Understanding and forecasting hypoxia using machine learning algorithms
Evan Joseph Coopersmith, Barbara Minsker and Paul Montagna

ABSTRACT
This study’s primary objective lies in short-term forecasting of where and when hypoxia may transpire to enable 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 (daily, seasonal, and long-term). To predict hypoxic events, these three mathematical trends are isolated and extracted to obtain unbiased forecasts using a sequential normalization approach. Next, machine learning algorithms are constructed employing the continuous, normalized values from a variety of sensor locations. By including latitude and longitude coordinates as additional variables, a spatial depiction of hypoxic conditions can be illustrated effectively, allowing for more efficient summer data collection and more accurate, near-real-time projections. Using $k$-nearest neighbor and regression tree 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 exceeded 80%; Conversely, during periods in which no hypoxia was observed, the model’s estimate remained below 20%. These results indicate that the modeling approach produces reasonable forecasts for this case study.

Key words | forecasting, hypoxia, $k$-nearest neighbors, machine learning

INTRODUCTION
Hypoxia is a condition in which aquatic dissolved oxygen (DO) levels drop to levels below 2 mg/L (Dauer et al. 1992). Generally, research has found hypoxia resulting from anthropogenic eutrophication (Campbell & Goodman 2007), biological oxygen demand (Mallin et al. 2006), and more specifically, the leaching of nitrates from fertilizers into local groundwater (Booth & Campbell 2007).

Hypoxia has been observed in watersheds from the Great Lakes (Loewen et al. 2007), to European lakes (Galkovskaya & Mityanina 2005), and has spread considerably in marine ecosystems since the 1960s (Diaz & Rosenberg 2008). Corpus Christi Bay, Texas (USA) hypoxia results from a seasonal periodicity of dissolved oxygen fluctuations (Ritter & Montagna 1999), a diurnal cycle resulting from tides (Hagy & 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 & Furnans 2007b).

Hypoxic conditions were first observed in Corpus Christi Bay in 1988 (Montagna & 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 (Montagna & Ritter 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 & 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 (HABHRC) 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 deepest water layer, the hypolimnion (Galkovskaya & Mityanina 2005). In Corpus Christi Bay, the onset of hypoxia is frequently coincident with the entrance of hypersaline water from Laguna Madre or Oso Bay into the less saline water of Corpus Christi Bay (Hodges & Furnans 2007a). 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 & 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. Before machine learning algorithms can be executed in the manner that will best explain the phenomena in question, data must first be appropriately prepared (Pyle 1999). 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). Other popular techniques employed in data-driven modeling efforts are illustrated in the context of river basin management (Solomatine & Ostfeld 2008). A newer blend of traditional regression and genetic algorithm approaches is employed to integrate scientific insight and numerical fitting in knowledge discovery (Giustolisi & Savic 2006). This new methodology can be used for time-series analysis or even to augment an existing dataset via interpolation (Giustolisi et al. 2007). Currently, machine learning algorithms are widely used for the purpose of hydrological modeling of remote sensing data using artificial neural networks (Evora & Coulibaly 2009), fuzzy inference systems for modeling river flow (Jaccquin & Shamseldin 2009), and evolutionary polynomial regression to model soil moisture (Elshorbagy & El-Baroudy 2009).

In this paper, 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, the methodology used to analyze the data, the most important results of that analysis, and finally, the conclusions drawn from those results.

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 225 km south-southeast of San Antonio (Figure 1). For more information on Corpus Christi Bay, please see http://www.serf.tamus.edu/.

The bay’s bathymetry is characterized by an extremely flat floor at an average depth of 3.6 m (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 the seventh largest port in the United States. 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 & 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 wind-driven water masses moving northward from Laguna Madre may
provide the mechanism by which saline water enters the bay and generates hypoxic conditions (Hodges & Furnans 2007a). In addition to the salinity currents, wastewater treatment plants on the Oso Creek (to the west of the bay with “WWTP” labels in Figure 1), discharge into Oso Bay and from there to Corpus Christi Bay, providing a potential anthropogenic source of contaminants which, along with the aforementioned agricultural runoff, can lead to nitrate-based eutrophication.

