ABSTRACT

This study aims to create more accurate and efficient near-real-time forecasts of hypoxia that will give researchers advance notice for manual sampling during hypoxic events. Hypoxic or dead zones, which occur when dissolved oxygen levels in water drop below 2 mg/L, are prevalent worldwide. An example of such an hypoxic zone forms intermittently in Corpus Christi Bay (CC Bay), Texas, a USEPA-recognized estuary of national significance. Hypoxia in CC Bay is caused by inflow of hypersaline waters that enter from adjacent bays and estuaries, natural fluctuations in oxygen levels due to the oxygen production-consumption cycle of the aquatic flora and fauna, seasonal fluctuations, and discharges from several wastewater treatment plants.

The hypoxia forecasting method tested in this work involves a suite of data-driven model fusion techniques such as historical scenario modeling and boosting both a k-nearest neighbor (KNN) algorithm and the historical scenario model. Existing data-driven k-nearest neighbor and physics-based valve models are used as the basis for the model fusion. The historical scenario model combines the k-nearest neighbor algorithm with the valve model to predict the probability of hypoxia twenty-four hours ahead. Boosting involves training the model repeatedly on subsets of the training dataset.

The results of the fused models are compared with those of the individual models to test the effectiveness of model fusion in predicting the estuarine conditions. The results showed that the valve model, which has been hitherto computing oxygen profiles, can be extended to forecast probabilities of hypoxia when combined with the k-nearest neighbor algorithm to form the historical scenario model. The findings also show that boosting significantly enhances the performance of the k-nearest neighbor algorithm and the historical scenario model, although further testing on more extensive continuous datasets is needed to verify the findings in other locations. The results show promise for model fusion to be effective for real-time forecasting in hypoxia-affected water bodies.

1. Introduction

There are 169 hypoxic areas spread across every continent and only 13 are in the recovery phase (Selman et al., 2008). In all of these regions, including Corpus Christi Bay (CC Bay), Texas, a bay connected with the Gulf of Mexico that is pivotal to this study, the stressed ecosystems have primarily been a consequence of two reasons – human-induced eutrophication and/or naturally occurring density-induced water column stratification (Diaz, 2001).

The environmental impacts of hypoxia on theecosystems are perceivable. The spectrum of feedback mechanisms that evolved in these ecosystems varies with species and severity of the problem. Certain species in deep waters migrate to different depths in the water column.Others in shallow waters simply migrate to the surface (Montagna and Ritter, 2006; Ekau et al., 2010) to escape the hypoxic hypolimnion, the lowest layer of water that is most susceptible to hypoxia (Galkovskaya and Minyanina, 2005), where they could be vulnerable to sunlight and predators. Hypoxic regions also have lower biodiversity and are home predominantly to low-oxygen-tolerant organisms (Levin, 2003).

The sustainability of the ecosystem therefore depends upon addressing hypoxia related issues. The research on hypoxic conditions in Corpus Christi Bay is a case in point. Daily data collected by sondes placed in CC Bay reflects substantial fluctuations in dissolved oxygen levels just over the course of a week's time. Twenty-four-hour-ahead forecasts are needed for researchers to prepare and send field crews to place the sondes and take grab samples when the likelihood of hypoxia is high. This study aims at creating such near-real-time forecasts of hypoxia while exploring the applicability of model fusion techniques.

Since1969, when model fusion was first implemented (Bates and Granger, 1969), both statistical and machine learning fusion techniques have been created. In this research, a new approach, historical scenario modeling, has been developed and tested for combining two individual models – a k-nearest neighbor (KNN) algorithm (Coopersmith et al., 2010) and a valve hypoxia model (To, 2009). An existing method, boosting, is also applied to test its ability to improve the performance of the above-mentioned KNN algorithm and the historical scenario model.

