Forecasting hypoxia in the Chesapeake Bay and Gulf of Mexico: model accuracy, precision, and sensitivity to ecosystem change



2011 Environ. Res. Lett. 6 015001

(http://iopscience.iop.org/1748-9326/6/1/015001)

Environ. Res. Lett. **6** (2011) 015001 (11pp) doi:10.1088/1748-9326/6/1/015001

# Forecasting hypoxia in the Chesapeake Bay and Gulf of Mexico: model accuracy, precision, and sensitivity to ecosystem change

**Mary Anne Evans**[1] **and Donald Scavia**[1,2]

[1] School of Natural Resources and Environment, University of Michigan, Ann Arbor, MI 48109, USA
[2] Graham Sustainability Institute, University of Michigan, Ann Arbor, MI 48104, USA

E-mail: mevans@umich.edu and scavia@umich.edu

Received 25 August 2010
Accepted for publication 26 November 2010
Published 23 December 2010
Online at stacks.iop.org/ERL/6/015001

**Abstract**
Increasing use of ecological models for management and policy requires robust evaluation of model precision, accuracy, and sensitivity to ecosystem change. We conducted such an evaluation of hypoxia models for the northern Gulf of Mexico and Chesapeake Bay using hindcasts of historical data, comparing several approaches to model calibration. For both systems we find that model sensitivity and precision can be optimized and model accuracy maintained within reasonable bounds by calibrating the model to relatively short, recent 3 year datasets. Model accuracy was higher for Chesapeake Bay than for the Gulf of Mexico, potentially indicating the greater importance of unmodeled processes in the latter system. Retrospective analyses demonstrate both directional and variable changes in sensitivity of hypoxia to nutrient loads.

**Keywords:** model-data comparison, coastal systems, nitrogen loading, eutrophication

## 1. Introduction

Ecological models are increasingly moving from heuristic to applied, and this movement requires rigorous analysis and optimization of accuracy, precision, and sensitivity to system change. Ecological systems are subject to sporadic changes caused by internal dynamics (Bronmark *et al* 2010), shifts in drivers (climate (Scheffer and van Nes 2007), human inputs (Goolsby *et al* 2001, Rabalais *et al* 2002a)), invasive species (Higgins and Zanden 2010), and other factors. Some of these changes can be included in models explicitly, but others are beyond the scope of most modeling activities. These unmodeled changes and processes are generally parameterized through key model coefficients, and because systems change, those parameterizations are subject to change, therefore it is important for model calibrations to reflect the current state of the system.

Ecosystems are also subject to relatively high 'random' short-term variability (e.g. weather) that does not necessarily reflect directional change. Robust model parameterization thus also requires sufficiently long time frames to capture the range of system variability to both detect mean behavior and undertake reasonable uncertainty analysis. There is a potential tension between the goals of providing high accuracy and high precision and between the challenges of incorporating information about both random variability and long-term system changes. So, it is important to develop model calibration approaches that optimize model performance (accuracy, precision) in the face of systems that are both undergoing directional change and are highly variable.

Models of varying degrees of complexity have been informative tools in understanding the controls on hypoxia occurrence in river-impacted coastal areas (Peña *et al* 2010). Hypoxia, low oxygen concentrations in bottom waters, occurs

 

when decomposition rates exceed those of oxygen diffusion and mixing. Hypoxia is a widespread and increasing phenomenon (Diaz and Rosenberg 2008, Zhang *et al* 2010) that can lead to widespread ecosystem changes including altered biogeochemical cycles (Kemp *et al* 2005, Turner *et al* 2008), fish kills (Diaz and Rosenberg 2008), decreased or displaced fish production (Rabalais and Turner 2001), and decreased value to human use through recreation and fisheries harvest losses (Renaud 1986).

Two major river-impacted coastal hypoxic areas of the United States occur in the Gulf of Mexico (GOM) along the Louisiana–Texas coasts and in Chesapeake Bay (CB). Hypoxia has been heavily studied in these areas (Justić *et al* 1993, Bierman *et al* 1994, Rabalais *et al* 1994, 1998, Boesch *et al* 2001, Rabalais and Turner 2001, Hagy 2002, Rabalais *et al* 2002a, 2002c, Childs *et al* 2002, Rabalais *et al* 2004, Kemp *et al* 2005, Rabalais 2006, Walker and Rabalais 2006, Scully 2010, etc), due in part to concern over potential fisheries impacts (Renaud 1986, Rabalais and Turner 2001), and management goals have been set to limit hypoxia severity. Models have been used successfully in both systems to explore the underlining causes of hypoxia and to make specific management recommendations (Cerco and Cole 1993, Rabalais *et al* 2002b, Justić *et al* 2003, Scavia *et al* 2003, Hagy *et al* 2004, Turner *et al* 2005, Scavia *et al* 2006, Turner *et al* 2006, Justić *et al* 2007, Rabalais *et al* 2007, Turner *et al* 2008, Greene *et al* 2009, Penta *et al* 2009, Wang and Justić 2009, Bianchi *et al* 2010, Liu *et al* 2010, Liu and Scavia 2010, Peña *et al* 2010). Models and empirical data indicate that hypoxia in these systems is caused by a combination of nutrient-driven, mostly nitrogen, production of phytoplankton organic matter; decomposition; freshwater-driven stratification of the water-column; and storm mixing. Management recommendations have generally focused on control of nitrogen loading to these systems due to evidence that it is an important driver of hypoxia and its susceptibility to management compared to other drivers. However, phosphorus load control has also been addressed (Boesch *et al* 2001, Environmental Protection Agency (EPA) Science Advisory Board (SAB) 2007, Mississippi River/Gulf of Mexico Watershed Nutrient Task Force 2008).