Within Corpus Christi Bay lies a spatial grid of oxygen sensors (shown in detail in Figure 2) that provides oxygen levels to an accuracy of 0.1 mg/L 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 et al. 2008). Moreover, in Corpus Christi Bay and other locations studied 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 other sources of data, referred to as ‘continuous’ samples. For the continuous samples, 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 shows the sampling locations in Corpus Christi Bay, with grab sample locations given in black and continuous sensor locations given in white.

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) is used primarily as a means of numerical control since that location virtually never experiences a hypoxic event due to its distance from the influx of saline water from the Gulf of Mexico. The black and white gradient below illustrates dissolved oxygen levels, measured in mg/L.

**METHODOLOGY**

To improve understanding and create near-term forecasts of hypoxia in Corpus Christi Bay, this study uses a data mining approach. Data mining is a process for extracting knowledge from data, with the steps illustrated in Figure 3. First, data are gathered, cleaned, and processed. Next, those data are transformed appropriately. Iterative algorithms are performed on those data and output is evaluated. At this point, the patterns obtained from mathematical analysis can become useful knowledge.

For the purposes of studying the occurrence 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). 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. Analysis indicates that before these trends are extracted, over 25% of the total variance associated with our historical database of dissolved oxygen values was due solely to temporal factors. Normalization leaves transformed data (Figure 3) 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. Additional relevant variables such as salinity, pH, temperature, conductivity, etc. are examined to determine if they are useful in predicting dissolved oxygen levels. The machine learning models for real-time predictions
such as simple auto-correlations, KNN, and regression trees are assessed, the pros and cons of each are weighed, and once a model is selected (in this case, KNN), 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. The steps in this process are detailed in the following four sub-sections.

Removing time-dependent trends

This section addresses the process of removing time-dependent trends. First, over 25 years of historical average daily dissolved oxygen readings taken by the Texas Parks and Wildlife Commission in Corpus Christi Bay (see Figure 2) are examined, shown in Figure 4. A cursory examination reveals that not only is there a slow, yet persistent decline in dissolved oxygen levels (Applebaum et al. 2005), 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.

The sinusoidal curve in Figure 4 was calculated by minimizing the sum of squared errors from the surrounding data points. The curve is comprised of the sum of a linear function and the harmonics of a sinusoidal wave with a period of one year. Detrending is accomplished via sequential normalization (Maidment & 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. Though Maidment and Parzen analyzed trends of monthly water usage, the notion
of eliminating the longest spanning trends first and proceeding sequentially to shorter cycles is analogous.

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

\[ \text{DO}_{\text{long-term}}(t) = k_1 t + k_2 \]  

(1)

Note: In Equations (1)–(9), \( k_1 \) through \( k_{12} \) are simply numerical constants which are obtained by minimizing the sum of squared residuals. For instance, in Equation (1), \( k_1 \) and \( k_2 \) are the constants from a linear regression. A linear model was selected over other exponential and logarithmic decay functions due to higher correlation coefficient values.

To remove this trend, the dataset is normalized to smooth the long-term decrease and remove cyclical functions. In Equations (2)–(9), \( n_1 \), \( n_2 \), and \( n_3 \) represent the number of dissolved oxygen points \( \text{DO}_i \):

\[ \text{DO}_{\text{norm.}} = \frac{\text{DO}(t)}{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \right)} = \frac{k_1 t + k_2}{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \right)} \]  

(2)

The normalized DO values are then adjusted by comparing each data point to the value we might expect from that point in our historical dataset of continuous sensors (beginning in 1999), creating a DO time series which does not contain the long-term decay:

\[ \text{DO}_{\text{adjusted}} = \frac{\text{DO}_i}{\text{DO}_{\text{norm.}}} = \frac{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \right)}{k_1 t + k_2} \]  

(3)

At this stage, the long-term trend is gone. Since the remaining trends are both periodic, modeling them via cyclical functions is the next step. By constructing a discrete Fourier transform with two harmonics, capturing the annual periodicity within this adjusted data is now possible:

\[ \text{DO}_{\text{seasonal}}(t) = k_3 \sin \left( \frac{2\pi t}{365} - k_4 \right) + k_5 \sin \left( \frac{2\pi t}{182.5} - k_6 \right) + k_7 \]  

(4)

\( t \) is measured in units of days.