2. Corpus Christi Bay, Texas – Case study

Two existing individual models and three combined models are evaluated inCorpus Christi Bay to predictthe course of hypoxic events. With an open water surface area of 432.9 km^2 (Flint, 1985), of which 57 km^2 are hypoxic, Corpus Christi Bay is chronically subject to this estuarine condition every summer (Martin and Montagna, 1995; Ritter and Montagna, 1999; Applebaum et al., 2005). Located off the southeast coast of Texas (see Figure 1), the bay is a shallow urban estuary with limited inflow of freshwater and distinct climatic conditions (Montagna and Kalke, 1995). Its only sources of freshwater drainage are the Nueces River and Oso River, and the adjacent Upper Laguna Madre and Oso Bay are sources of highly saline water, up to 60 psu (Islam et al., 2007).

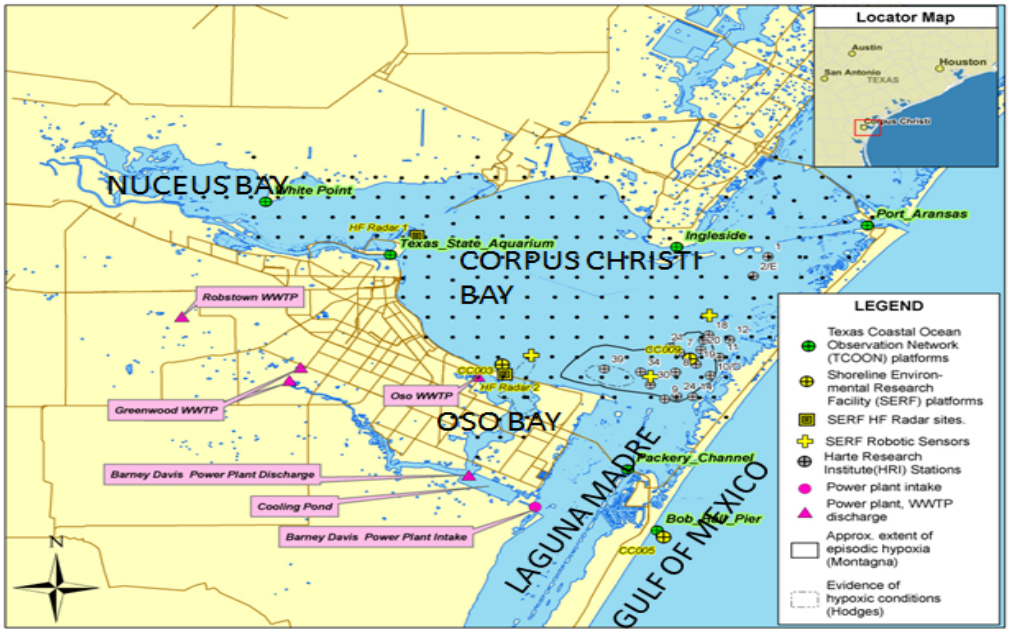

Figure 1 Corpus Christi Bay (To,2009)

3. Methodology

Hypoxia in Corpus Christi Bay is a consequence of salinity-induced, density stratification caused by wind-driven gravity currents, salinity and wind (Hodges et al., 2010); therefore these parameters have been used as the primary variables for modeling hypoxia events in all models.

Additionally, the solubility of oxygen is determined by temperature, which is an additional input variable to the k-nearest neighbor algorithm. Predicting tomorrow's oxygen levels in the k-nearest neighbor algorithm also requires today's dissolved oxygen levels. Finally, to extend the forecasts across the bay from a single location to multiple locations, spatial coordinates are needed as well (Coopersmith et al., 2010). Using these data, model fusion is examined as a tool to improve existing knowledge and understanding of the conditions leading to and resulting from hypoxia.

A brief summary of these individual hypoxia models is followed by a description of the three model fusion approaches below.

3.1 K-nearest neighbor algorithm (Coopersmith et al., 2010)

This machine learning model predicts the likelihood of occurrence of hypoxic events in the Corpus Christi Bay a day ahead.

$$\min\left\{\left[\frac{(DO_{norm}-DO_i)}{\sigma_{DO_{norm}}}\right]^2+\left[\frac{(Sal-Sal_i)}{\sigma_{Sal}}\right]^2+\left[\frac{(Temp-Temp_i)}{\sigma_{Temp}}\right]^2+\left[\frac{(Wind_{NS}-Wind_{NS})}{\sigma_{NS}}\right]^2+\left[\frac{(Wind_{EW}-Wind_{EW})}{\sigma_{EW_m}}\right]^2+\left[\frac{(Lat-Lat_i)}{\sigma_{Lat}}\right]^2+\left[\frac{(Long-Long_i)}{\sigma_{Long}}\right]^2\right\},\forall_i$$