Both systems have also undergone significant ecosystem changes in hypoxia sensitivity to nutrient loads over the last 30 years, such that in both systems the severity of hypoxia for a given nitrogen load is now approximately twice what it was in the early 1980s (Hagy *et al* 2004, Turner *et al* 2008, Liu *et al* 2010, Liu and Scavia 2010). Ongoing research and management scenarios are thus complicated by the need to account for this varying ecosystem sensitivity and by speculation about how the systems will respond as nutrient loads change. Shifts in system sensitivity can appear abrupt when viewed retrospectively (Hagy *et al* 2004, Environmental Protection Agency (EPA) Science Advisory Board (SAB) 2007, Turner *et al* 2008, Greene *et al* 2009, Liu *et al* 2010), however, because of significant interannual variability, they can be impossible to recognize contemporaneously. This delayed recognition of sensitivity change is a challenge to both short- (annual) and long- (management scenarios) term results, and highlights the need for models and model calibration approaches that optimize model performance in changing and highly variable systems.

In this study we test different model calibration approaches for fitting similar models of the GOM and CB to subsets of historical data that include system changes and periods of high variability. A wide range of modeling approaches, from simple regressions to 3D coupled hydrodynamic–biogeochemical and earth system models have been applied to hypoxia for both management and scientific investigation (Peña *et al* 2010). More complex models are generally able to resolve finer scale ecological mechanisms and provide process based insight. Simpler models, however, are often better predictors of system state and have proven very useful for management applications (Peña *et al* 2010). Within this range we use a relatively simple, mechanistically based, model that treats estuary and coastal currents as 'rivers' with point source organic matter loads. We selected this model because it has proven useful for management guidance and because the computational simplicity allows the explicit incorporation of uncertainty analysis (Scavia *et al* 2003, 2004, 2006, Scavia and Donnelly 2007, Stow and Scavia 2009, Liu *et al* 2010, Liu and Scavia 2010). For a description of this model's use in the GOM in the context of other modeling approaches, see the recent review by Peña *et al* (2010).

We test for accuracy, precision, and model sensitivity to system changes by hindcasting parts of the historical dataset. We then compare optimal model calibrations between these two systems and discuss its implications for both ecological interpretation and management. Finally, we use our optimal calibrations to forecast outcomes under different nutrient reduction scenarios.

## 2. Methods

### *2.1. Models*

We use versions of the Streeter–Phelps (SP) river model (Chapra 1997) developed for CB and the GOM. The model is described in greater depth and its assumptions justified in earlier publications (Scavia *et al* 2003, 2004, 2006, Scavia and Donnelly 2007, Stow and Scavia 2009, Liu *et al* 2010, Liu and Scavia 2010). These models share the same basic structure but are adapted to each system. Both models treat the estuary or coastal current as a 'river' and calculate longitudinal profiles of dissolved oxygen (DO) concentration downstream of an organic matter (BOD) point source (described below for each system). This organic matter point source is assumed to be proportional to the spring total nitrogen (TN) loading to the system with a proportionality constant equal to the product of the Redfield carbon to nitrogen ratio, the respiration ratio of oxygen consumption per organic carbon, and the dilution of inputs within the receiving water body. Spring TN loads were used because spring loads are the dominant drivers of hypoxia in these systems (Cerco 1995, Scavia *et al* 2003, Hagy *et al* 2004, Turner *et al* 2006).

DO profiles are calculated at steady state, for each location along the profile, DO is calculated by:

$$\mathrm{DO} = \mathrm{DO_s} - \frac{k_\mathrm{d}\mathrm{BOD_u}F}{k_\mathrm{r} - k_\mathrm{d}}(\mathrm{e}^{-k_\mathrm{d}\frac{x}{v}} - \mathrm{e}^{-k_\mathrm{r}\frac{x}{v}}) - D_\mathrm{i}\mathrm{e}^{-k_\mathrm{r}\frac{x}{v}} \quad (1)$$

where: DO = dissolved oxygen (mg $\mathrm{l}^{-1}$), $\mathrm{DO_s}$ = oxygen saturation (mg $\mathrm{l}^{-1}$), $k_\mathrm{d}$ = BOD decay coefficient (1/day), $k_\mathrm{r}$ = reaeration coefficient (1/day), $\mathrm{BOD_u}$ = initial BOD (mg $\mathrm{l}^{-1}$), $x$ = downstream distance (km), $F$ = fraction of BOD sinking below the pycnocline (unitless), $D_\mathrm{i}$ = the initial oxygen deficit (mg $\mathrm{l}^{-1}$), and $v$ = net downstream advection (km/day). While in the original SP formulation, $v$ represents net downstream advection, in this application it also parameterizes the combined effect of horizontal transport and subsequent settling of organic matter from the surface. Therefore, it has no simple physical analog.

The length of the hypoxic zone is summed across the part of the profile with DO at hypoxic levels and converted to a measure of hypoxic area or volume by empirical relationships developed from measurements of the hypoxic area or volume in each system (see below). The model was calibrated by fitting predicted and measured area or volume and minimizing error terms. During calibration, each parameter can be assumed to be either constant across all years or adjusted each year. If a parameter is adjusted each year, we assume that its variability includes the effects of all unmodeled processes.