An analogous normalization (the first normalization removed the long-term trend while this second normalization removes the annual cycle) is then performed to generate yet another adjusted dataset, which lacks both a long-term trend and an annual oscillation:

\[ \text{DO}_{\text{twice-norm.}} = \frac{\text{DO}_{\text{norm.}} \text{DO}_{\text{seasonal}}(t)}{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_{\text{norm.}}}{n_2} \right)} \]  

\[ = \frac{k_1 t + k_2}{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \right)} \frac{k_3 \sin \left( \frac{2\pi t}{365} - k_4 \right) + k_5 \sin \left( \frac{2\pi t}{182.5} - k_6 \right) + k_7}{\left( \frac{\sum_{i=1}^{n_2} \text{DO}_{\text{norm.}}}{n_2} \right)} \]  

(5)

Finally, to remove the only remaining cycle, 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):

\[ \text{DO}_{\text{twice-adjusted}} = \text{DO}_i \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \frac{\sum_{i=1}^{n_2} \text{DO}_i}{n_2} \]  

\[ = \text{DO}_i \frac{k_3 \sin \left( \frac{2\pi t}{365} - k_4 \right) + k_5 \sin \left( \frac{2\pi t}{182.5} - k_6 \right) + k_7}{\left( \frac{\sum_{i=1}^{n_2} \text{DO}_{\text{norm.}}}{n_2} \right)} \]  

(6)

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

\[ \text{DO}_{\text{diurnal}}(t) = k_8 \sin \left( \frac{2\pi t}{24} - k_9 \right) + k_{10} \sin \left( \frac{2\pi t}{12} - k_{11} \right) + k_{12} \]  

(7)

\( t \) is now measured in units of hours.

Continuing:

\[ \text{DO}_{\text{thrice-norm.}} = \text{DO}_{\text{twice-norm.}} \frac{\text{DO}_{\text{diurnal}}(t)}{\left( \frac{\sum_{i=1}^{n_3} \text{DO}_i}{n_3} \right)} \]  

\[ = \frac{k_1 t + k_2}{\left( \frac{\sum_{i=1}^{n_1} \text{DO}_i}{n_1} \right)} \frac{k_3 \sin \left( \frac{2\pi t}{365} - k_4 \right) + k_5 \sin \left( \frac{2\pi t}{182.5} - k_6 \right) + k_7}{\left( \frac{\sum_{i=1}^{n_2} \text{DO}_{\text{norm.}}}{n_2} \right)} \frac{k_8 \sin \left( \frac{2\pi t}{24} - k_9 \right) + k_{10} \sin \left( \frac{2\pi t}{12} - k_{11} \right) + k_{12}}{\left( \frac{\sum_{i=1}^{n_3} \text{DO}_i}{n_3} \right)} \]  

(8)
At this point, normalization has removed three time dependencies. We verified this by re-analyzing the normalized data for trends; the normalized data exhibited minimal correlations with respect to time. Consequently, any value from any historical record can be compared to any other by adjusting thrice, as follows:

\[
\text{DO}_{\text{thrice--adjusted}} = \text{DO} \left( \frac{\sum_{i=1}^{s_1} \text{DO}_i}{n_1} \right) \left[ k_1 t + k_2 \right] \left[ k_3 \sin \left( \frac{2 \pi t}{365} - k_4 \right) + k_5 \sin \left( \frac{2 \pi t}{182.5} - k_6 \right) + k_7 \right] \left[ k_8 \sin \left( \frac{2 \pi t}{1} - k_9 \right) + k_{10} \sin \left( \frac{2 \pi t}{3.5} - k_{11} \right) + k_{12} \right]
\]

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 (normalized DO, salinity, temperature, pH, conductivity, wind, etc) 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 \textit{k}-nearest neighbor algorithm, abbreviated ‘KNN’. KNN remains among the most straightforward methods for classification, analysis, and forecasting. (Kumar et al. 2006). With KNN, 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. In implementing this algorithm, only continuous sensor data are used. KNN’s approach aims to classify any record by locating the nearest records via some previously defined metric for distance based on the similarity of all independent variables. In this instance, a simple Euclidean distance function \( d \) is applied to determine the similarity between one data point \( x_i \) and another historical value \( x_{ij} \) (Kumar et al. 2006):

\[
d(x_i, x_{ij}) = \sqrt{\sum_{i} (x_{ij} - x_i)^2}, \forall i, j
\]