Equation (1)

where the input variables at point (x, y), the coordinates of the desired forecast location are: DOnorm = dissolved oxygen (mg/L), Sal = salinity (PSU), Temp = temperature (oC), WindNS = wind in North-South direction (m/s), WindEW = wind in East-West direction (m/s), Lat = latitude; and Long = longitude. Historical occurrences in the dataset are represented by: DOi = dissolved oxygen (mg/L) at point i, Sali = salinity at point i, Temp = temperature at point i, Wind NSi = wind in North-South direction at point i, Wind EWi = wind in East-West direction at point i, Lati = latitude of point i, and Longi = longitude of point i. The corresponding sigma variables (xxx, etc.) are the variance of each variable, which normalizes each distance term to avoid scaling problems.

The KNN model serves to establish a correlation between these inputs, which were the parameters most correlated with oxygen levels (Coopersmith et al., 2010), and the output that is a

probability estimate of dissolved oxygen twenty-four hours later. Using the distance-metric function given in equation 1, the algorithm classifies nearest neighbors in parameter space that are most likely to be hypoxic in twenty-four hours. The model is described in detail by Coopersmith et al (2010).

3.2 Valve model (To, 2009)

The 2-D model predicts salinity and dissolved oxygen concentrations over the time and distance traversed by the gravity currents in the presence of wind events only. The inputs to this model are wind data, to determine the opening and closing of the valve, and water quality data, i.e., salinity and dissolved oxygen, for both Corpus Christi Bay and Upper Laguna Madre. The model also requires parameters characterizing the formation and transport of currents, including the ranges of wind speed and direction causing gravity currents, average gravity current speed, net oxygen demand rate, and the initial thickness of the gravity current. However, the domain of the model is restricted to the southeast corner of Corpus Christi Bay, which receives the hypersaline inflows from Upper Laguna Madre, which is the only source of gravity currents in the area (To, 2009). The oxygen profile is calculated assuming the oxygen depletion rate within the gravity current to be a constant 0.18 mg/L/hr (Hodges et al., 2010).

3.3 Historical scenario modeling

The historical scenario model combines the k-nearest neighbor model with the valve model to forecast the probability of hypoxia twenty-four hours ahead. The valve model traces the path of a single gravity current from the time it originates at the mouth of Upper Laguna Madre and flows into Corpus Christi Bay. The combined historical scenario model tracks several gravity plumes released into Corpus Christi Bay on consecutive days, with each plume generated by using the k-nearest neighbor model to identify the most similar historical scenarios to the current conditions. Essentially, this turns the deterministic valve model into a probabilistic one, enabling it to provide probability estimates of dissolved oxygen at individual data stations in the region where the gravity plumes occur. For each historical scenario, the valve model forecasts dissolved oxygen concentrations at sampling locations in CC Bay that are within the domain of the valve model.

3.4 Boosting the k-nearest neighbor algorithm

KNN boosting is performed iteratively. A single iteration involves training the KNN on the dataset to find five hundred nearest neighbors to the given set of conditions (the query instance), comprising dissolved oxygen, salinity, temperature, wind, latitude and longitude. Then, for each of these neighbors obtained from the first iteration, the weight term (w_i), which is initialized to zero, is updated. For this purpose, a weight update term λ is used. The criterion for weight adjustment is whether the value of dissolved oxygen of the neighbor instance is ±0.5 within the range of that of the dissolved oxygen of the query instance. The weight is then increased by a factor of $\lambda/(1+e\text{-}w_i^t)*d$, where $(1+e\text{-}w_i^t)$ is the sigmoid function, for each instance that is in the correct class and decreased by the same factor for an instance that is misclassified. The sigmoid function is incorporated to avoid drastic increase or decrease in the weight modification term (Neo, 2007).

At the end of an iteration, a modified dataset is created which has weights on the k historical datapoints obtained from the previous iteration. The total number of iterations, T, is halted when all neighbor instances are classified correctly. The resulting T classifier models are then combined to form one composite boosted model.