As in prior applications to CB and the GOM (Stow and Scavia 2009, Liu *et al* 2010, Liu and Scavia 2010), the model was calibrated using Bayesian fitting through Markov Chain Monte Carlo methods (Lunn *et al* 2000, Gill 2002, Gelman and Hill 2007). All model calibration was conducted in WinBUGS (version 1.4.3), called through R (version 2.6.0, R2WinBUGS, version 2.1-8), using the same methods and inputs described elsewhere (Stow and Scavia 2009, Liu *et al* 2010, Liu and Scavia 2010). In prior applications of both models, either $v$ or $F$ was allowed to vary by year, and all other parameters were fit as constants across years or determined from empirical data (see below).

Model application to the two systems differed in four ways:

(1) The location of the organic matter point source was determined by the geography and physics of each system. In CB, summer surface waters flow seaward and bottom waters flow landward. The primary nutrient input to the modeled area of CB is the Susquehanna River at the head of the bay and most hypoxia occurs in the mid-bay region. Thus, the model origin and organic matter point source were assigned to the lower end of the mid-bay region (220 km down bay from the Susquehanna River mouth) and distance in the model is following the landward flowing bottom water. Organic matter loading was based on Susquehanna River spring TN loading. In the GOM, hypoxia occurs below a westward flowing coastal current along the Louisiana and Texas costs. Because there are two main nutrient inputs to the GOM (the Mississippi and Atchafalaya Rivers), we model two organic matter point sources, one at the model origin (Mississippi River) and one at 220 km down current (Atchafalaya River). Organic matter is proportional to spring TN load with 50% of the Mississippi River and 100% of the Atchafalaya River TN load assumed to be entrained in the westward current.

(2) The initial oxygen deficit ($D_\mathrm{i}$) was assumed to be 0 in the GOM because there is little oxygen depletion in waters east of the delta. $D_\mathrm{i}$ in CB was estimated each year based on measured bottom-water oxygen concentrations at the model origin and a stochastic term based on measurement variation.

(3) In CB, the reaeration coefficient is known to vary with distance down estuary (Hagy 2002). Our model uses this observed variation in distance ($x$) and calculates $k_{rx} = b_x K$ where $b_x$ is a location specific constant accounting for spatial variation (Scavia *et al* 2006) and $K$ is a fit model parameter scaling reaeration.

(4) In CB, the volume of water with DO $< 2$ mg $\mathrm{l}^{-1}$ is determined each year, so the model hypoxia cutoff is set to 2 mg $\mathrm{l}^{-1}$ when determining length ($L$), and volume ($V$) is calculated using the empirical relationship $V = 0.003\,91L^2$ (Scavia *et al* 2006). In the GOM, hypoxia is reported as the area of hypoxic bottom water, with measurements taken just above the sediment water interface. Because the model simulates the entire sub-pycnocline layer and because available DO profiles show that when near-bottom DO is 2 mg $\mathrm{l}^{-1}$, average sub-pycnocline DO approaches 3 mg $\mathrm{l}^{-1}$, the GOM model hypoxia cutoff is set to 3 mg $\mathrm{l}^{-1}$. Hypoxic area ($A$) is calculated using the empirical relationship $A = 38.835L$ (Scavia and Donnelly 2007).

### 2.2. Data sources

We use spring total nitrogen (MT TN/d) loading data from the USGS to drive both models. Average January through May TN loads from the Susquehanna River (at Conowingo, MD gauging station) are used for CB and May TN loads from the Mississippi (at St Francisville) and Atchafalaya (at Melville) Rivers are used for the GOM (USGS 2007, 2009, 2010).

Model calibrations and tests are conducted using empirically measured hypoxic area (GOM) or volume (CB). GOM hypoxic area has been interpolated from near-bottom DO measurements collected by shelf-wide cruises in late July or early August (Rabalais *et al* 2002b, Rabalais 2009). Cruises have been conducted yearly since 1985 with the exception of 1989. Because the measured hypoxic area was potentially impacted by tropical storms in 1996, 1998, 2003, and 2005, we removed these years from both our calibration and test datasets (see Turner *et al* 2008) because the model is incapable of accounting for these extreme conditions. Such tropical storms can disrupt water-column stratification, mixing oxygenated water downward, and thus temporarily break the link between production and hypoxia observed in non-storm years. CB hypoxic volume is determined from DO profiles taken on four cruises in July and August each year since 1984 and sporadically before then (Hagy *et al* 2004, Chesapeake Bay Program (CBP) 2008). We use the July CBP cruise data from the consistent record since 1984 for model calibration and testing.

Initial oxygen deficit ($D_\mathrm{i}$) for CB was based on average July bottom-water oxygen concentration measured at stations in the mid-bay region (Scavia *et al* 2006, Chesapeake Bay

Program (CBP) 2008). The difference between saturated and measured oxygen concentration was used for mean $D_i$ in calibration years. For hindcasts and forecasts, $D_i$ was drawn from a normal distribution with average and standard deviation equal to that in the measured calibration data.

