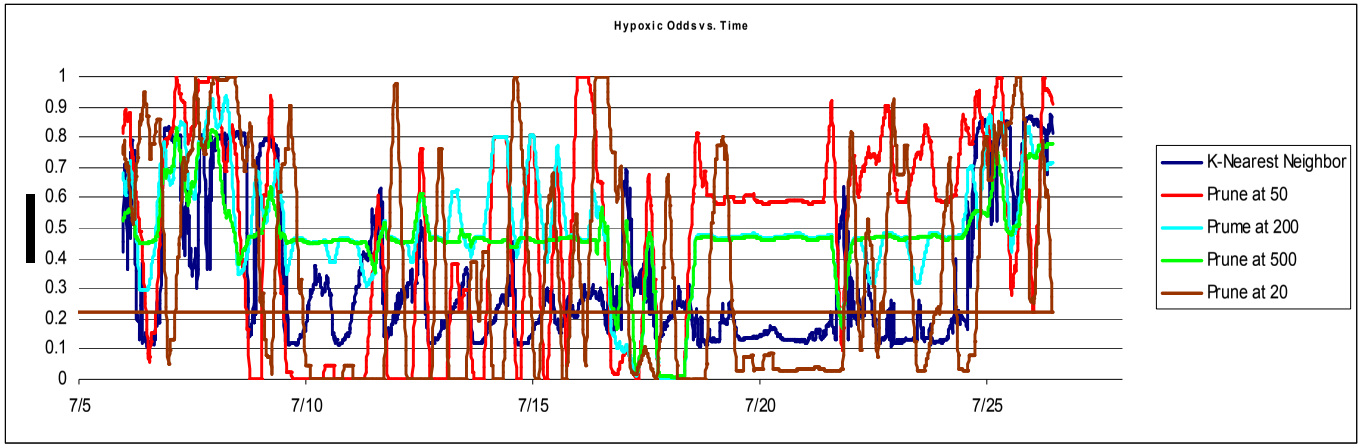
4.5 Regression Tree Results (The Model *Not* Chosen)

These discussions have centered around the k-nearest-neighbor algorithm. In this section KNN and regression trees are compared by applying the regression tree algorithm over the same continuous period as the previous section (see Figures 4.14 and 4.15).

As discussed, one of the weaknesses of regression trees is that the model will lack sensitivity to incremental changes in the independent variables if the node size is too large. Indeed, in the two graphs, the regression tree forecasts vary little for extended periods of time, especially for the larger node sizes of 200 and 500. Conversely, when node sizes shrink, the assumptions of normality within a similar sample are weakened, and large fluctuations in the probability estimates may be observed. This is observed in Figure 4.14 when the node size shrinks to 20 or 50.

Though KNN algorithms behave like regression trees by generating a similar subset of the original database which represents the event to be forecasted, KNN allows its similar set to vary dynamically in time. For example, if a subset contains 500 matches, fifteen minutes later, a new 500-sample subset is created, but it likely that the vast majority of this subset overlaps with the previous, allowing for subtle sensitivity to remain. Regression trees require total replacement or none at all. That is to say, from one point in time to the next, when locating the records which will fill the similar set, either the same node is retained as a similar set (no change in the forecast) or another node consisting of completely different records replaces it. Any attempts to mitigate this problem by creating increasing numbers of smaller population nodes will necessarily cause the analysis to threaten assumptions of the normality of the underlying node distribution. Nodes are assumed to behave as normal distributions and very small

number of records may not define well a Gaussian distribution. Understanding this underlying distribution is required to generate a probability estimate. For this reason, the conclusion is that using a discontinuous algorithm like regression trees to solve a continuous time-series forecast is less effective.