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

\[
\min \left\{ \frac{(\text{DO}_{\text{norm}} - \text{DO}_i)^2}{\sigma_{\text{DO}_{\text{norm}}}} + \frac{(\text{Sal} - \text{Sal}_i)^2}{\sigma_{\text{Sal}}} + \frac{(\text{Temp} - \text{Temp}_i)^2}{\sigma_{\text{Temp}}} \right\}, \text{Given Input} : (\text{DO}_{\text{norm}}, \text{Sal}, \text{Temp})
\]

In this equation, \( \text{DO}_{\text{norm}} \) represents a normalized value of dissolved oxygen, \( \text{Sal} \) represents a salinity reading, and \( \text{Temp} \) refers to a temperature value. The values of sigma are calculated as the standard deviation of all values of that particular variable within the historical dataset.

Then the \( k \) best matches are located (with \( k \) to be chosen prior to the model’s execution based upon which value yields the strongest empirical results) and used as a plausible distribution of outcomes known as the ‘similar set’. For each past record in the similar set, it is observed what happened at that location in one day’s time and those results are used as an empirical prediction. Therefore it is important for each member of the training set to contain not only the historical dissolved oxygen value, 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 continuous sampling, when one would know the historical dissolved oxygen reading, but not the value twenty-four hours ahead. For instance, if a continuous sensor is placed in the water for ten days, for the first nine days, we would have access to the dissolved oxygen value at the time of sampling AND the value twenty four hours later.
On the tenth day, we would not know the dissolved oxygen value for twenty-four hours into the future. Consequently, only the first nine days of the ten-day sample can become part of the training set.

In addition to the previous normalized values of dissolved oxygen, salinity, wind 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, as in Equation (11), and thereby account for numerous other factors such as spatial variables of latitude and longitude which are introduced later. Please note, each difference factor (opposed to an overall spatial forecast) with hind-casting, including 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.

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 in Figure 10(a–c), a variable’s importance is quite easily determined by either allowing that one variable to change 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, 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.

**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 24 during the summer of 2005. This will be the location used for hind-casting and model validation.

**Spatial interpolation**

Having validated forecasts at individual locations (as opposed to an overall spatial forecast) 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

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 user-defined coordinates 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.

To calculate the baseline estimates of mean and standard deviation:

\[
\mu_{x,y}^v = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\sum_{j=1}^{m} \frac{1}{d_{ij}}} \right) \mu_{i,j}^v
\]

\[
\forall \left\{ x_{\min} + c_1 \frac{(x_{\max} - x_{\min})}{x_{\text{precision}}} \right\}
\forall \left\{ y_{\min} + c_2 \frac{(y_{\max} - y_{\min})}{y_{\text{precision}}} \right\},
\]

And similarly:

\[
\sigma_{x,y}^v = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\sum_{j=1}^{m} \frac{1}{d_{ij}}} \right) \sigma_{i,j}^v
\]

\[
\forall \left\{ x_{\min} + c_1 \frac{(x_{\max} - x_{\min})}{x_{\text{precision}}} \right\}
\forall \left\{ y_{\min} + c_2 \frac{(y_{\max} - y_{\min})}{y_{\text{precision}}} \right\},
\]

\((x_{\min}, y_{\min})\) and \((x_{\max}, y_{\max})\)—The boundaries of the user-defined rectangular grid
\(x_{\text{precision}}\) and \(y_{\text{precision}}\)—Latitudinal and longitudinal resolution over grid
\(\mu_{x,y}^v\) = Estimated stationary mean for variable \(v\) at point \(x,y\)
\(\sigma_{x,y}^v\) = Estimated stationary standard deviation for variable \(v\) at point \(x,y\)

The historical database consists of discrete stations (Figure 2), each with a contingent of historical data points. For each station, the following is defined:

\(\mu_{x,y}^v\) = Stationary mean for variable \(v\) at station \(i\)
\( \delta^i \) = Stationary standard deviation for variable \( v \) at station \( i \).