### *2.3. Hindcast and forecast tests*

We tested several model calibration algorithms to optimize model performance. Each test used a calibration dataset and a test dataset. Model performance was measured by assessing precision, accuracy, robustness, and sensitivity to system change, although some tests focus on a subset of these measures. Precision was assessed as the size of the coefficient of variation and the 95% credible interval (CI) of the model prediction. Accuracy was assessed as the percentage of observations in the test dataset that fell within the 95% CI of the model prediction. It is expected that this value could differ from 95% because the test dataset contains observations for years that are not included in the model calibration and thus the model is predicting outside its statistical sample. Robustness was based on the impact of individual years in the calibration data set on calibrated parameter values. Sensitivity to system change was assumed to be maximized when few, recent years of calibration data were used because averaging across larger numbers of data points decreases the impact of any given point on the average.

To test the impact of increasing the number of years in the calibration set on model precision and robustness (Test 1), calibrations began with the first three years of data and we progressively added years for successive calibrations (calibration dataset). To test precision, we used each calibrated model to predict hypoxia for each year in the full dataset (test dataset) and the average CV and 95% CI were calculated. To test robustness, we examined variation in parameter values over time from each calibration test. Model accuracy was quantified by calculating the percentage of observed hypoxic areas or volumes that fell within the model's 95% CI for that year's prediction. We repeated this test (Test 2) beginning with the three most recent years and adding years in reverse order.

Because the above comparisons are confounded by overlap between the calibration and test datasets, they were used only to narrow the range of years for which a more complete test was conducted. In these tests of accuracy, forecasts were conducted using 3, 5, and 7 year windows of calibration data (Test 3, range selected based on the results of Tests 1 and 2, see below). Precision was quantified by the CV of the hypoxia forecast in the year following each calibration and the average of these CVs across all calibrations using the same window size. The accuracy of these calibrations was tested by forecasting hypoxia in the year following each calibration window and calculating the percentage of observed hypoxic areas or volumes that fell within the forecast's 95% CI across all calibrations using that window (test dataset).

To prepare for forecasts where all model coefficients are to be held constant, we tested two methods of parameter calibration. In prior work with this model (Liu *et al* 2010, Liu and Scavia 2010), $v$ was estimated as the year-specific term $v_i$ and then forecasts used the mean and standard deviation of $v_i$ through time, ignoring the Bayesian fit parameter distribution. We compared this approach with one that estimated all parameters (including $v$) as constant distributions through time such that forecasts could be conducted directly from the calibrated distributions (Test 4). For both methods (using $v_i$ and $v$) in CB, the parameter $D_i$, which is not calibrated but calculated from observed values for calibration years, was estimated using the average and variation in the previously observed values.

### *2.4. Response curves*

Response curves of predicted hypoxic area or volume versus spring TN load, were constructed in the same way as hindcasts and forecasts but using evenly spaced spring TN loads spanning the observed range rather than exact historical loads (Scavia *et al* 2006, Liu *et al* 2010, Liu and Scavia 2010). As in prior publications (Liu *et al* 2010, Liu and Scavia 2010), response curves were constructed using 50% CIs to better constrain conditions in typical or average years.

## 3. Results

### *3.1. Full dataset calibration*

When the model is fit to the entire CB and GOM datasets, allowing $v_i$ to vary in each year and then averaging $v_i$ for hindcasts, hindcasting accuracy is high (100% and 80% of the observations are within the 95% CI of the hindcast, respectively, see right-most symbols associated with the full dataset in figures 1(a) and (c)). These percentages differ from 95% because of additional variability added to the model when the parameter $v_i$, and the parameter $D_i$ in the CB model, are averaged across years. Model precision, as measured by the CV of the predicted hypoxic region, is better for CB (33%) than for the GOM (41%) (see right-most symbols associated with the full dataset in figures 1(a) and (c)). Model parameters have mean values of $F = 0.91$, $k_d = 0.14$, $K = 0.58$, $v_A = 2.5$ (where $_A$ indicates the average across years), $D_A = 1.2$ for CB; and $F = 0.51$, $k_d = 0.006$, $k_r = 0.012$, $v_A = 0.64$ for the GOM (see right-most symbols associated with the full dataset in figures 1(b) and (d)). These values are consistent with previously published model calibrations and, as in prior calibrations, estimated process rates based on these parameters are consistent with observed rates (Scavia *et al* 2006, Scavia and Donnelly 2007, Liu *et al* 2010, Liu and Scavia 2010).

### *3.2. Effect of number of years in the calibration*

The precision and accuracy of the models calibrated to the full datasets represent a goal for calibrations using sub-datasets. However, they may not represent the best overall model calibration because using the full dataset ignores temporal trends and regime shifts within the system and thus sacrifices model sensitivity to system change. It also confounds calibration and test datasets, causing a possible overestimation of model accuracy in predicting novel conditions. So, we tested model performance by calibrating to subsets of data