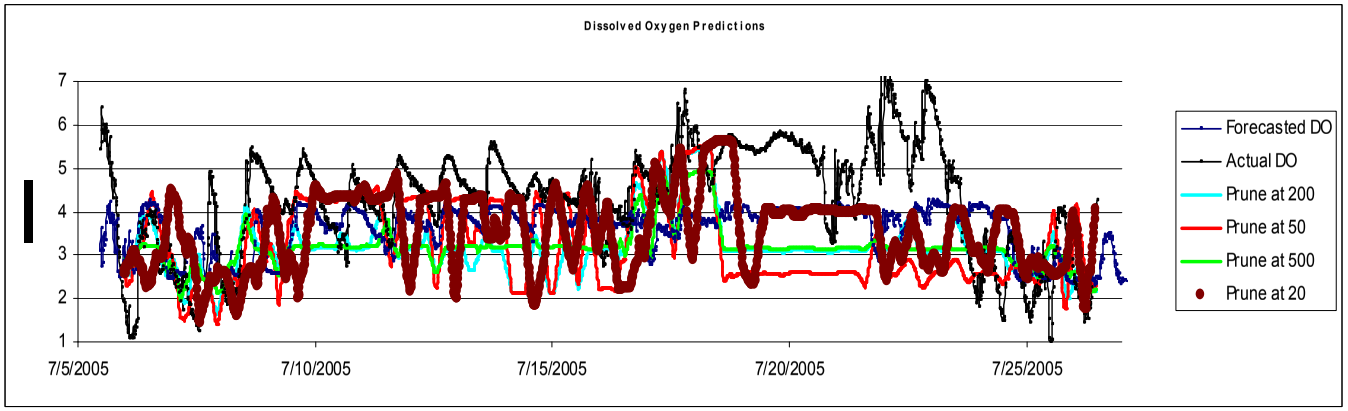


Figure 4.14 (Above) – Hypoxic Odds, Station 24, Summer '05, Regression Trees

Figure 4.15 (Below) – Predicted Dissolved Oxygen Level, Regression Trees

5. Conclusion

The analysis indicates that k-nearest neighbor algorithms can produce only reasonable estimates of the probability of observing a hypoxic event the following day given inputs of dissolved oxygen, salinity, water temperature, wind speed, and wind direction for the present bay status. These results corroborate the hydrodynamic observations of previous researchers. By implementing machine learning algorithms to isolate those variables most correlated with future dissolved oxygen levels, confirmations of hypotheses regarding the mechanisms that cause hypoxia in Corpus Christi Bay are aided. K-nearest-neighbor models provide insight into these mechanisms by displaying sensitivity to those factors which alter dissolved oxygen levels most.

Using spatial interpolation techniques in conjunction with sequential normalization of time-dependent data at an array of continuous benthic sensors, it is possible to generate a bay-wide forecast on any coordinate system. Real-time forecasting is entirely possible with incomplete data, though potentially compromised in its accuracy. Using the k-nearest neighbor technique (which outperforms regression trees considerably), the probabilistic estimates are verifiable and facilitate the informed positioning of future sensors. This. adaptive sampling approach has already identified that the phenomenon stretches farther west than had been suspected over the first two decades of observing Corpus Christi Bay hypoxia.

Ultimately, in addition to the gains in mathematical forecasting, this analysis fits well with proposed hydrodynamic mechanisms. The impacts of saline water entering the bay, wind-effected dissolved oxygen levels, and fluctuating water temperatures are observed. In the future, one can envision statistically defensible real-time forecasts,

optimized sensor placements, and collaboration on further research between those working with mechanistic modeling of hydrodynamics and those producing machine learning algorithms for the purposes of better understanding those mechanisms which drive hypoxia in Corpus Christi bay and elsewhere.

References

Booth, M.S. and Campbell, C. 2007. Spring Nitrate Flux in the Mississippi River Basin: A Landscape Model with Conservation Applications. Environmental Science & Technology; 8/1/2007, Vol. 41 Issue 15, p5410-5418.

Brammer, A.J. and Z.R. Del Rey, E.A. Spalding, and M.A. Poirrier. Effects of the 1997 Bonnet Carré Spillway Opening on Infaunal Macroinvertebrates in Lake Pontchartrain, Louisiana. Journal of Coastal Research; Sep2007, Vol. 23 Issue 5, p1292-1303.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA. 1984.

Breiman, L., et al., *Classification and Regression Trees*, Chapman & Hall, Boca Raton, 1993.

Campbell, J.G. and Goodman, L.R. 2007. Lethal levels of hypoxia for gulf coast estuarine animals. Marine Biology; Aug2007, Vol. 152 Issue 1, p37-42.

Dauer, D and A.J. Jr. Rodi, J.A. Ranasinghe. 1992. Effects of low dissolved oxygen events on the macrobenthos of the lower Chesapeake Bay, *Estuaries*, 15, p384-391.

Donnelly, K.A. and Scavia, D. 2007. Reassessing Hypoxia Forecasts for the Gulf of

Mexico. Environmental Science & Technology; 12/1/2007, Vol. 41 Issue 23, p8111-8117.

Fayyad, U, and G. Piatetsky-Shapiro, P. Smyth. 1996. From data mining to knowledge discovery in databases. Ai Magazine, 17(3), p37-54.

Galkovskaya, G.A. and Mityanina, I.F. 2005. Structure Distinctions of Pelagic Rotifer Plankton in Stratified Lakes With Different Human Impact. Hydrobiologia; Sep2005, Vol. 546 Issue 1-3, p387-395.