3.5 Boosting the historical scenario model

As the name suggests, this method is a continuation and a combination of the two previous methods – historical scenario modeling and boosting the k-nearest neighbors model. As in historical scenario modeling, the KNN is run for seven continuous days of data. For each day's input vector of variables (dissolved oxygen, salinity, temperature, wind, latitude and longitude), the KNN algorithm is boosted until the k most accurate historical scenarios are obtained. These historical scenarios are then used as inputs to the valve model. The rest of the process is the same as that described for the historical scenario modeling.

4. Results

The forecasts of the three best-performing models - k-nearest neighbor algorithm, boosted k-nearest neighbor model, and boosted historical scenario model - for 23 July 2006 are compared with the actual dissolved oxygen values at the corresponding stations on that day in Figure 2.

Data are available at stations 2, 19, and 24. In the week of 18 July – 25 July 2006 when continuous data were collected, 23 July 2006 was observed to be hypoxic. Hence, the five days of data prior to 23 July 2006, i.e., 07/18/2006 to 07/22/2006, were used to compare the three models, which were the only dates in the available data set with five continuous days of data that are required for the valve model to fully transport a plume from the mouth of the bay to the end of the hypoxic region.

At Station 2, where hypoxia never occurs, boosted k-nearest neighbor algorithm (2 percent) predicts a lower probability of hypoxia than the k-nearest neighbor algorithm (7 percent). The boosted historical scenario model performs best at Station 19, but does not perform well at Station 24. This implies that the boosted historical scenario model is only accurate in the region for which the valve model is designed to predict hypoxia, from 2000 m to 5000 m from the mouth of Laguna Madre. In this case, Station 24 is at 1432.8 m and Station 19 is at 4700 m from the mouth of Laguna Madre.

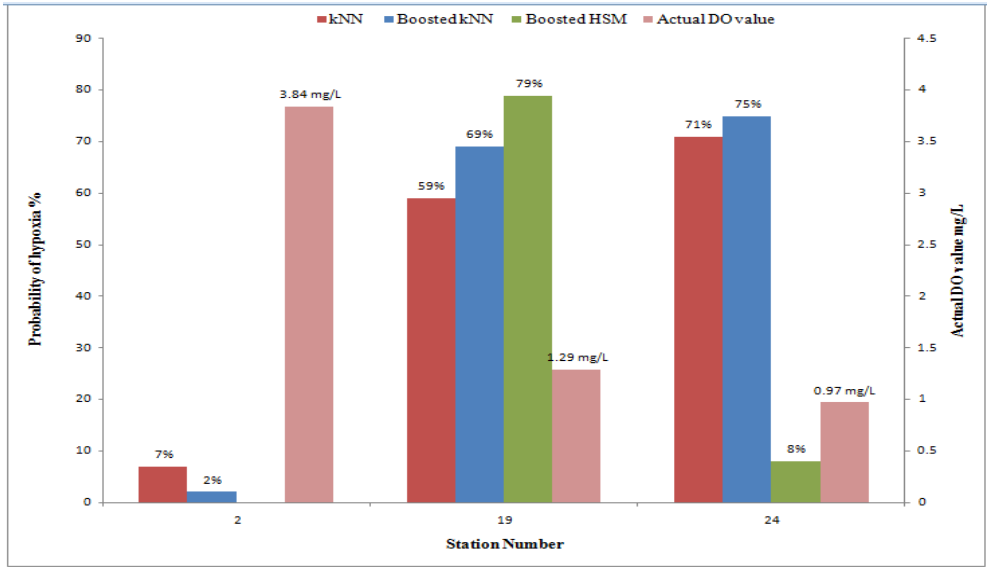

Figure 2 Forecasts of hypoxia versus actual dissolved oxygen values on 23-Jul-2006

5. Conclusion

The two major contributions of this study are the following. First, model fusion added a probabilistic dimension to the otherwise physics-based valve model, combining the strengths of the k-nearest neighbor algorithm and the valve model in the process. Instead of tracking a single gravity plume, the combined model can track several plumes being issued continuously, more

closely simulating the natural phenomenon, and providing forecast probabilities of hypoxia twenty-four hours ahead. Second, boosting improved the performance of both the k-nearest neighbor algorithm and the historical scenario model. The boosted historical scenario model is most accurate in the specific forecasting region for which it was designed, while the boosted KNN model provides reasonable forecasts at all of the stations.Considering the paucity of data available on five consecutive days, as required for the valve model, the boosted KNN model seems to be the most promising approach for further investigation with more extensive datasets.