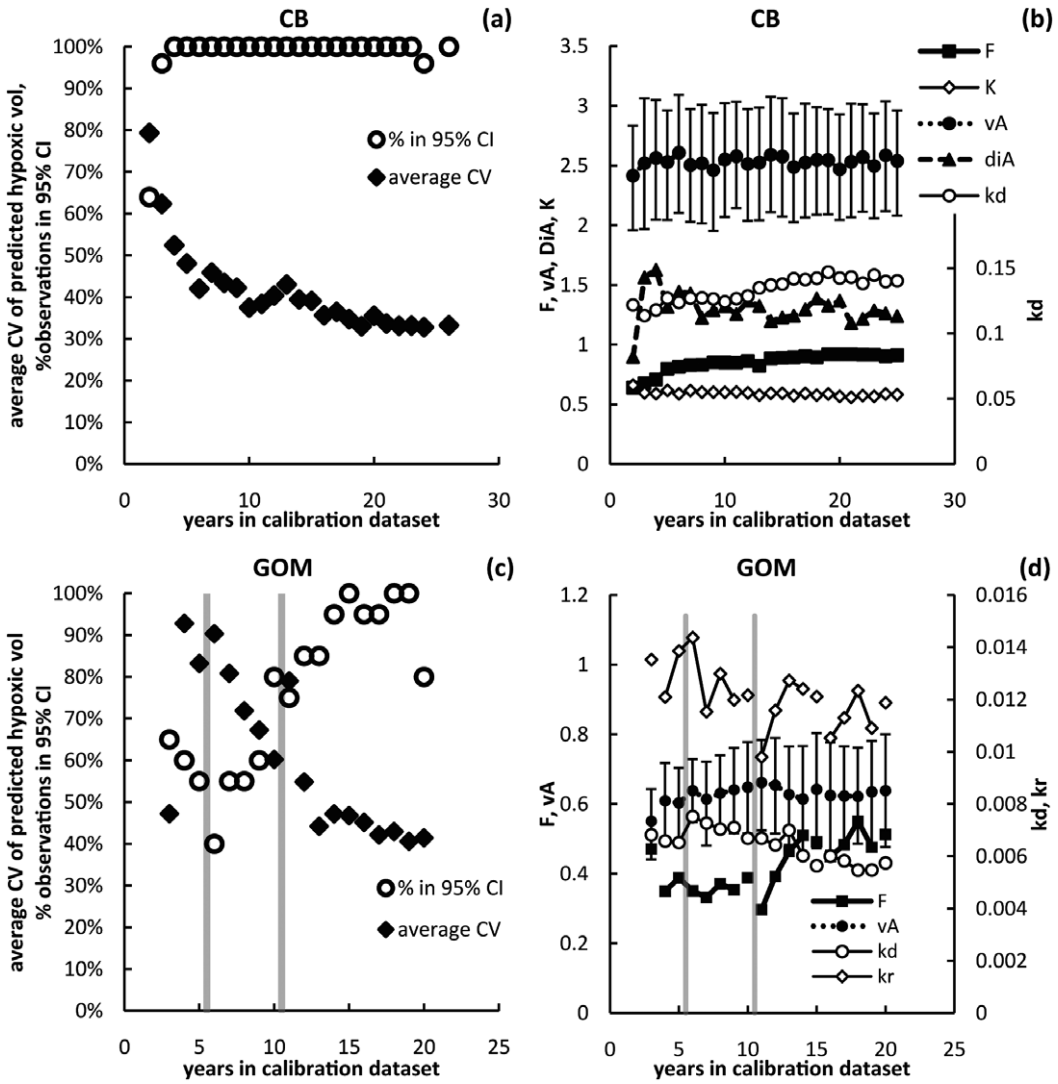


**Figure 1.** (Test 1) Average CV of predicted hypoxic volume or area and % observations in 95% CI using (a) a test dataset of Chesapeake Bay hypoxic volume from 1984–2008 and calibration datasets of increasing numbers of years starting in 1984 and (c) a test dataset of Gulf of Mexico hypoxic area from 1985–2009 and calibration datasets of increasing numbers of years starting in 1985. Parameter estimates (panels (b) and (d); means, with 50% CI bars for $v_A$) corresponding to the model calibrations used in panels (a) and (c). Note that results from calibrating to the full datasets are shown as the right-most symbols in each graph.

with increasing number of years. Beginning at the start of each dataset (Test 1), adding years causes a rapid improvement for the CB model performance in both precision and accuracy until about 5 years of data are used (figure 1(a)). Beyond this point, precision and accuracy asymptote toward values of the full dataset. Similarly, CB parameter values were highly variable in the beginning until about 5 years of data were used (figure 1(b)). Beyond 5 years, there was little change in values, despite known changes in system behavior, indicating that model calibration to long datasets loses sensitivity to these changes. These same patterns of precision, accuracy, and parameter variability were observed when calibration was started using the three most recent years and adding years in reverse order (Test 2, data not shown). Thus, using about 5 years of calibration data seems to offer an optimal combination of model precision, accuracy, and robustness, while avoiding a loss of sensitivity to system change in this system.

Determining optimal calibration for the GOM involves a greater compromise between precision and sensitivity. As in CB, adding calibration years (Test 1) causes a rapid improvement in model precision and accuracy (figure 1(c)). However, precision and accuracy continue to improve until about 15 years of calibration data. The model also maintained more sensitivity to system change with the addition of calibration years for longer datasets than the CB model (figure 1(d)). Parameter values, especially $F$ and $k_r$ continue to change as years are added up to at least 15 years of calibration data. When calibration started using the three most recent years and years were added in reverse order (Test 2, data not shown), model accuracy was high (100% of observations in the 95% CI) using even 3 years of data and remained at this level as years were added. Model precision improved quickly until about 5 years and then saturated, and parameter values varied in a similar pattern. Though results were more mixed

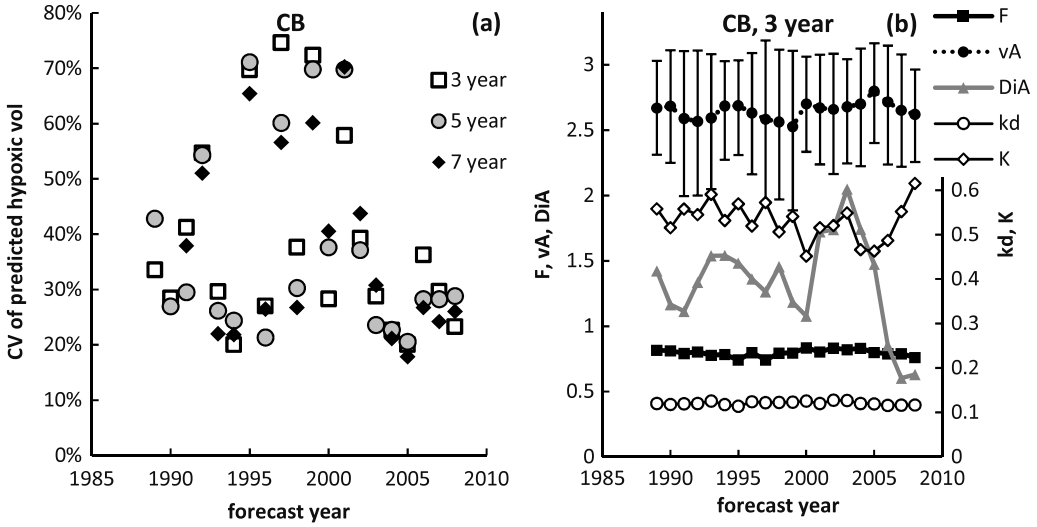


**Figure 2.** (Test 3) (a) CV of predicted hypoxic volume in CB for the year following each 3, 5, and 7 year calibration period. (b) Parameter estimates with error bars showing the 50% CI of average $v$ over the 3 year calibration period.

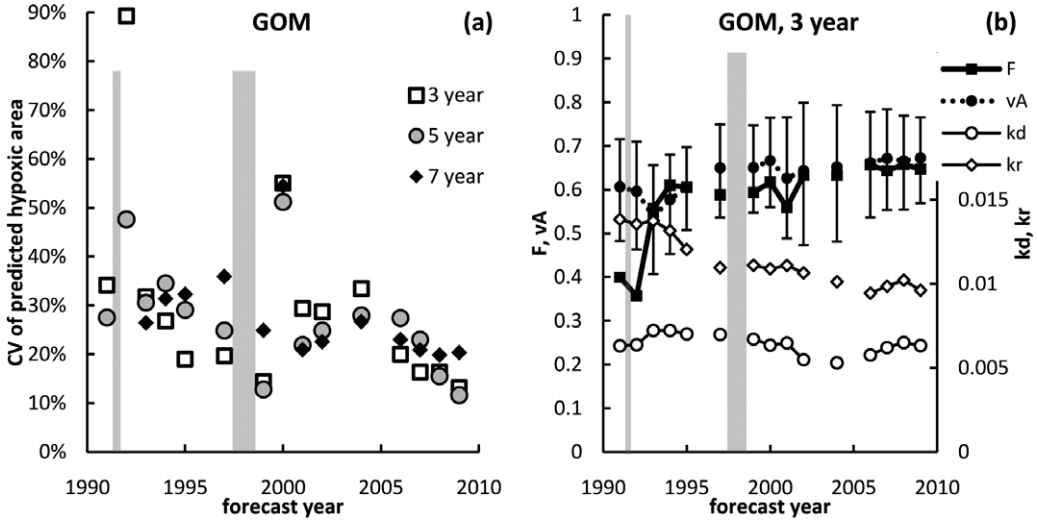


**Figure 3.** (Test 3) (a) CV of predicted hypoxic area in the GOM for the year following each 3, 5, and 7 year calibration period. (b) Parameter estimates with error bars showing the 50% CI of average $v$ over the 3 year calibration period.

for the GOM, with continued improvement in model precision and accuracy beyond the first 5 years of calibration data in the forward though not the backward calibration tests, we decided to further test models using 5 years of calibration data because using 15 or 20 years lost sensitivity to system changes which have been observed on shorter time scales (Turner *et al* 2008, Liu *et al* 2010).

### *3.3. Moving window calibrations: case 1—averaging* $v_i$

Previous applications of this model estimated $v_i$ for each year in a calibration dataset and then averaged it for forecasts. So we first test the moving window calibrations (Test 3) with this method and then compare it below to the case where $v$ is estimated as a constant over the calibration widow period (Test 4). Tests with 3, 5, and 7 year moving windows showed little difference in precision (CV) when forecasting CB (figure 2(a)) or GOM (figure 3(a)) hypoxia in the year following the calibration window and no overall improvement in precision using larger windows. A change in precision could indicate over or under fitting the model, but this does not seem to be taking place. For all window sizes, parameter values changed over time; however, parameter variability was highest using the smallest (3 year) window (figures 2(b) and 3(b)). This increased variability indicates higher model sensitivity to system state because more of the underlying variability is reflected in the parameters. Model accuracy was high for all window sizes in CB and decreases with window size in the GOM. In CB the per cent of observations in the 95% CI is 100%, 95%, and 100% for the 3, 5, and 7 year calibration periods, respectively, compared to 93%, 80%, and 73% for these calibrations in the GOM. Thus, the results support the use of a 3 year moving window.

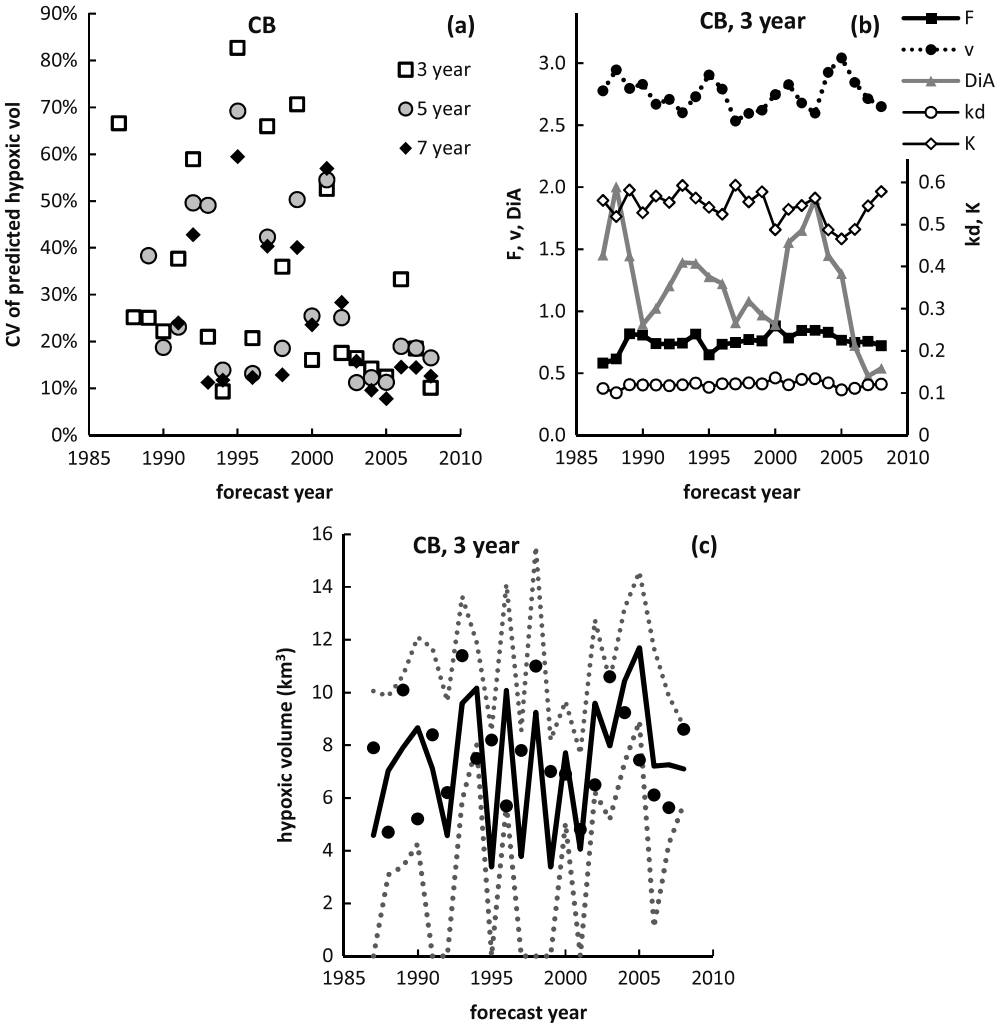


**Figure 4.** (Test 4) (a) CV of predicted hypoxic volume in CB for the year following each 3, 5, and 7 year calibration period. (b) Parameter estimates for the 3 year calibration period. (c) Forecast hypoxic volume (black line, forecast mean; gray dotted lines, forecast 95% CI) and observed hypoxic volume (black dots) for each forecast year using the 3 year calibration period.

Compared to using the full dataset, 3 year moving window calibrations resulted in the same accuracy for CB (100%) and improved accuracy for GOM (95% versus 80% of observations within the 95% CI). Average model precision (CV) for the 3 year moving window calibration was slightly poorer in CB (39% versus 33%), but was slightly improved in the GOM (30% versus 41%). Using a moving window allows the model precision to vary over time based on recent system variability. Precision is higher (lower CV) during periods of system stability, such as the late 1990s in the GOM, and lower (higher CV) following regime shifts (figure 3(a)).

### *3.4. Moving window calibrations: case 2—constant v*

Fitting $v_i$ for each year and then averaging it for forecasts introduces arbitrary variation into the model. As an alternative, we tested moving window calibrations of 3, 5, and 7 years fitting all parameters, including $v$, as constants (Test 4).

With 3, 5, and 7 year moving window calibrations, CB hypoxia forecast accuracy is lower than expected. Accuracy is highest for the 3 year window and decreases with increasing window size (82%, 70%, and 68% of observed hypoxic volume were within the model 95% CI, figure 4(c) compares forecast and observed hypoxia for the 3 year moving window calibration). There was very little difference in precision (CV) among window sizes (figure 4(a)) and no overall improvement in precision using larger windows. For all window sizes, parameter values changed over time; however, as in prior tests, parameter variability was highest using the smallest (3 year) window (figure 4(b)), indicating the highest model sensitivity to system state.

Test for the GOM resulted in lower accuracy, with 73%, 68%, and 46% of the observed hypoxic areas within the model's 95% CI for 3, 5, and 7 year windows, respectively (figure 5(c) compares forecast and observed hypoxia for the 3 year moving window calibration). There was very little difference in precision (CV) among window sizes (figure 5(a)), no overall improvement in precision using larger windows, and parameter values changed over time with the highest variability associated with the smallest (3 year) window (figure 5(b)).

The CV of predicted hypoxic area or volume varies with time in all moving window calibrations. However, the average

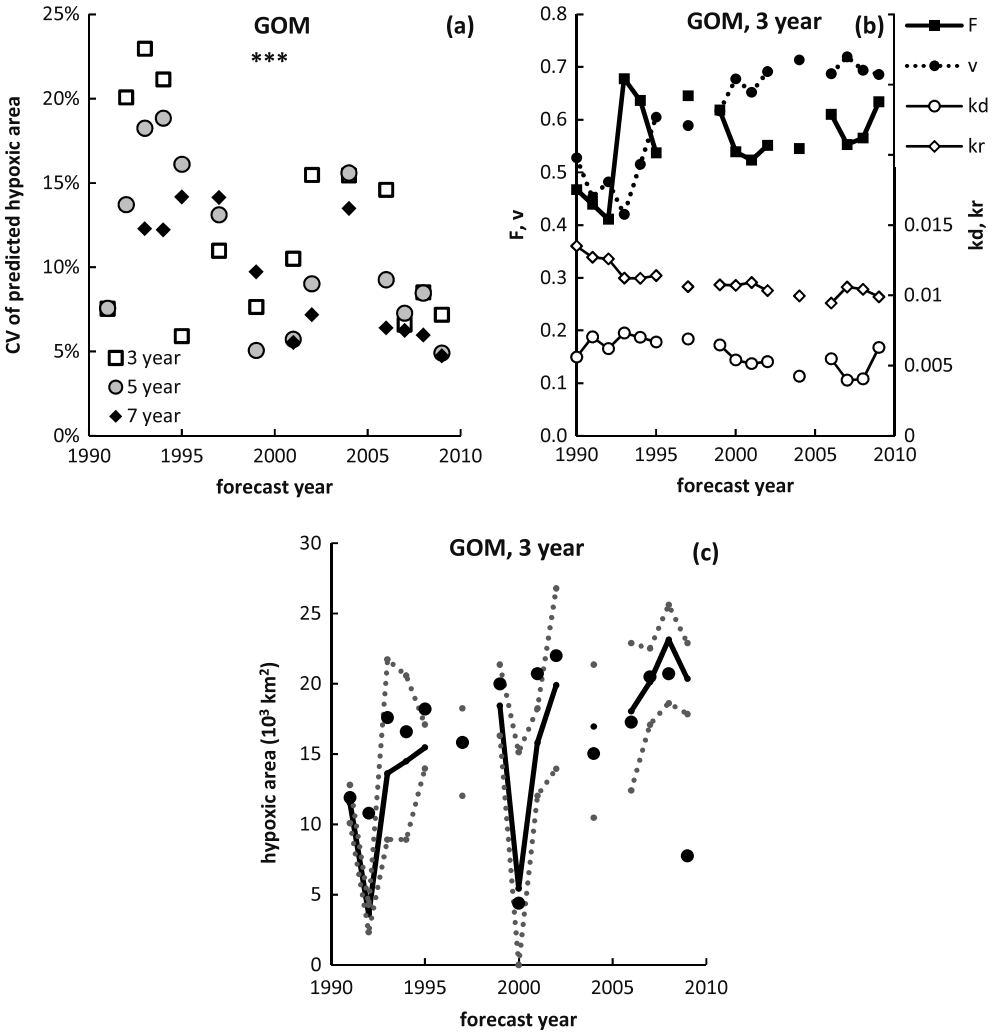


**Figure 5.** (Test 4) (a) CV of predicted hypoxic area in the GOM for the year following each 3, 5, and 7 year calibration period with all parameters treated as constants. *** In 2000, CV of predicted hypoxic area was 98%, 45%, and 54% for the 3, 5, and 7 year calibration periods, respectively, and is not graphed. (b) Parameter estimates for the 3 year calibration period. (c) Forecast hypoxic area (black line, forecast mean; gray dotted lines, forecast 95% CI) and observed hypoxic area (black dots) for each forecast year using the 3 year calibration period.

CV is improved by calibrating with a constant $v$ in both systems. Using the 3 year window, the CV for CB is improved from 39% to 33% and for GOM from 30% to 18%, compared to moving window tests averaging $v_i$. This is a substantial improvement in model precision. This increase in model precision is accompanied by a decrease in model accuracy. However, because the increased variability introduced into the model by averaging year-specific $v_i$ is not related to a specific mechanism or known ecological process, the lower forecasting accuracy is likely a better representation of true model performance. The SP model is a vast simplification of nature and the accuracy cost of using this model (95% − 82% = 13% for CB and 22% for the GOM) reflect unmodeled variation in these systems.

## 4. Discussion

The dataset for the GOM included years in three distinct and previously observed system states with varying sensitivity to hypoxia formation (Environmental Protection Agency (EPA) Science Advisory Board (SAB) 2007, Turner *et al* 2008, Greene *et al* 2009, Liu *et al* 2010). Similar changes in system state have been observed in CB, however, data limitations prevented us from including the historic CB system state (Hagy *et al* 2004, Liu and Scavia 2010) in the current model tests. Model accuracy was poorer for the GOM than for CB and one of the reasons could be the attempt to calibrate the model across multiple system states. Including multiple states is minimized when using short calibration windows and improved model accuracy in the shortest windows are a result. GOM model accuracy for the 3 year moving window calibration, fitting $v$ as a constant, is further improved to 78% when excluding calibrations that overlap more than one system state. