MODEL FUSION FOR IMPROVING HYPOXIA FORECASTS IN CORPUS CHRISTI BAY, TX, USA: A STUDY OF BOOSTING AND HISTORICAL SCENARIO MODELING

BY

INDU CHINTA

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, 2010

Urbana, Illinois

Adviser:

Professor Barbara Minsker

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.

To Mom, Dad and to Suma

ACKNOWLEDGEMENTS

I am deeply blessed to have the guidance of Professor Barbara Minsker and am grateful to her beyond words. She gave me the best break as an international learner and “independent researcher” in an altogether different field of Environmental Engineering.

I could not have written the first sentence of this thesis without David Hill and Evan Coopersmith. David spent long hours familiarizing me with data mining concepts. A regular course or a professor could not have taught me as much. Evan’s advice is without peer. He is a real authority in machine learning and his thesis, my bible.

I was quite wary about programming which forms a significant portion of my research. I am especially thankful to Ankeeth Ved, who helped me tirelessly translate my logic into the several programming languages I used to develop my models. Thanks also to Dr. Ernest Sin Chit To who patiently answered and endured my myriad phone calls and emails from as distant as Texas.

A special thanks to David Waterman for introducing me to AutoCAD and ensuring I presented my results most accurately and precisely.

Mary Pearson and Marvis Orzek are two of the finest people I have known. At the U of I, I could not have asked for a better research group – Andrea Zimmer, Brian Payne, Xiaowen Wu and Chuan Li. Thanks all!

Finally, all of this would not have been possible without Suma. A terrific doctor and even wonderful sister. Thank you!

TABLE OF CONTENTS

1. Introduction…………..……………………...……………………………………………...1

2. Corpus Christi Bay, Texas – Case study………...……………………………………………..4

3. Methodology…………..……………………………………………………………………8

   3.1 K-nearest neighbor algorithm..………...…………………………………………………..9

   3.2 Valve model...………………………………………………………………………….10

   3.3 Historical scenario modeling...…………………………………………………………...11

   3.4 Boosting the k-nearest neighbor algorithm……………………………………...………..14

   3.5 Boosting the historical scenario model...…………………………………………………15

4. Results……………………...……………………………………………………………16

5. Conclusion……………...…………………………………………………………………22

References………………………………………………………………………………….24

1. Introduction

The World Resources Institute identified several hundreds of eutrophic and hypoxic areas around the world. Out of these, as many as 169 are hypoxic and only 13 are in the recovery phase (Selman et al., 2008). These hypoxic areas, also called dead zones, are spread across every continent and particularly concentrated along the Asian, European, and North American coastlines. Hypoxia occurs in areas ranging from Chinhae Bay, Korea (Lim et al., 2006) and Mikawa Bay, Japan (Suzuki, 2001) in the eastern hemisphere, to the Baltic Sea, where the hypoxic zone has increased fourfold from the 1960s (Jonsson et al., 1990; Zillén et al., 2008), and the extensively studied and the once largest dead zone in the Black Sea (Mee, 1992; Cociasu et al., 1996; Daskalov, 2003; Langmead, 2009), the Northern Adriatc (Justíc, 1987), and Kattegat (Baden et al., 1990) in the central hemisphere, to the now largest dead zone in the Northern hemisphere, the Gulf of Mexico (Rabalais et al., 2002; Bianchi, 2010). 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 the affected ecosystems are perceivable, if not calamitous. 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 where they can avoid the hypoxic zone. 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; Zhang et al., 2010).

The sustainability of the ecosystem thus depends upon understanding and addressing the issues related to hypoxia. 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 by fusing existing models using model fusion techniques.

Model fusion started as early as 1969 with Bates and Granger combining forecasts of airline passenger data to “form a composite set of forecasts” that would lower the mean-square error compared to the individual forecasts. Some of the other earlier contributions in this area include Dickinson (1971) and Newbold and Granger (1974). Later, statistical techniques such as Bayesian model averaging were introduced by Bunn (1975, 1977) and Bordley (1982), which were pursued by Duan et al. (2007), Raftery et al. (1997), Vrugt and Robinson (2007), and more recently by Hsu et al. (2009) to predict “daily watershed streamflow”

Machine learning strategies are as popular as the statistical techniques described above. Simple averaging, neural networks, fuzzy logic, M5 model trees, and instance-based learning were illustrated for fusing flow forecasting models built over the River Ouse catchment in the United Kingdom (See, 2008). Prior to this, See and Abrahart, in 2001, tested four data fusion “experiments” - combining mean and median of four individual forecasts and two others using neural networks - and observed that the two neural network fusion methods produced better results. Ajami et al. (2006) shed light on the use of simple averaging method, multimodel superensemble method, modified multimodel superensemble method, and the weighted averaging method for streamflow forecasting. Simple averaging method, weighted averaging method, and neural network method have also been employed for integrating rainfall-runoff models (Shamseldin and O’Connor, 1999, and, Shamseldin et al., 1997).

More closely related to this study is the work of Coulibaly et al. (2005), wherein the authors combined “three dynamically different” hydrological models – a nearest neighbor model, conceptual model, and artificial neural network model - using the improved weighted average method. This type of combination provided an improved “4-day ahead prediction” of daily reservoir inflow.

Apart from hydrology, model fusion has also been extensively applied in geophysics and meteorology (Chakraborthy et al., 2007) for combining climate models. Dietrich et al. (2008) transformed meteorological ensemble forecasts into “discharge ensemble forecasts of rainfall-runoff models” to effectively predict flood situations. Furthermore, “multiple climate and hydrological models” were integrated for predicting uncertainty in streamflow using techniques like pooling and linear regression weighting (Block et al., 2009).

This research explores the applicability of model fusion techniques for hypoxia forecasting. 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.

The study is structured as follows. Section Two is a description of Corpus Christi Bay, which is the focus of this research for testing the model fusion approaches. Section Three discusses the individual models and methods that will be applied to fuse them. Section Four presents the results of these approaches and finally Section Five contains conclusions.

2. Corpus Christi Bay, Texas – Case study

Two existing individual models and three combined models are evaluated over Corpus Christi Bay to interpret the 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, 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). The location of CC Bay is shown in Figure 2.1.

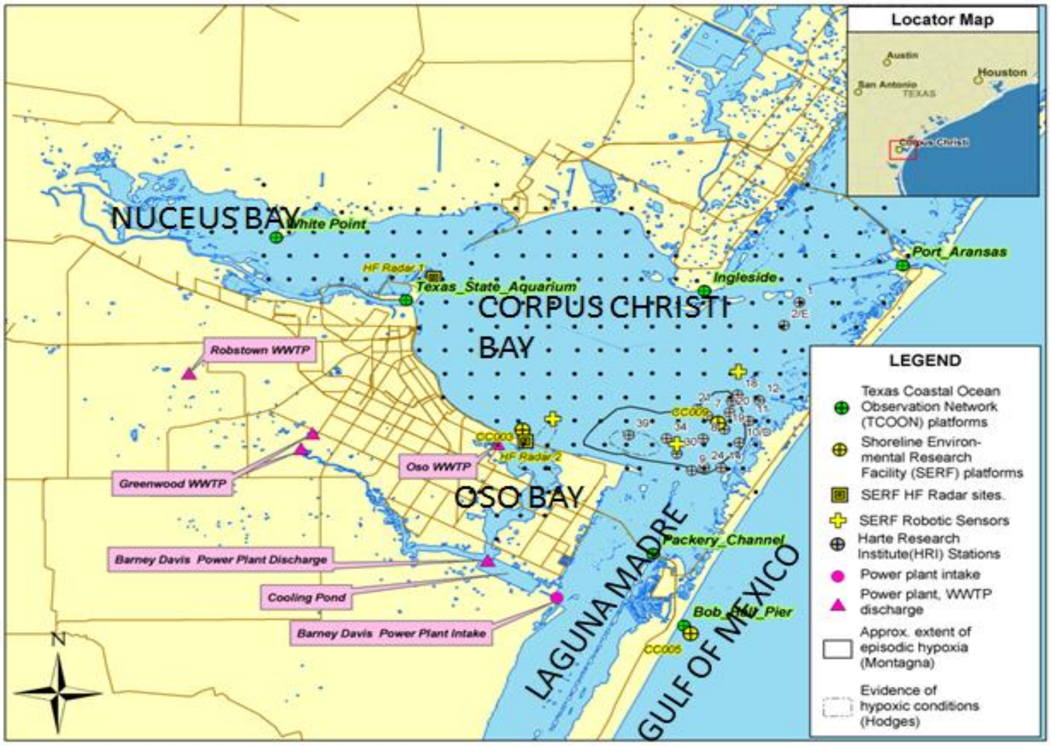
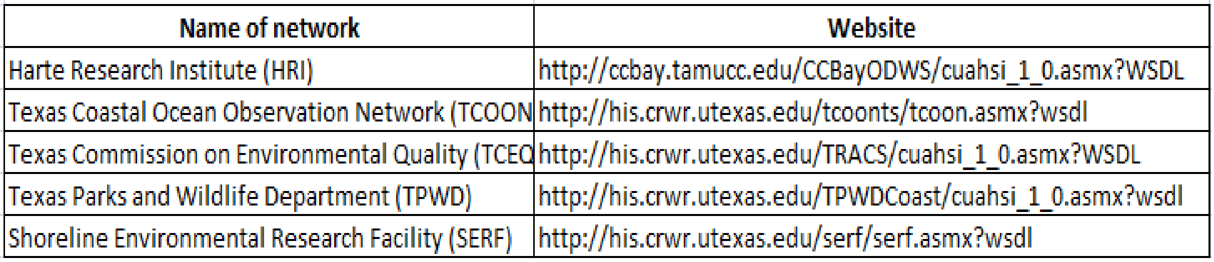


