# HARNESSING THE POWER OF SENSORS AND CYBERINFRASTRUCTURE TOWARDS ENVIRONMENTAL SUSTAINABILITY: THE WATERS NETWORK VISION AND TESTBEDDING RESEARCH

BARBARA SPANG MINSKER AND EVAN COOPERSMITH
*Department of Civil and Environmental Engineering, University of Illinois, MC-250 Urbana, IL 61822, USA*

PAUL MONTAGNA
*Harte Research Institute, Texas A&M University-Corpus Christi Corpus Christi, TX 78412, USA*

The WATERS Network is a U. S. National Science Foundation-supported environmental observatory initiative to transform science and engineering of the water environment through new infrastructure investments. WATERS Network seeks to answer the following overarching question: How do we establish a framework to more reliably predict and manage water quantity and quality as climate changes, populations grow, land use evolves, and individual and societal choices are made? Research enabled by the WATERS Network will provide the basic knowledge needed to understand, engineer, manage, and set policy for water resources systems and infrastructure that are critical for life and society. To assist with identifying needs and potential research outcomes from WATERS Network observatories, testbedding activities are underway in many locations around the U.S. This paper highlights ongoing research and infrastructure activities in the Upper Illinois River Basin and Corpus Christi Bay, Texas, and focuses on recent findings from the Corpus Christi Bay testbed. New approaches for enabling real-time forecasting of hypoxia have been developed using data integration and nearest-neighbor algorithms. Early forecasts were used to support adaptive sampling that identified far more widespread hypoxia than previously expected. These results show that the real-time, adaptive sampling envisioned under the WATERS Network is feasible and useful for understanding dynamic environmental systems.

## INTRODUCTION

The WATERS Network (http://watersnet.org) is evaluating how a combination of measurements from observatories and facilities, innovative technologies, and models will address two overarching research questions: How will regional-scale fresh water availability and demand change in the future? How will human behavior, land-use change, and the water management infrastructure affect the quality of water? Supported by the U.S. National Science Foundation (NSF), WATERS Network is in the conceptual

design phase of NSF's Major Research Equipment and Facilities Construction (MREFC) program, which builds major national community research and education infrastructure.

To demonstrate the value of investing in the WATERS Network, numerous testbedding projects are exploring various aspects of network and observatory design and their benefits to research. The sections below highlight two testbeds in the Upper Illinois River Basin and Corpus Christi Bay, followed by a brief summary of methodology and results of adaptive hypoxia forecasting in Corpus Christi Bay.

## UPPER ILLINOIS RIVER BASIN TESTBED

The Upper Illinois River Basin (UIRB) is in the humid continental climate region of the United States and encompasses diverse agricultural, suburban, and urban land uses, including the City of Chicago. The UIRB testbed, shown in Figure 1, includes the UIRB and a few additional watersheds along Lake Michigan, which serves as water supply for the City of Chicago. UIRB contains nearly all aspects of the water cycle, including groundwater aquifers used for water supply, extensive engineered infrastructure, a Great Lake used for water supply and swimming, rivers, tile drainage systems that are typical of Midwestern agriculture, etc. The UIRB receives numerous inputs of contaminants and nutrients from manmade sources that include municipal and industrial releases, urban and agricultural runoff, and atmospheric deposition. Urban development of the Chicago and Milwaukee metropolitan areas has affected the UIRB extensively, particularly straining groundwater supplies in suburban areas. Withdrawals from Lake Michigan are regulated by international treaty and Supreme Court Decree, hence this area is expected to face significant water supply challenges in the future. The effects from converting previously agricultural land to new residential land are not fully known and may affect riparian ecosystems thorough accelerated erosion and channel instability; loss of aquatic habitat; increase in peak flow, duration of peaks, and flood volumes; and loss of base flow from ground-water pumping and change in surface water drainage networks. Surface waters in the UIRB currently face significant water quality challenges due to urban and agricultural runoff throughout the region, as well as combined sewer overflows in the Chicago region. Lake Michigan beach closings due to high E. coli measurements are also frequent.