And for the spatial baseline estimates:

\( d_{x,y} \) = 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 \( c_1 \) and \( c_2 \) are integers between 0 and \( x_{\text{precision}} \) and 0 and \( y_{\text{precision}} \) (inclusive) respectively. Inspection of the above formulae reveals that all combinations of \( c_1 \) and \( c_2 \) 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, \( |1/(d_{x,y})^2| \) is set to a very large number to ensure that the historical data values at that station are used.

II. Integrating today’s input

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. We calculate \( v_{x,y} \), the value of variable \( v \) and point \( x,y \), using the same construction as Equation (12), replacing \( \mu_{x,y} \) with \( v_{x,y} \) and \( \mu^v \), and all counters \( i \) with \( j \). In this case, instead of \( n \) stations with their own data distributions, there are \( m \) current data points, each containing a vector of relevant independent variable readings.

Next, the elements of the matrix \( v_{x,y} \) are normalized by comparing these values to the baseline distributions at the same locations, as follows:

\[
N^v_{x,y} = \frac{(v_{x,y} - \mu^v)}{\delta^v_{x,y}},
\]

\[
\forall \left\{ x_{\text{min}} + c_1 \frac{(x_{\text{max}} - x_{\text{min}})}{x_{\text{precision}}} \right\}, \left\{ y_{\text{min}} + c_2 \frac{(y_{\text{max}} - y_{\text{min}})}{y_{\text{precision}}} \right\} \quad (14)
\]

where

\( N^v_{x,y} \) = 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:

\( d_{x,y} = \) 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 the section on ‘Isolating parameters that influence DO/ model testing’ using the \( k \)-nearest neighbors approach for all relevant variables.

RESULTS AND DISCUSSION

This section mirrors the progression described in the previous discussion of methodology, showing the findings at each milestone: detrending, determining relevant variables, and creating single and multiple location forecasts. These forecasts are then used to test the hydrodynamic hypothesis by exploring the role of wind. Wind is treated differently than the discussion of salinity and temperature as it is a vector, not a scalar quantity. All of these results are shown only for \( k \)-nearest-neighbor algorithm. Regression trees produced poor results: using a discontinuous algorithm like regression trees for this continuous time-series forecast was less effective. Regression tree results and comparisons with KNN are given by Coopersmith (2008). During the initial analysis of data, some auto-regressive models were constructed. Tests were run using staggered data, for instance, examining the correlation of data points with data at the same location, one hour, two hours, three hours later, etc. However, the correlation coefficients decayed rapidly for any prediction longer than an hour or two using a simple autoregressive model. Replacing the staggered data with moving averages as predictive variables showed little improvement. Therefore the ARMA approach was not pursued further for this application.
Detrending

First, as before, begin with the detrending process. Figure 5 represents the same data as in Figure 4 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 5 and 6, which is then removed from the data.

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 7 and 8. Yet another Fourier transform with two harmonics can be employed to specify this particular equation, facilitating the normalization described in Equation (8). Figure 7 contains all Texas Parks and Wildlife data, with the long-term and annual cycle removed and all data sorted by time of day to illustrate the diurnal cycle alone.

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 7). Moreover, dissolved oxygen levels reach their low points during the early morning hours (Figure 8). This is understandable given the photosynthesis and respiration mechanism for hypoxia described in the introduction. Hours of viable sunlight allow for oxygen to be produced by flora, which is subsequently consumed by fauna in the night-time hours, leaving a minimum just after sunrise when the rate of photosynthetic oxygen production surpasses the rate of consumption. Figure 8 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.

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 24 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. Wind also plays a key role, which is explored separately below due to its two-dimensional nature. Though salinity and temperature play a clear role in the forecasting of future dissolved oxygen levels, some variables either lacked the necessary correlation with DO, or were products of variables already included in our analysis. Conductivity was excluded as a variable because it is largely a function of salinity and temperature and pH was excluded due to its weaker correlation.
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, 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 9 for three different temperature levels.

Figure 9(a–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 9(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.

As shown in both Figures 10 and 11, 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 10, a regression over the scatterplot would yield an insignificant coefficient of correlation, though Figure 11 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 10 was recreated using only data from specific time periods of the day (i.e. only records gathered between 12 and 2 a.m., 2 and 4 a.m., 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.

**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 a single location forecast like the one at station 24 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, which ensures that the forecasts are spatial interpolations rather than extrapolations beyond the data. 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 rectangular grid of latitudinal and longitudinal coordinates and generate a spatial map (see, e.g. Figures 12–14). These figures were generated using dissolved oxygen, salinity, temperature, wind speed and wind direction over the latitudinal and longitudinal ranges illustrated.

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 14), reflecting more uncertainty, as well as lower expected dissolved oxygen levels (Figure 12) are prime candidates for an additional sensor. Based on the results given in Figures 12–14, researchers introduced new sensors (#8, #34, #199, and #202 in Figure 15) in summer 2007, from which they learned that hypoxic risk is far more widespread than initially believed (Figure 15). 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.
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 have not yet been addressed. As mentioned in the introduction and the methodological description of spatial mapping above wind is a variable whose impact must be explored and verified in more detail. High winds can drive more saline water into bottom waters of Corpus Christi Bay from Laguna Madre, increasing
stratification, but can also aerate surface waters. Low winds can cause stagnation, which allows respiration at the bottom to overcome diffusion of oxygen at the surface. Because wind is a two-dimensional 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 24 is used, for a period within the summer of 2005 during which continuous readings were taken every 15 minutes for approximately 20 days.
Figure 16 shows the projections of dissolved oxygen levels with and without the inclusion of wind data (darkest and lightest lines, respectively) as well as the actual observed result. Figure 17 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.

To verify model accuracy, we classified each of our 2,012 readings (one every fifteen minutes from July 6th to July 27th, 2005) from continuous sensor #24 as hypoxic or non-hypoxic based on previously collected historical data. Assuming a threshold probability estimate of 0.7 as a forecast of hypoxia, the model predicts hypoxia for 318 individual readings. Of those, we find that 61.6% of our data points estimated to be hypoxic do show hypoxic behavior, as composed to 17.8% of the sample as a whole. Additionally, of points with probability estimates below 0.7 (the remaining 1694 points), only 9.6% presented hypoxic behavior. Two-proportion z-tests find that the sample of points with hypoxia forecast is more likely to be hypoxic than the overall sample, with a statistically significant p-value less than $10^{-10}$. Furthermore, the sample of points not estimated to be hypoxic ($p < 0.7$) are less likely to be hypoxic than the overall sample with a statistically significant p-value of $< 10^{-10}$.

**CONCLUSIONS**

The analysis indicates that $k$-nearest neighbor algorithms can produce reasonable estimates of the probability of observing a hypoxic event the following day given inputs of dissolved oxygen, salinity, water temperature (Figures
wind speed, and wind direction (shown to be meaningful in Figure 17(a,b)) for the present bay status. Hind-casting at other locations apart from sensor 24 could be useful to provide a basis of comparison—in the current study, sensor 24 was the only continuous sensor able to provide a sufficient stretch of data for this task. 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 near-real-time bay-wide forecast on any coordinate system if sufficient near-real-time sensors are available. Real-time forecasting is entirely possible with incomplete data (i.e. only a few locations), 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 hypoxia phenomenon stretches considerably 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 further research to link mechanistic modeling of hydrodynamics with machine learning algorithms for the purposes of better understanding those mechanisms which drive hypoxia in Corpus Christi bay and elsewhere.

ACKNOWLEDGEMENTS

The authors would like to thank Andrew Collier for programming the visualizations of hypoxic events in Corpus Christi Bay. Thanks are due to Professor David Maidment from the University of Texas at Austin for his insight regarding the process of detrending and sequential normalization. Thanks are given to Professor Ben Hodges and Paula Kulis from the University of Texas at Austin for helping to illustrate wind’s role in governing the dissolved oxygen levels in CC bay. Finally, this work was supported by the National Science Foundation under award number CBET–0609545.

REFERENCES


Texas Parks and Wildlife Department, 4200 Smith School Road, Austin, TX 78744t http://www.tpwd.state.tx.us/. Data requested December 8, 2005 via Mark Fisher, Texas Parks and Wildlife Department, Coastal Fisheries Division, Rockport, Texas, USA.