Figure 2.1 Corpus Christi Bay (To,2009)

Hypoxia is a result of eutrophication or excessive nutrient discharges in most aquatic systems around the world (Conley, 2009). However, eutrophication has been ruled out as a major cause of hypoxia in Corpus Christi Bay over the past fourteen years for the bay indicates a declining trend in freshwater inflow and an insignificant variation in nutrient levels (Islam, 2006). For this reason, hypoxia is believed to be primarily correlated with the entrance of hypersaline waters from Upper Laguna Madre (Figure 2.1), driven by southeasterly winds in the summer months (Hodges and Furnans, 2007). The density of these waters is higher than the water in Corpus Christi Bay which results in the formation of buoyant plumes, also called gravity currents (Kulis and Hodges, 2005). These plumes move into Corpus Christi Bay because of southeasterly winds that lead to water column stratification, which results in benthic hypoxia. Additionally, discharges from nearby wastewater treatment plants (WWTPs) and the Barney Davis Power Plant that are transported through Oso Bay into Corpus Christi Bay may compound the problem.
The area most prone to hypoxia is the southeast region of Corpus Christi Bay where the bay meets the Upper Laguna Madre and Oso Bay (Ritter and Montagna, 1999). Systems for monitoring water quality parameters that are correlated to the occurrence of hypoxia have been installed in various locations around the bay to better understand the causes and impacts of hypoxia through discrete sampling (Islam, 2007). To advance real-time observations and to gain insight into the changes occurring in the bay during hypoxic events, robotic profilers measured not only multiple water quality parameters in real time but also at multiple depths at high temporal resolution (Islam et al., 2010). These systems have since been decommissioned and are not used in this work.

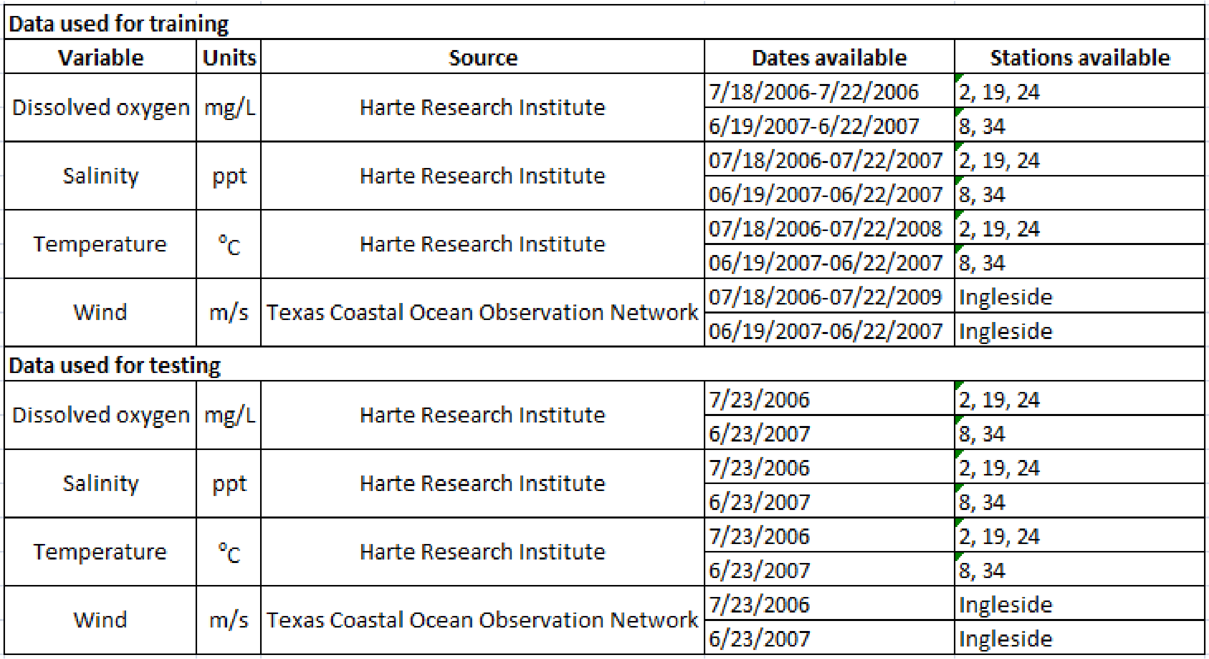
The data used in this work comprise dissolved oxygen (mg/L), salinity (Practical Salinity Units), temperature (^{o}C), wind speed (m/s) and wind direction (degrees clockwise from North) for Corpus Christi Bay, and dissolved oxygen (mg/L) and salinity (PSU) for Upper Laguna Madre. They are derived from the following sensor networks: (1) Harte Research Institute (HRI) grab samples of water quality (dissolved oxygen, salinity, temperature), (2) Shoreline Environmental Research Facility (SERF) continuous samples of water quality and surface current velocities, (3) Texas Coastal Ocean Observation Network (TCOON) wind and tide data along the coast of Texas, (4) Texas Commission on Environmental Quality (TCEQ) water quality data, and (5) Texas Parks and Wildlife Department (TPWD) grab samples of water quality. The websites of the networks are given in Table 2.1. These data forming the historical dataset are available for the summer months, June to August, from 1999 to 2007. Table 2.2 presents the data used to train and test the models in this study. Sampling stations given in Table 2.2 are shown in Figure 2.2.

Table 2.1 Web addresses of sensor networks in Corpus Christi Bay (To, 2009)



Name of network	Website
Harte Research Institute (HRI)	http://ccbay.tamucc.edu/CCBayODWS/cuahsi_1_0.asmx?WSDL
Texas Coastal Ocean Observation Network (TCOON	http://his.crwr.utexas.edu/tcoonts/tcoon.asmx?wsdl
Texas Commission on Environmental Quality (TCEQ	http://his.crwr.utexas.edu/TRACS/cuahsi_1_0.asmx?WSDL
Texas Parks and Wildlife Department (TPWD)	http://his.crwr.utexas.edu/TPWDCoast/cuahsi_1_0.asmx?wsdl
Shoreline Environmental Research Facility (SERF)	http://his.crwr.utexas.edu/serf/serf.asmx?wsdl

Table 2.2 Data used for modeling



Data used for training				
Variable	**Units**	**Source**	**Dates available**	**Stations available**
Dissolved oxygen	mg/L	Harte Research Institute	7/18/2006-7/22/2006	2, 19, 24
			6/19/2007-6/22/2007	8, 34
Salinity	ppt	Harte Research Institute	07/18/2006-07/22/2007	2, 19, 24
			06/19/2007-06/22/2007	8, 34
Temperature	°C	Harte Research Institute	07/18/2006-07/22/2008	2, 19, 24
			06/19/2007-06/22/2007	8, 34
Wind	m/s	Texas Coastal Ocean Observation Network	07/18/2006-07/22/2009	Ingleside
			06/19/2007-06/22/2007	Ingleside
Data used for testing				
Dissolved oxygen	mg/L	Harte Research Institute	7/23/2006	2, 19, 24
			6/23/2007	8, 34
Salinity	ppt	Harte Research Institute	7/23/2006	2, 19, 24
			6/23/2007	8, 34
Temperature	°C	Harte Research Institute	7/23/2006	2, 19, 24
			6/23/2007	8, 34
Wind	m/s	Texas Coastal Ocean Observation Network	7/23/2006	Ingleside
			6/23/2007	Ingleside

Two forms of historical data have been used in this research: grab samples and continuous samples. Grab samples, which represent the majority of available records, are collected by hand at pre-defined locations, depths, and time intervals. Continuous samples are collected from stationary sensors over a week to fortnight, at a regular interval of fifteen minutes. The grab sampling locations are indicated in black and continuous locations are in white as shown in figure 2.2.