Initial testbedding activities in the UIRB have focused on real-time forecasting and management of combined sewer overflows in the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC). MWRDGC's Tunnel and Reservoir Plan (TARP), whose completion is anticipated in 2019, consists of over 100 miles of underground tunnels connecting several reservoirs, which can be used during wet weather events to store excess wastewater until it can be treated and discharged into the Chicago waterways system. The TARP project was initiated to address two primary issues facing Chicago: basement flooding and pollution from CSOs. However, recent observations indicate that both flooding and CSOs continue to occur despite excess storage within the TARP system. For example, on August 22–23, 2007, a storm passed over Chicago, dropping 1.9 inches of rain in 7.5 hours. The storm resulted in significant flooding in the

Chicago area, caused 189 CSO events, and forced the reversal of the North Branch of the Chicago River into Lake Michigan. At the conclusion of this storm, however, there were approximately 500 million gallons (25%) of storage still available in TARP. Events such as this suggest that conveyance of wastewater to and within the TARP system needs to be analyzed for infrastructure deficiencies and control decisions optimized.

Initial work has identified high-risk locations for CSO overflows and a real-time CSO forecasting system is now being developed that will combine physics-based models with dynamic Bayesian networks and support vector machines. Using these models, a real-time decision support system is being created that combines multiple sensor data sources with advanced optimization to identify operator strategies that will minimize CSO events, as shown in Figure 2.

## CORPUS CHRISTI BAY TESTBED

Corpus Christi Bay (CCBay) is located along the southeastern coast of Texas, just west of the Gulf of Mexico with a dividing barrier island, as shown in Figure 3 (Google Earth). The bay itself represents an urban estuary, home to the city and Port of Corpus Christi, which serves as this country's seventh largest port. The relatively low tidal range within the Gulf of Mexico, combined with the restricted channel entrances, causes water circulation to be defined more by meteorological phenomena like wind, rather than the traditional mechanism of tides (Kulis and Hodges [1]). Since 1988, CCBay has had numerous incidents of hypoxia, during which dissolved oxygen fell below 2 mg/L (Montagna and Kalke [2]). These events threaten benthic life, reducing diversity and biomass and often causing the relocation of sensitive benthic organisms to surface waters (Ritter and Montagna [3]). Hydrodynamic analysis suggests that water masses moving northward from Laguna Madre may provide a mechanism by which saline water enters the bay and generates hypoxic conditions through salinity stratification (Hodges and Furnans [4]). Coincident with the salinity currents are the wastewater treatment plants on the Oso Creek (shown in purple in Figure 3), which discharge to Oso Bay and from there to Corpus Christi Bay, providing a potential anthropogenic source of outside contaminants, along with agricultural runoff, that can lead to nitrate-based eutrophication.

Within Corpus Christi Bay lies a large number of sensors operated by agencies and researchers, shown in Figure 3. Particularly critical to this work is a spatial grid of 50 oxygen measurement locations that have provided oxygen levels over the last five to ten years. Because hypoxic events are observed predominantly in the benthic zone (Osterman *et al.* [5]), this analysis focuses solely on grab and continuous sample data from the bottom layer of the water column. 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. Continuous sensors are deployed at a location and depth for a period of one to two weeks, during which they log readings every fifteen minutes. These data provide a critical record of the temporal dynamics needed for 24-hour ahead hypoxia forecasting, the goal of this work.

**METHODOLOGY**

This study followed the steps below to create a short-term hypoxia forecasting model using available historical datasets from CCBay:

1. *Removing time-dependent trends.* Because oxygen and temperature sensor readings exhibit natural cycles, it was necessary to remove cycles from the data prior to fitting the model. Detrending the dataset was accomplished via sequential normalization (Maidment and Parzen [6]), in which the longest-periodicity trend is removed first, followed by smaller periodicities. In the case of oxygen, a linear long-term trend of slowly reducing oxygen levels was removed first, followed by seasonal and daily trends that were modeled by discrete Fourier transforms with two harmonics. Temperature was detrended similarly.
2. *Isolating Parameters That Influence DO / Model Testing.* Once the trends were removed, the resulting stationary data were then used to identify which parameters most influence dissolved oxygen levels and to fit machine learning algorithms to the data to forecast hypoxia based on those parameters. For this purpose, two distinct, nonparametric machine learning algorithms were examined: k-nearest neighbors (KNN) and regression trees. With KNN, given an input vector consisting of a variety of variables relevant to the prediction of hypoxia, a distance function is computed to determine the similarity between the input data point and another, historical value. The *k* best matches are located (with k chosen *a priori*) and used to predict a plausible distribution of outcomes. For example, if 80 of 100 similar records showed hypoxia 24 hours later, then a forecast of 80% probability of hypoxia is made. Regression trees (Breiman et al. 1984) split the dataset into branches by maximizing the ratio of information gain. At the end of each path of branches is a node, within which is a set of historical records which, like KNN, should be similar to the current data point whose future is unknown. This is used to predict the likelihood of hypoxia in a similar fashion to the KNN approach. Unfortunately the regression tree

approach did not perform well for this application, and results are not shown here (see Coopersmith [7]).

3. *Calibration and Validation Using Historical Data at a Single Location.* To calibrate the models at a single location, a sliding window approach was used. A segment of data (the most recent segment) was reserved for validation and the remaining data were used to fit the models.
4. *Spatial Interpolation.* Having validated forecasts at individual locations with hind-casting, the next step was to interpolate spatially from these results to multiple locations using the following three-step procedure: (1) At each of the location coordinates in a spatial interpolation grid, the historical database was used to create the best estimate of the mean and standard deviation for each independent variable (called the "baseline expectation"). (2) Next today's data were used to generate tomorrow's forecast for each independent variable at each location in the interpolation grid using inverse-distance weighting of forecasts at nearby sensors, which are each normalized by the baseline expectation at that location. (3) Lastly, given the spatially normalized values computed for each of the independent variables, the machine learning algorithms were then implemented to predict all relevant dependent variables (in this case oxygen).

**RESULTS**

Dissolved oxygen, salinity, temperature, and wind (speed and direction) were found to be most predictive of hypoxia 24 hours ahead, but wind proved to be particularly critical to accurate forecasting. The top graph in Figure 4 shows the projections of dissolved oxygen levels with and without the inclusion of wind data (dark and light blue lines, respectively) as well as the actual observed result (red). The bottom graph in Figure 4 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 20$^{th}$ 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.

Spatial maps of hypoxia forecasts were also generated using data from each sampling location (shown in Figure 5, where the black numbers represent grab sampling locations and the white numbers represent "continuous" embedded sensor locations) and incorporating latitude and longitude into the k-nearest neighbor algorithm as independent variables (see, e.g., Figures 5 through 7). 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 7), reflecting more uncertainty, as well as lower expected dissolved oxygen levels (Figure 5) are prime candidates for an additional sensors. Based on the results given in Figures 5-7, researchers introduced new sensors (#8, #34, #199, and #202 in Figure 8) in summer

2007, from which they learned that hypoxic risk is far wider spread than initially believed (Figure 8).

## CONCLUSIONS

This work demonstrates the value of integrating sensor data with machine learning algorithms to improve understanding and management of complex environmental phenomena. Based on early forecasts from the CCBay hypoxia model that predicted more widespread hypoxia beyond the current sampling area, in Summer 2007 researchers implemented a broader sampling strategy in CCBay over a longer period and showed that hypoxia occurs in late spring and early fall as well as summer, and the spatial extent of the hypoxic area is double the size of previous observations. This demonstrates the value of adaptive sampling using models of sensor data to enhance understanding of dynamic water quality issues. The hypoxic area now extends from Laguna Madre to Oso Bay, which is important because three wastewater treatment plants discharge to Oso Creek (see Figure 3). The significance of this finding is that for the first time since 1988, there is evidence that nutrient discharges may be playing a role in causing hypoxia in Corpus Christi Bay. If that is true, then there are management options that could contribute to solving the problem, which previously had not been thought possible. The role of nutrients is being investigated further in the coming year.

These findings illustrate the advantages of further investment in integrated real-time adaptive sensing, modeling, and cyberinfrastructure such as that proposed under the WATERS Network. Although the analysis in this work simulated a real-time forecast using historical data, with real-time infrastructure the researchers in CCBay would be able to better schedule their manual sampling campaigns to periods with high risk of hypoxia. Such investment would also enable more sensor data at additional spatial locations, which should further improve understanding of the causes of hypoxia, as illustrated by the additional data collection activities in Summer 2007. Ongoing research will explore whether similar advantages can be gained in the urban setting of UIRB with combined sewer overflow modeling.

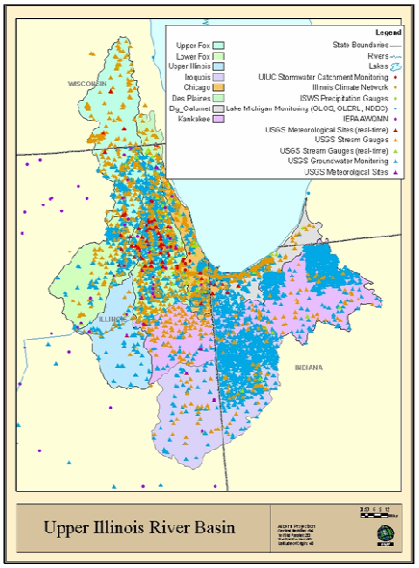


Figure 1. Upper Illinois River Basin and some of the many data collected by multiple agencies.

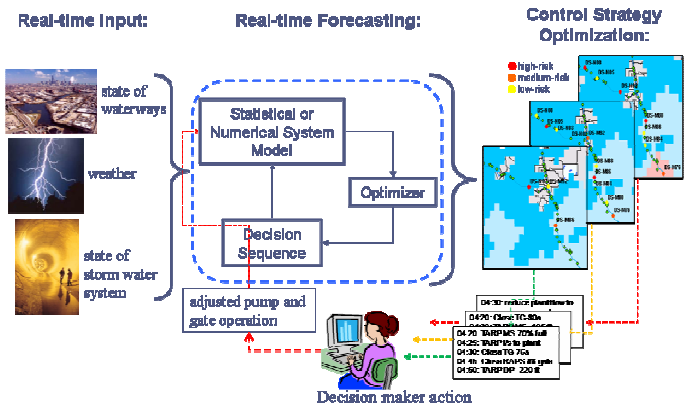


Figure 2. Real-time CSO decision support system.

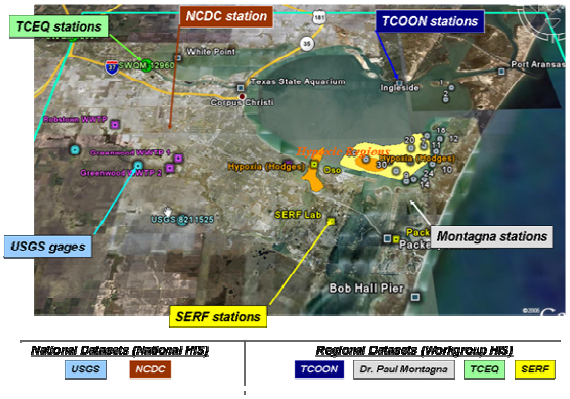


Figure 3. Corpus Christi Bay and its sensor systems.

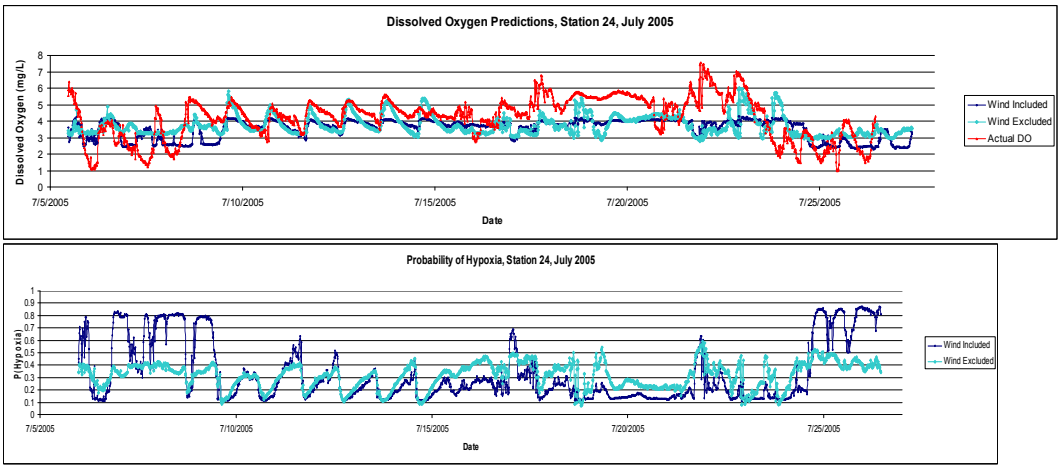


Figure 4. Top: Dissolved oxygen levels; bottom: estimated probability of hypoxia

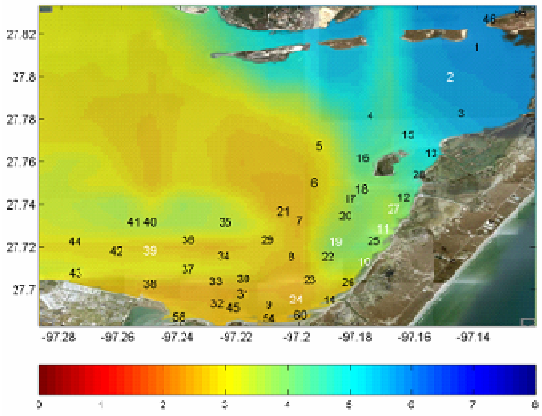

Figure 5. Expected dissolved oxygen levels (mg/l)