Further research to extend this work could include the following. Instead of boosting the k-nearest neighbor algorithm directly, other approaches such as boosting by instance selection (García-Pedrajas, 2009) might be tested. Furthermore, the two assumptions in the valve model – a constant gravity current speed and a constant oxygen depletion rate – are simplifications of the complex physical and biological processes in the Bay and deserve closer study. Physical models could be created that would take into consideration wind speed, wind direction, difference in densities of the plume and its ambient fluid, slope of the bay bottom, and plume dissipation. Also, it would be useful to compute the actual oxygen depletion rate with the help of biological models that would examine diurnal fluctuations due to photosynthesis and respiration (To, 2009). Overcoming these limitations of the valve model could prove to be of value in increasing the accuracy of the fused historical scenario model, but additional data would be needed to support increased complexity in the models.

6. References

Applebaum, S., Montagna, P.A., and Ritter, C., 2005. Status and trends of dissolved oxygen in Corpus Christi Bay, Texas, U.S.A. *Environmental Monitoring and Assessment*, Vol. 107, pp. 297-311.

Bates, J. M., and Granger, C.W.J. 1969.The combination of forecasts.*Operational Research Quarterly*, Vol. 20, No.4, 451-468.

Coopersmith, E., Minsker, B.S., and Montagna, P., 2010. Understanding and forecasting hypoxia using machine learning algorithms. *Journal of Hydroinformatics*, in press.

Diaz, R.J., 2001. Overview of hypoxia around the world.*Journal of Environmental Quality*, Vol. 30, No. 2, pp. 275-281.

Ekau, W., Auel, H., P□rtner, H.O., and Gilbert, D., 2010.Impacts of hypoxia on the structure and processes in pelagic communities (zooplankton, macro-invertebrates and fish).*Biogeosciences*, Vol. 7, pp. 1669-1699.

García-Pedrajas, N., 2009.Constructing ensembles of classifiers by means of weighted instance selection.*IEEE Transactions on Neural Networks*, Vol. 20, No. 2, pp. 258-277.

Hodges, B.R., Furnans, J.E., and Kulis, P., 2010. "Case Study: A thin-layer gravity current with implications for desalination brine disposal,".*ASCE Journal of Hydraulic Engineering*, in press.

Levin, L.A, 2003. Oxygen minimum zone benthos: Adaptation and community response to hypoxia. *Annual Review of Oceanography and Marine Biology*, Vol. 41, pp. 1-45.

Montagna, P.A., and Kalke, .R.D., 1995.Ecology of infaunal Mollusca in south Texas estuaries.*American Malacological Bulletin*, Vol. 11, pp. 163-175.

Montagna, P., and Ritter, C., 2006. Direct and indirect effects of hypoxia on benthos in Corpus Christi Bay, Texas, U.S.A. *Journal of Experimental Marine Biology and Ecology*, Vol. 330, No. 1, pp. 119-131.

Neo, T.K.C., 2007. A direct boosting algorithm for the k-nearest neighbor classifier via local warping of the distance-metric.MS thesis, Brigham Young University.

Ritter, C., and Montagna, P.A., 1999. Seasonal hypoxica and models of benthic response in a Texas Bay.*Estuaries*, Vol. 22, No. 1, pp. 7-20.

Rokach, L., 2010. Ensemble-based classifiers.*Artificial Intelligence Review*, Vol. 33, pp. 1-39.

Selman, M., Greenhalgh, S., Diaz, R., and Sugg, Z., 2008.*Water Quality: Eutrophication and Hypoxia*, No. 1, pp. 1-6.

To, E.S.C, 2009.Hypoxia modeling in Corpus Christi Bay using a hydrologic information system. PhD diss., University of Texas at Austin, May 2009.