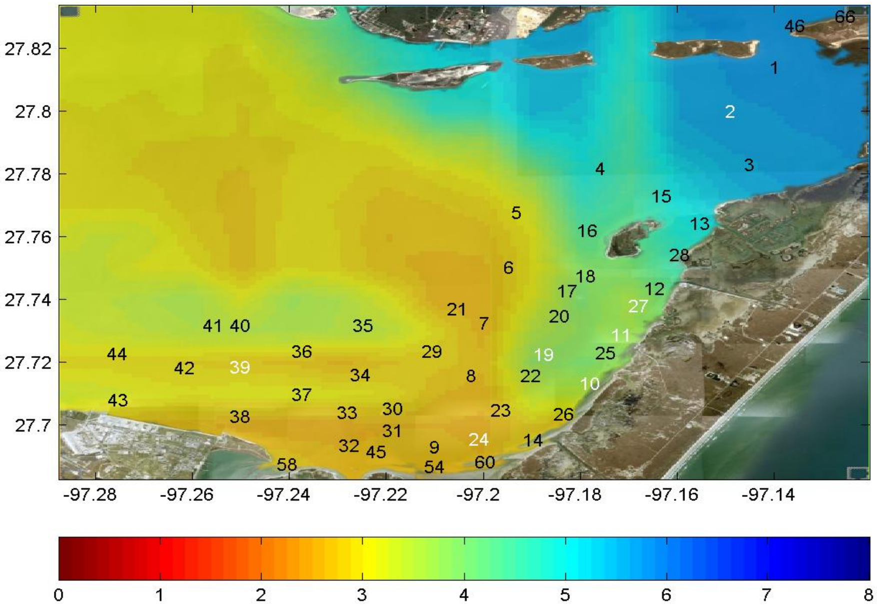


Figure 2.2 Sensor locations in Corpus Christi Bay (Coopersmith et al., 2010)

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.

Model fusion is the process of building a single robust model by integrating multiple models for improving the accuracy of prediction and performance of individual models (Rokach, 2010). It also bears terms such as combination of multiple classifiers (Lam and Suen, 1995; Sharkey, 1999; Woods et al., 1997; Xu et al., 1992; Kittler et al., 1998), classifier fusion (Cho and Kim, 1995; Gader et al., 1996; Grabisch and Dispot, 1992; Keller et al. 1994; Bloch, 1996), and mixture of experts (Jacobs et al., 1991; Jacobs, 1995; Jordan and Xu, 1995; Nowlan and Hinton, 1991) to name a few.

In this study, the two existing hypoxia models for CCBay, k-nearest neighbor (Coopersmith et al., 2010) and valve (To, 2009) models, are fused using two different approaches termed historical scenario modeling and boosting. Boosting is applied to both the original k-nearest neighbor model and the fused historical scenario model. These two techniques of model fusion are representative of the two major categories of model fusion – (i) combining individual models (historical scenario modeling), and (ii) combining using different training sets (Kuncheva, 2001) (boosting the k-NN and historical scenario model).

The dataset available did not contain data at all locations on all days. Rather, very few stations had data collected on overlapping dates. While the k-nearest neighbor algorithm requires a minimum of three inputs from three different locations, the historical scenario model requires continuous days of data to track gravity plumes released on consecutive days. In order to meet these requirements, two sets of contiguous data were identified for testing the models. The data were from three stations during 18-23 July 2006 and from two stations during 19-23 June 2007.

A brief summary of the 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.

Equation (3.1.1)

$$\min\left\{\begin{array}{l}\left[\frac{\left(DO_{norm}-DO_{i}\right)}{\sigma_{DO_{norm}}}\right]^{2}+\left[\frac{\left(Sal-Sal_{i}\right)}{\sigma_{Sal}}\right]^{2}+\left[\frac{\left(Temp\ -Temp_{i}\right)}{\sigma_{Temp}}\right]^{2}+\\ \left[\frac{\left(Wind_{NS}-Wind_{NS}\right)}{\sigma_{NS}}\right]^{2}+\left[\frac{\left(Wind_{EW}-Wind_{EW}\right)}{\sigma_{EW_{m}}}\right]^{2}+\left[\frac{\left(Lat-Lat_{i}\right)}{\sigma_{Lat}}\right]^{2}+\left[\frac{\left(Long-Long_{i}\right)}{\sigma_{Long}}\right]^{2}\end{array}\right\},\forall_{i}$$

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 3.1.1, 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 valve model belongs to the category of physics-based models. It is built on the gravity-current theory, proposed by Dr. Ben Hodges and Paula Kulis at the University of Texas at Austin (Kulis and Hodges, 2005). The gravity current theory states that hypoxia occurs as a consequence of the entrance of hypersaline waters, which are also highly dense, from Laguna Madre. These currents are driven forward and downward into Corpus Christi Bay by southeasterly winds and their own weight in the general direction shown in Figure 3.1. As To (2009) observes, there is a difference in energy in the hypersaline water of the gravity current and the less saline water above the current. Hence, transfer of oxygen from the layer above the current to the layer below is negatively affected. With a reduced supply of oxygen, in the course of time, the bottom layer of the bay becomes hypoxic.

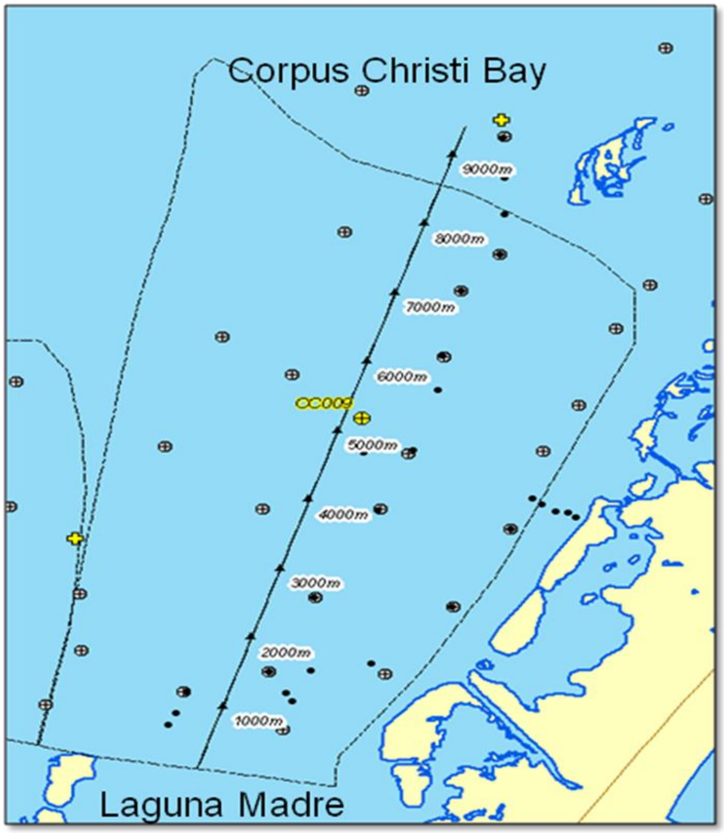


Figure 3.1 Domain of the valve model (To, 2009)

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 valve, representative of a plug, is turned on only when winds of a certain specified speed and direction occur, which leads to release of a gravity current (To, 2009). 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 (Figure 3.2.1), 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). This model is restricted to tracking the "fate and transport" of the gravity currents as they develop. Hypoxia was predicted to occur predominantly from 2,000m to 5,000m from the mouth of Laguna Madre (To, 2009).

The domain of the valve model will also be the domain of the subsequent fused models, as it is the only area that can be analyzed using both models.

3.3 Historical scenario modeling

The historical scenario modeling combines the k-nearest neighbor model with the valve model to forecast the probability of hypoxia twenty-four hours ahead. As discussed in the previous section, 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 model tracks several gravity plumes released into Corpus Christi Bay on consecutive days, with each plume generated from historical scenarios found with the k-nearest neighbor model. 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.

Using this approach, a step-by-step description of historical scenario modeling follows. First, the k-nearest neighbor model is run for input vectors of data from consecutive days comprising normalized (detrended or time-independent) values of dissolved oxygen, salinity, temperature and wind. The ‘k’ best matches, or historical scenarios, for each of these days are then identified from the historical dataset used to calibrate the model. These historical scenarios then serve as a set of plausible conditions for generating gravity currents with the valve model.

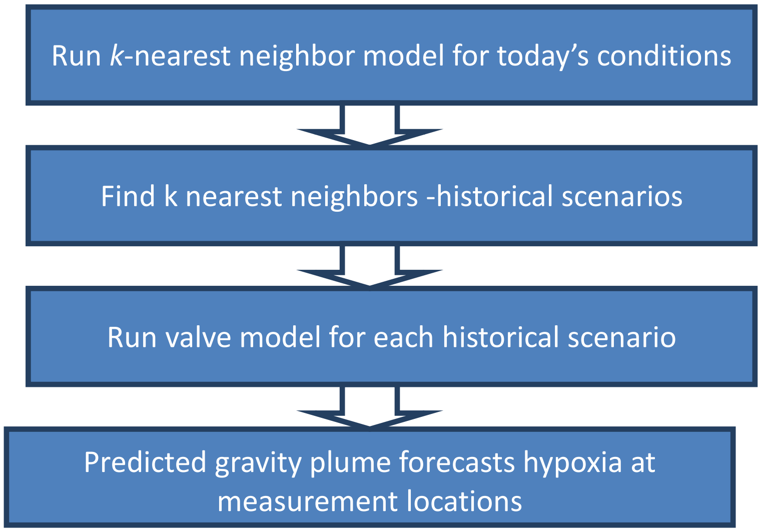

Run *k*-nearest neighbor model for today’s conditions
Find k nearest neighbors -historical scenarios
Run valve model for each historical scenario
Predicted gravity plume forecasts hypoxia at measurement locations

Figure 3.2 Algorithm for historical scenario modeling

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 (Figure 3.3).

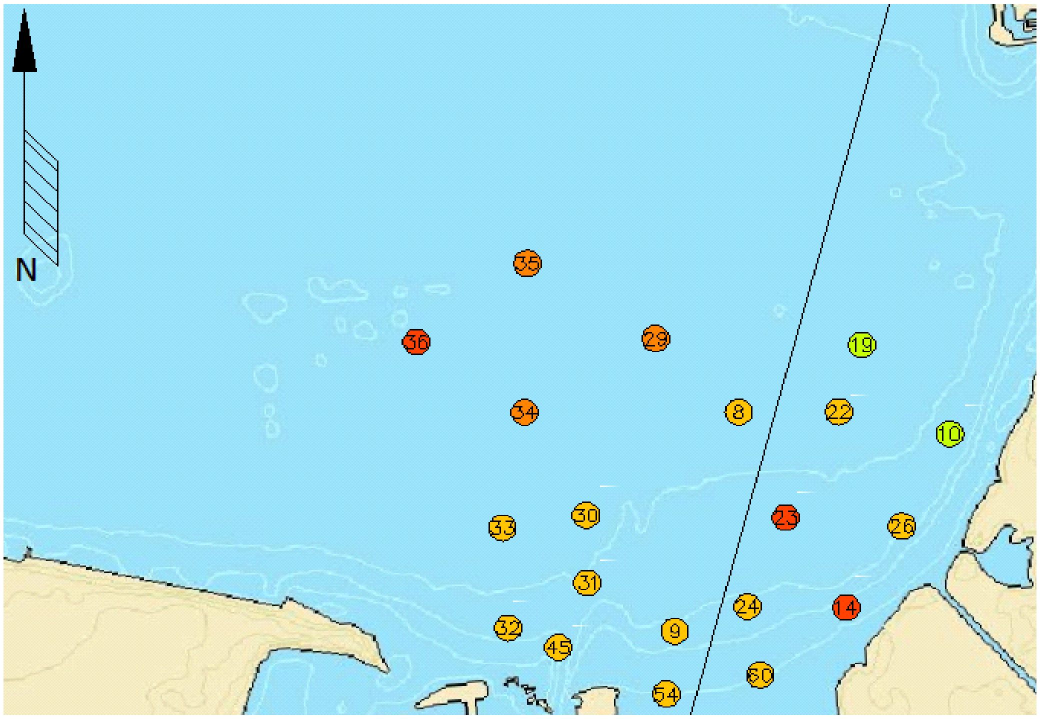


Figure 3.3 Stations within the domain of the valve model

The valve model assumes the speed of the gravity current to be approximately 1 km/day. Hence, in every twenty-four hours, a plume travels about 1000 m. This implies that a single day's input can predict up to 1000 m, and consequently, the total spatial length of prediction of the fused model is determined by the total number of days used to compute historical scenarios from the k-nearest neighbor algorithm. This is explained in the example in the following paragraphs. The stations are projected on to the transect line along which the gravity currents are transported (seen in Figure 3.2.1), since the valve model is designed to compute the variation in dissolved oxygen only along the length of the line.

This limitation was taken into consideration while fusing the models. For example, the study examined hypoxia on 23 July 2006 over a length of 5000m. The data pertaining to five successive days, 18 - 22 July 2006, were used to run the k-nearest neighbor algorithm and find historical scenarios. These historical scenarios were then used as inputs to the valve model.

Since the gravity plume travels approximately one kilometer per day, a plume that originated on 18 July 2006 would be at 5000m on 23 July 2006, a plume that originated on 19 July 2006 would be at 4000m on 23 July 2006, and so on. Hence, for consistency in forecasts for 23 July 2006 and comparability with the k-nearest neighbor algorithm, forecasts at 1000m would correspond to output from 22 July 2006, those at 2000m would correspond to output from 21 July 2006, and so on.

3.4 Boosting the k-nearest neighbor algorithm

Boosting the k-nearest neighbor model is the second fusion method and an example of the second class of model fusion, combining using different training sets. Although there are several means to boost the k-NN classifier (Amores et al.,2006; García-Pedrajas and Ortiz-Boyer,2008; García-Pedrajas, 2009; Yang et al., 2009), a direct boosting algorithm is chosen that trains the KNN model repeatedly on a dataset and produces an ensemble of classifier models (Neo, 2007). Direct boosting was chosen because it is the most generally applicable and widely applied boosting 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, or the query instance, comprising dissolved oxygen, salinity, temperature, wind, latitude and longitude. Then, at 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 same as that described for the historical scenario modeling.

4. Results

The results of applying the approaches described in the previous sections to the Corpus Christi case study, before and after model fusion, are examined in this section.

Figure 4.1 shows KNN forecasts of the probability of hypoxia twenty-four hours ahead on July 23, 2006. Note that only those stations within the domain of the valve model are shown to enable comparison among all of the models.