Goldshmid, R and R. Holzman, D. Weihs, and A. Genin. 2004. Aeration of corals by sleep-swimming fish. Limnology & Oceanography; Sep2004, Vol. 49 Issue 5, p1832-1839.

Google Earth, Image from NASA, © 2007 Tele Atlas, © 2007 Europa Technologies, Image © 2008 DigitalGlobe.

Hagy, J.D. and Murrell, M.C. 2007. Susceptibility of a northern Gulf of Mexico estuary to hypoxia*:* An analysis using box models. Estuarine Coastal & Shelf Science; Aug2007, Vol. 74 Issue 1/2, p239-253.

Hetland, R.D. and 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; Mar2008, Vol. 70 Issue 1/2, p49-62.

Hodges, B. R. and J. E. Furnans. 2007. Thin-layer gravity currents in a shallow estuary. Proc., 18th Engineering Mechanics Division Conference (EMD2007) June 3-6, Blacksburg, VA, USA.

Hodges, B.R**.** and J.E. Furnans (2007), "Linkages between hypoxia and thin-layer stratification in Corpus Christi Bay," manuscript in revision for *Environmental Fluid Mechanics* (July, 2007).

Kulis, P. and Hodges, B. R. (2006b) "Three Dimensional Circulation in Corpus Christi Bay", Gulf Estuarine Research Society Meeting, November 2-4, 2006, Corpus Christi, Texas.

Kumar, P. and J. Alameda, P. Bajcsy, M. Folk, M. Markus. *Hydroinformatics*. CRC Press, Taylor & Francis Group, New York, New York, 2006.

Loewen, M.R. and J.D. Ackerman, P.F. Hamblin. Environmental implications of stratification and turbulent mixing in a shallow lake basin; Jan2007, Vol. 64 Issue 1, p43-57

Lopes, J.F. and C. I. Silva, A. C. Cardoso. Validation of a water quality model for the Ria

de Aveiro lagoon, Environmental Modelling & Software; Apr2008, Vol. 23 Issue 4, p479-494.

M.A. Mallin and V.L. Johnson, S.H. Ensign, T.A. MacPherson. 2006. Factors contributing to hypoxia in rivers, lakes, and streams. Limnology & Oceanography; Jan2006 Part2, Vol. 51 Issue 1, p690-701.

Maidment, D.R. and E. Parzen, 1984: Time Patterns of Water Usage in Six Texas Cities. *J. of Water Resour. Plann. Manage.,* 110, p90-106.

McRoberts, R.E. and E.O. Tomppo, A.O. Finley, and J. Heikkinen. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sensing of Environment; Dec2007, Vol. 111 Issue 4, p466-480.

Meliker, J.R. and G.A. AvRuskin, M.J. Slotnick, P. Goovaerts, D. Schottenfeld, G.M. Jacquez, and J.O. Nriagu. Validity of spatial models of arsenic concentrations in private well water. Environmental Research; Jan2008, Vol. 106 Issue 1, p42-50

Montagna, P.A. and R.D. Kalke (1992), The effect of freshwater inflow on meiofaunal and macrofaunal populations in the Guadalupe and Nueces Estuaries, Texas, *Estuaries*, 15, p307-326.

Montagna, P.A. and C. Ritter. 2006. Direct and indirect effects of hypoxia on benthos in

Corpus Christi Bay, Texas, U.S.A. *Journal of Experimental Marine Biology and Ecology* 330: p119-131.

Nemes, A. and R.T. Roberts, W.J. Rawls, Ya.A. Pachepsky, and M. Th. van Genuchten. Software to estimate −33 and −1500kPa soil water retention using the non-parametric k-Nearest Neighbor technique Environmental Modelling & Software; Feb2008, Vol. 23 Issue 2, p254-255.

Osterman, L.E. and R.Z. Poore, P.W. Swarzenski, The last 1000 years of natural and anthropogenic low-oxygen bottom-water on the Louisiana shelf, Gulf of Mexico. Marine Micropaleontology; Feb2008, Vol. 66 Issue 3/4, p291-303, 13p

Ritter, M.C. and P.A. Montagna. 1999. Seasonal hypoxia and models of benthic response in a Texas bay. *Estuaries* 22: p7-20.

Texas Parks and Wildlife Department**,** 4200 Smith School Road, Austin, TX 78744 http://www.tpwd.state.tx.us/

Ward, G.H. 1997. Processes and Trends of Circulation Within the Corpus Christi Bay National Estuary Program Study Area, Technical Report CCBNEP-21, Corpus Christi National Estuary Program, 286p.