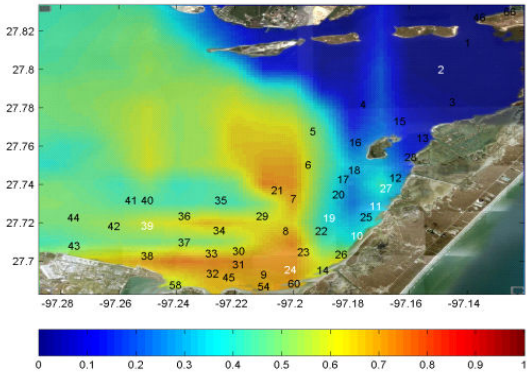

Figure 6. Probability of hypoxia

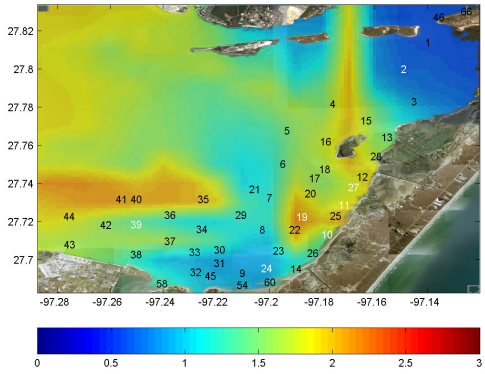

Figure 7. Standard deviation of oxygen forecast (mg/l)

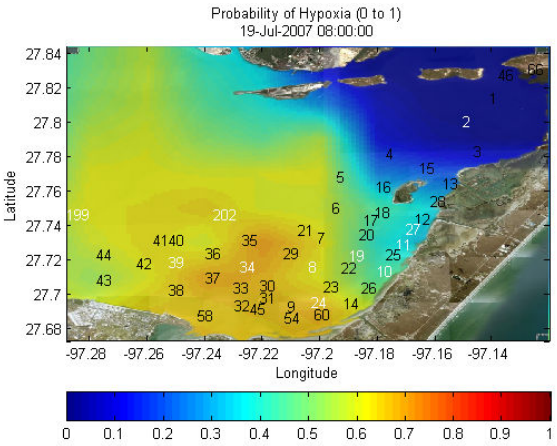


Figure 8. Hypoxic risk with new sensors included during Summer 2007

## REFERENCES


[1] Kulis P. and Hodges B. R., "Three dimensional circulation in Corpus Christi Bay", Gulf Estuarine Research Society Meeting, Corpus Christi, Texas, USA (2006).

[2] Montagna P.A. and Kalke R.D., "The effect of freshwater inflow on meiofaunal and macrofaunal populations in the Guadalupe and Nueces Estuaries," Texas, *Estuaries*, Vol. 15, (1992), pp 307-326.

[3] Ritter M.C. and Montagna P.A., "Seasonal hypoxia and models of benthic response in a Texas bay," *Estuaries*, Vol. 22, (1999), pp 7-20.

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

[5] Osterman L.E., Poore R.Z., and Swarzenski P.W., "The last 1000 years of natural and anthropogenic low-oxygen bottom-water on the Louisiana shelf, Gulf of Mexico," *Marine Micropaleontology*, Vol. 66, No. 3/4, (2008), pp 291-303.

[6] Maidment D.R. and Parzen E., "Time patterns of water usage in six Texas cities," *J. of Water Resour. Plann. Manage.*, Vol. 110, (1984), pp 90-106.

[7] Coopersmith, E., *Understanding And Forecasting Hypoxia Using Machine Learning Algorithms*, M.S. thesis, University of Illinois, Urbana, IL, (2008).


## ACKNOWLEDGMENTS


This work was supported by the U.S. National Science Foundation under grants CBET-0609545 and CBET-0533513. The authors gratefully acknowledge the many contributions of the CCBay and IRB testbed teams and the WATERS Network design team to this work, particularly Jeff Dozier for the WATERS Network background information, David Hill for the UIRB testbed description and graphics, David Maidment for assistance with the detrending, and Paula Kulis and Bed Hodges for their insights on hypoxia mechanisms.