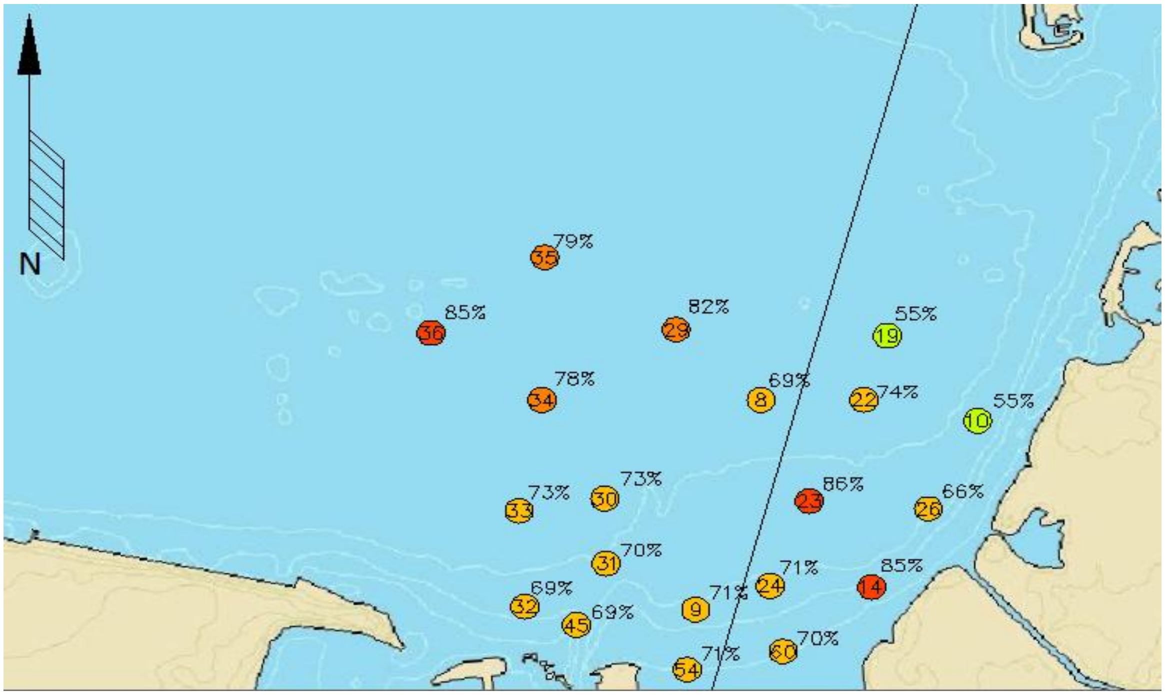


Figure 4.1 K-nearest neighbor algorithm 23-Jul-2006

The results from the valve model before model fusion consist of salinity and dissolved oxygen profiles of the gravity currents that originate only from Upper Laguna Madre, as shown in Figure 4.2. These profiles are computed for the one-dimensional length and depth traversed by the current in the bay along the transepts shown.

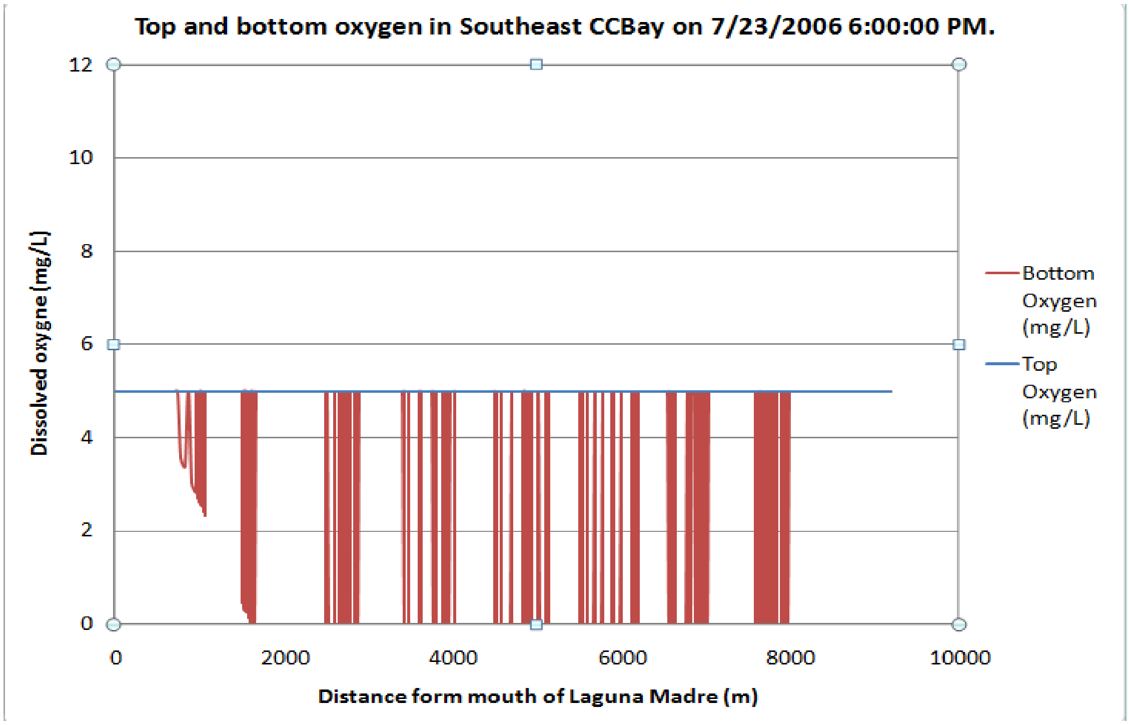


Figure 4.2 Dissolved oxygen profile for 23 July 2006 (To, 2009)

The forecasts from combining the k-nearest neighbor algorithm and the valve model, i.e., the historical scenario model, are given in Figure 4.3 for the same date.

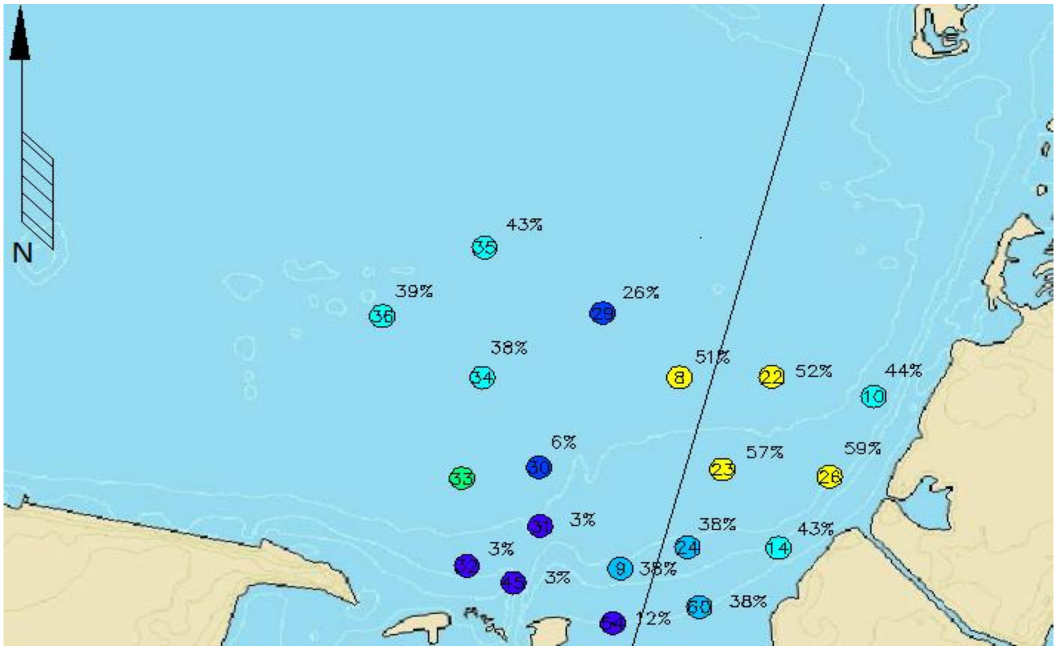


Figure 4.3 Historical scenario model 23-Jul-2006

Comparing Figures 4.1 and 4.4, before and after boosting the k-nearest neighbor algorithm, it is clear that the forecasts of hypoxia are substantially lower with boosting, which is more accurate given that there is no hypoxia. For instance, the forecast for the northern most portion of the bay (where Station 2 lies) is 2 percent (not shown in the figure), wheareas KNN predicts 7 percent. Hypoxia never occurs at this station, thus the boosted k-nearest neighbor model represents an improved forecast. It was also observed that the majority of gains from boosting occur in the first three iterations and the overall accuracy does not improve further beyond three iterations, therefore training can stop at T=3 iterations.. Where hypoxia occurs, the boosted KNN model predicts a probability of hypoxia well over 65 percent. For example, at Station 10, the actual value of dissolved oxygen on 23 July 2006 is 1.29 mg/L. In Figure 4.1, the KNN model forecast probability for this station is 54 percent, whereas the boosted KNN algorithm predicts 70 percent.

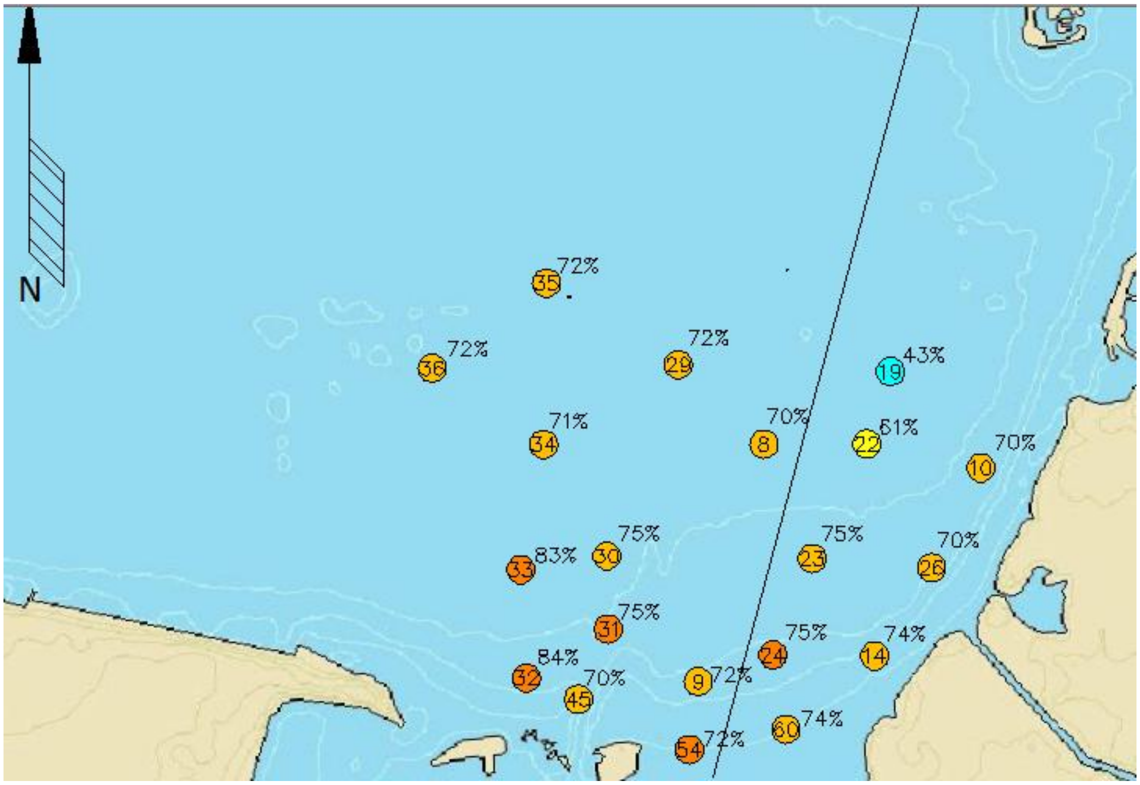


Figure 4.4 Boosted k-nearest neighbor algorithm 23-Jul-2006

The results for 23 July 2006 also show that the boosted historical scenario model is an improvement over the historical scenario model without boosting. At Station 10, where hypoxia occurred, the forecasts from the historical scenario model, the boosted historical scenario model and the boosted k-nearest neighbor model are 44 percent (Figure 4.3), 90 percent (Figure 4.5), and 70 percent (Figure 4.4), respectively.

The differences in performance of the historical scenario model and the boosted historical scenario model could be because of the inherent assumptions underlying the original valve model. The valve model assumes a constant gravity current speed of 1 km/day and a constant oxygen depletion rate. However, it does not take into account diurnal fluctuations due to photosynthesis and respiration of aquatic flora and fauna that could lead to variation in the oxygen depletion rate (To, 2009). Moreover, while the valve model tracks a single gravity current plume, the combined model tracks several plumes issued into CC Bay one after the other. Therefore, the overlap of these plumes, their different speeds, and their lifetimes in the bay could potentially affect the forecast probabilities (given in Figure 4.5).

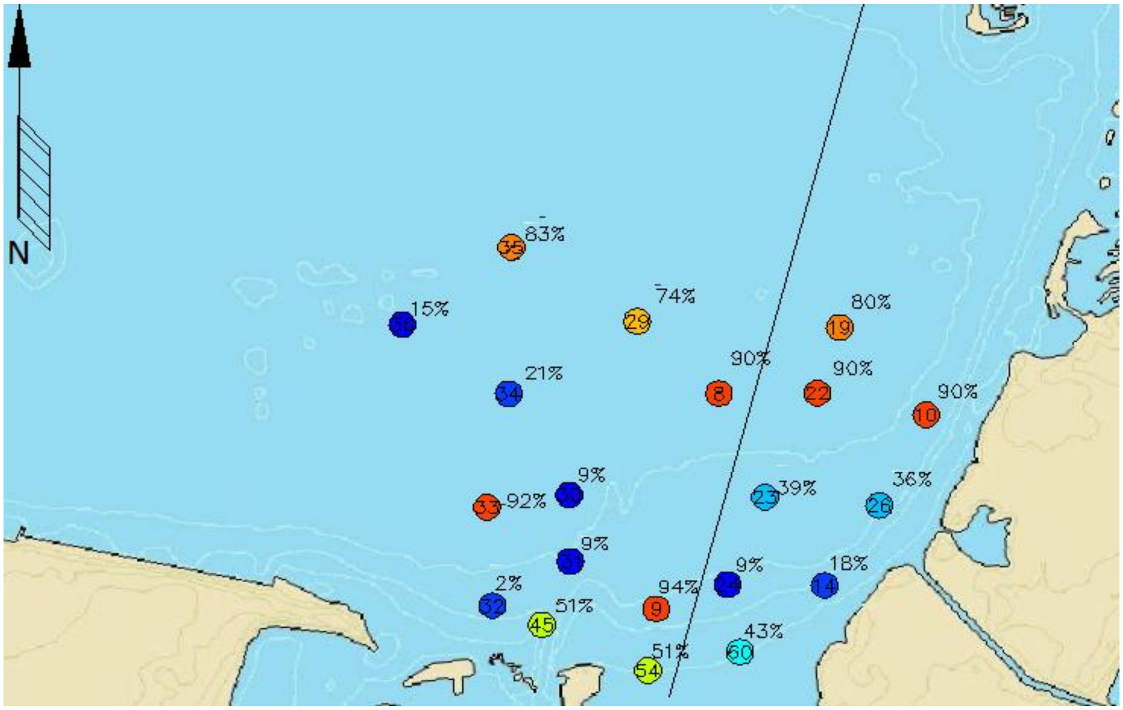


Figure 4.5 Boosted historical scenario model 23-Jul-2006

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 4.6. 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.

At Station 2, where hypoxia never occurs as discussed earlier, 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. This implies that the boosted historical scenario model is in agreement with the valve model that predicts 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.

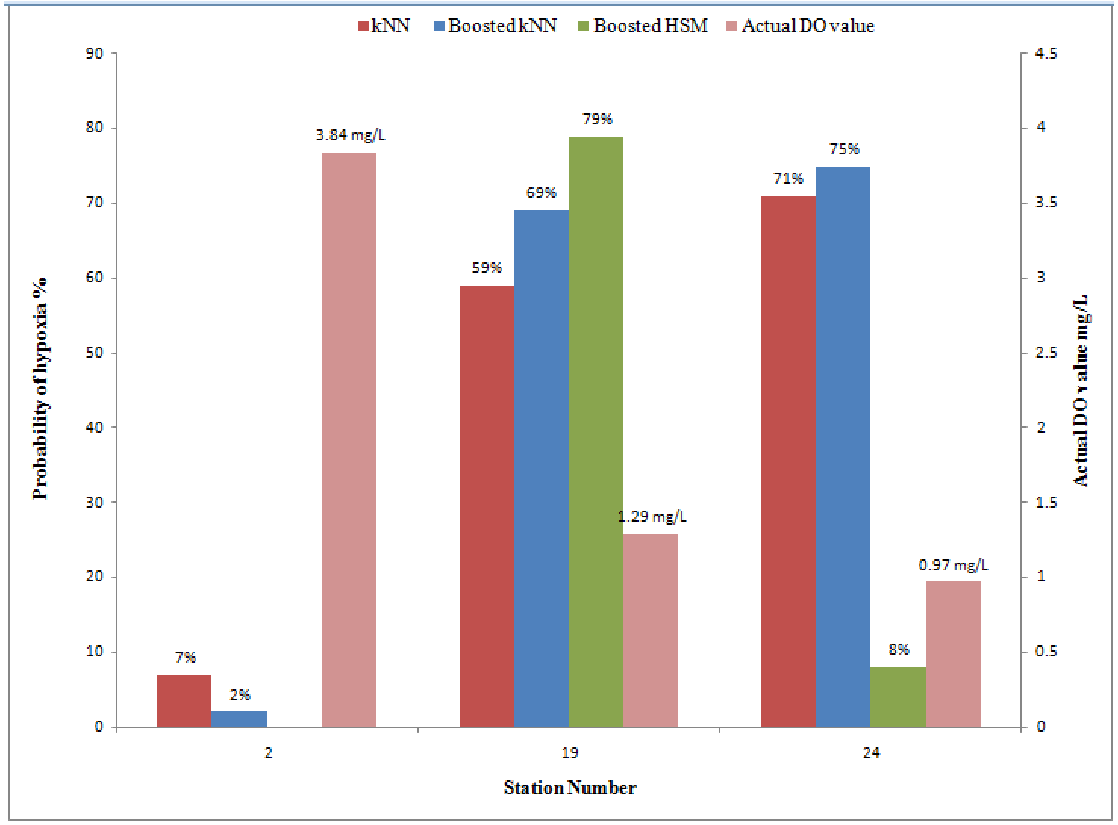


Figure 4.6 Forecasts of hypoxia versus actual dissolved oxygen values on 23-Jul-2006

Forecasts of hypoxia from all three models are also computed and compared (in Figure 4.7) for 23 June 2007, which had hypoxic conditions. On this date there were data from only two stations – Station 8 and Station 34 – for the week 06/19/2007 to 06/23/2007.

Figure 4.7 shows that the boosted historical scenario model gives more reliable results in the presence of both hypoxic and borderline hypoxic conditions. At station 34 with hypoxic conditions (DO value of 1.53 mg/l), the boosted historical scenario model predicts a higher value of 89 percent while k-nearest neighbor algorithm and boosted k-nearest neighbor algorithm show 68 percent and 71 percent respectively. At station 8 where there are borderline hypoxic conditions (DO of 2.52 mg/L), again the boosted historical scenario model predicts a lower value of 39 percent whereas the k-nearest neighbor algorithm and boosted k-nearest neighbor algorithm forecast 55 percent and 49 percent, respectively. Hence, in both cases the forecasts of the boosted historical scenario model are more definitive and the model appears to be more reliable.

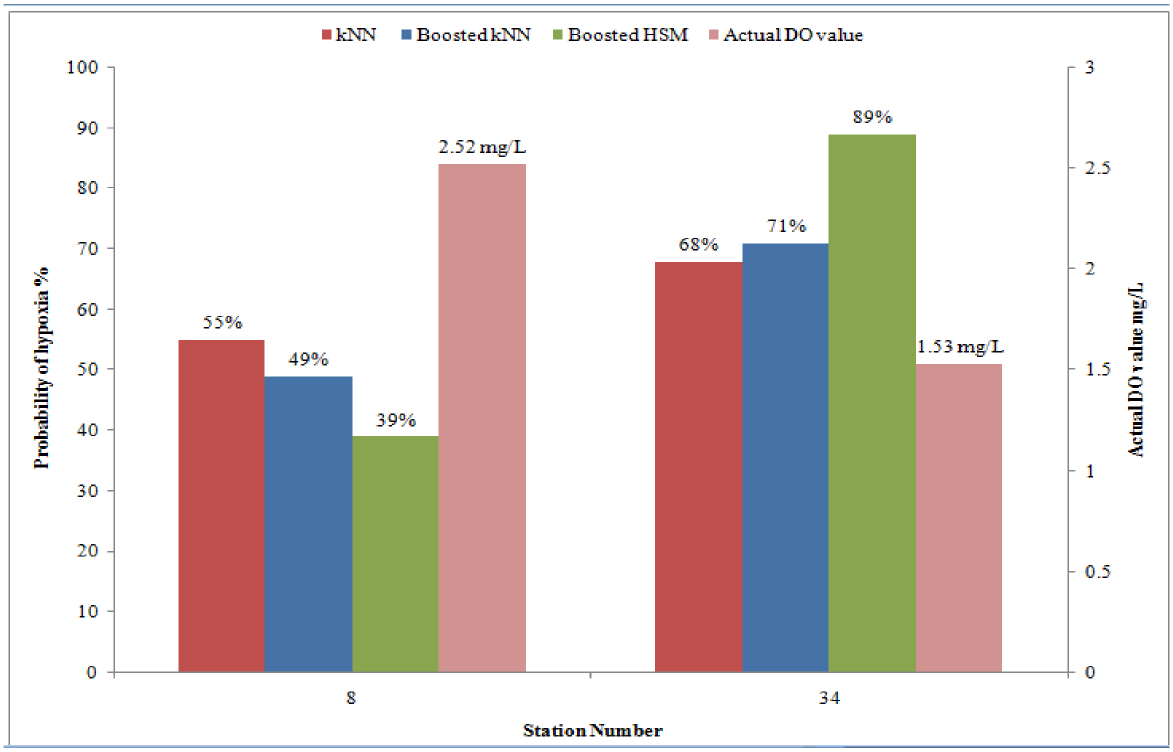


Figure 4.7 Forecasts of hypoxia versus actual dissolved oxygen values on 23-Jul-2006

5. Conclusion

The study presents three model fusion strategies – historical scenario modeling, boosting the k-nearest neighbor algorithm, and boosting the historical scenario model – and their outcomes. Where there were sufficient data, it was observed that the boosted k-nearest neighbor algorithm performed significantly better than the other fused as well as individual models. However, the performance of the boosted k-nearest neighbor was not equally reliable when there was not sufficient data from enough stations. In the latter case, the boosted historical scenario model was more reliable. In both cases, the results from the historical scenario model showed that more consistent data are necessary to come to any conclusions on its performance.

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. Boosting improved the performance of both the k-nearest neighbor algorithm and the historical scenario model.

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. The model testing was severely affected by a lack of daily data for successive weeks from multiple stations; hence additional data would improve the generalizability. For the individual and boosted k-nearest neighbor algorithm, there is a need for data, continuous or grab samples, at multiple stations for overlapping dates. On the other hand, the historical scenario model requires both continuous data at multiple stations and continuous data for overlapping dates. The fact that data were available for only two dates (23 July 2006 and 23 June 2007) limits the scope of the findings but indicates significant promise for these approaches if they can be tested with more data.

Lastly, 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 challenges in the valve model could prove to be of value in increasing the accuracy of the fused historical scenario model.

References

Ajami et al. 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. *American Meteorological Society*, 755- 768.

Amores, J., Sebe, N., and Radeva, P., 2006. Boosting the distance estimation: Application to the K-nearest neighbor classifier. *Pattern Recognition Letters*, Vol. 27, pp. 201-209.

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.

Baden, S.P., Loo, L-O., Pihl, L., and Rosenberg, R., 1990. Effects of eutrophication on benthic communities including fish: Swedish West Coast. *Ambio*, Vol. 19, No. 3, pp. 113-122.

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

Bianchi, T.S., DiMarco, S.F., Cowan Jr, J.H., Hetland, R.D., Chapman, P., Day, J.W., and Allison, M.A., 2010. The science of hypoxia in the Northern Gulf of Mexico: A review. *Science of the Total Environment*, Vol. 408, pp. 1471-1484.

Bloch, I., 1996. Information combination operators for data fusion: A comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 26, pp. 52-67.

Block, P.J., Filho, S., Sun, L., and Kwon, H.-H., 2009. A streamflow forecasting framework using multiple climate and hydrological models. *Journal of the American Water Resources Association*, Vol. 45, No. 4, 828-843.

Bordley, Robert F., 1982. The combination of forecasts: a Bayesian approach. *Journal of the Operational Research Society*, Vol 33, No. 2, 171-174.

Bunn, D.W., 1975. A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly*, Vol. 26, pp. 325-329.

Bunn, D.W., 1977. A comparative evaluation of the outperformance and minimum variance procedures for linear syntheses of forecasts. *Operational Research Quarterly*, Vol. 28, pp. 653-660.

Chakraborty et al. 2007. Prediction of the diurnal cycle using a multimodel superensemble. Part II: Clouds. *American Meteorological Society*, Vol. 135, 4097-4116.

Cho, S.-B., and Kim, J.H., 1995. Combining multiple neural networks by fuzzy integral and robust classification. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 25, pp. 380-384.

Cociasu, A., Dorogan, L., Humborg, C., and Popa, L., 1996. Long-term ecological changes in Romanian coastal waters of the Black Sea. *Marine Pollution Bulletin*, Vol. 32, No. 1, pp. 32-38.

Conley, D.J., Pearl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C., and Likens, G.E., 2009. Controlling eutrophication: Nitrogen and Phosphorus. *Science*, Vol. 323, pp. 1014-1015.

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

Coulibaly et al. 2005. Improving daily reservoir inflow forecasts with model combination. *Journal of Hydrologic Engineering*, 91-99.

Daskalov, G.M., 2003. Long-term changes in fish abundance and environmental indices in the Black Sea. *Marine Ecology Progress Series*, Vol. 255, pp. 259-270.

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

Dickinson, J.P. 1973. Some statistical results in the combination of forecast. *Operational Research Quarterly*, Vol. 24, No. 2, 253-260.

Dietrich et al. 2008. Combination of different types of ensembles for the adaptive simulation of probabilistic flood forecasts: hindcasts for the Mulde 2002 extreme event. *Nonlinear Processes in Geophysics*, 15, 275-286.

Ekau, W., Auel, H., Pörtner, 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.

Gader, P.D., Mohamed, M.A., and Keller, J.M., 1996. Fusion of handwritten word classifiers. *Pattern Recognition Letters*, Vol. 17, pp. 587-584.

Galkovskaya, G.A., and Mityanina, I.F., 2005. Structure distinctions of pelagic rotifer plankton in stratified lakes with different human impact. *Hydorbiologia*, Vol. 546, No. 1-3, pp. 387-395.

García-Pedrajas, N., and Ortiz-Boyer, D., 2008. Boosting random subspace method. *Neural Networks*, Vol. 21, pp. 1344-1362.

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.

Grabisch, M., and Dispot, F., 1992. A comparison of some for fuzzy classification on real data. In: *2nd International Conference on Fuzzy Logic and Neural Networks*, pp. 659-662, Iizuka, Japan, 1992.

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).

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.

Hodges, B., Furnans, J., and Kulis, P., 2008. Case study: A thin-layer gravity current with implications for desalination brine disposal. *Journal of Hydraulic Engineering*, in press.

Hsu, K-L., Moradkhani, H., and Sorooshian, S., 2009. A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research*, Vol. 45, pp. 1-15.

Islam, M.S., Bonner, J.S., Ojo, T., and Page, C., 2006. Using numerical modeling and direct observation to investigate hypoxia in a shallow wind-driven bay. *OCEANS 2006*, Art no. 4099028.

Islam, M.S., Bonner, J.S., Ojo, T., and Page, C., 2007. Real time monitoring of water quality parameters in Corpus Christi Bay to understand hypoxia. *OCEANS 2007 – Europe*, Art. No. 4302349.

Islam, M.S., Bonner, J.S., and Page, C.A., 2010. A fixed robotic profiler system to sense real-time episodic pulses in Corpus Christi Bay. *Environmental Engineering Science*, Vol. 27, No. 5, pp. 431-440.

Jacobs, R.A., 1995. Methods for combining experts' probability assessments. *Neural Computation*, Vol. 7, pp. 867-888.

Jacobs, R.A, Jordan, M.I., Nowlna, S.J., and Hinton, G.E., 1991. Adaptive mixtures of local experts. *Neural Computation*, Vol. 3, pp. 79-97.

Jordan, M.I., and Xu, L., 1995. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, Vol. 8, pp. 1409-1431.

Jonsson, P., Carman, R., and Wulff, F., 1990. Laminated sediments in the Baltic – a tool for evaluating nutrient mass balance. *Ambio*, Vol. 19, pp. 152-158.

Justić, D., 1987. Long-term eutrophication of the Adriatic Sea. *Marine Pollution Bulletin*, Vol. 18, No. 6, pp. 281-284.

Keller, J.M., Gader, P., Tahani, H., Chiang, J.-H, and Mohamed, M., 1994. Advances in fuzzy integration for pattern recognition. *Fuzzy Sets and Systems,* Vol. 65, pp. 273-283.

Kittler, J., Hatef, M., Duin, R.P.W., and Matas, J., 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp. 226-239.

Kulis, P. and Hodges, B.R., 2005. Improved techniques for gravity current modeling. *Proceedings of McMat 2005: 2005 Joint ASMEASCE/SES Conference on Mechanics and Materials* June 1-3, 2005, Baton Rouge, Louisiana, USA.

Kuncheva, L.I., 2001. Combining classifiers: Soft computing solutions. In: Pal, S.K., and Pal, A., (Eds)., *Pattern Recognition: From Classical to Modern Approaches*, World Scientific, pp. 427-451.

Langmead, O., McQuatters-Gollop, A., Mee, L.D., Friedrich, J., Gilbert, A.J., Gomoiu, M.T., Jackson, E.L., Knudsen, S., Minicheva, G., and Todorova, V., 2009. Recovery or decline of the northwestern Black Sea: A societal choice revealed by socio-ecological modeling. *Ecological Modelling*, Vol. 220, pp. 2927-2939.

Lam, L., and Suen, C.Y., 1997.Optimal combination of pattern classifiers. *Pattern Recognition Letters*, Vol. 16, pp. 945-954.

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.

Lim, H-S., Diaz, R.J., Hong, J-S., and Schaffner, L.C., 2006. Hypoxia and benthic community recovery in Korean coastal waters. *Marine Pollution Bulletin*, Vol. 52, pp. 1517-1526.

Mee, L.D., 1992. The Black Sea in crisis: A need for concerted international action. *Ambio*, Vol. 21, No. 4, pp. 278-286.

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.

Newbold, P., and Granger, C.W.J., 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A (General)*, Vol. 137, No. 2, pp. 131-165.

Nowlan, S.J., and Hinton, G.E., 1991. Evaluation of adaptive mixtures of competing experts. In: Lippmann, R.P., Moody, J.E., and Touretzky, D.S., (Eds), *Advances in Neural Information*

Processing Systems 3, pp. 774-780.

Raftery, A.E., Madigan, D., and Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, Vol. 92, No. 437, pp. 179-191.

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.

Rogova, G., 1994. Combining the results of several neural network classifiers. *Neural Networks*, Vol. 7, pp. 777-781.

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

See, L. 2008. Data fusion methods for integrating data-driven hydrological models. *Studies in Computational Intelligence* (SCI) 79, 1-18.

See, L., Abrahart, R. 2001. Multi-model data fusion for hydrological forecasting *Computers & Geosciences* 27 (2001) 987-994.

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

Shamseldin, Assad Y., and O'Connor, Kieran M. 1999. A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences-Journal,* Vol. 44, No. 6, 895-912.

Shamseldin, Assad Y., O'Connor, Kieran M., and Liang, G.C. 1997. Methods for combining the outputs of different rainfall-runoff models. *Journal of Hydrology,* Vol. 197, 203-229.

Suzuki, T., 2001. Oxygen-deficient waters along the Japanese coast and their effects upon the estuarine ecosystem. *Journal of Environmental Quality*, Vol. 30, pp. 291-302.

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

Vrugt, Jasper A., and Robinson, Bruce A. treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*, Vol. 43, W01411, doi:10.1029/2005WR004838.

Woods, K., Kegelmeyer, W.P., and Bowyer, K., 1997. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 405-410.

Xu, L., Krzyzak, A., and Suen, C.Y., 1992. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, pp. 418-435.

Yang, X., Yuan, B., and Liu, W., 2009. The effects of distance metrics on boosting with dynamic weighting schemes. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 320-324.

Zhang, J., Gilbert, D., Gooday, A.J., Levin, L., Naqvi, S.W.A., Middelburg, J.J., Scranton, M., Ekau, W., Peña, A., Dewitte, B., Oguz, T., Monteiro, P.M.S., Urban, E., Rabalais, N.N., Ittekkot, V., Kemp, W.M., Ulloa, O., Elmgren, R., Escobar-Briones, E., and Van der Plas, A.K., 2010. Natural and human-induced hypoxia and consequences for coastal areas: synthesis and future development. *Biogeosciences*, Vol. 7, pp. 1443-1467.

Zillén, L., Conley, D.J., Andrén, T., Andrén, E., and Björck, S., 2008. Past occurrences of hypoxia in the Baltic Sea and the role of climate variability environmental change and human impact. *Earth Science Reviews*, Vol. 91, pp. 77-92.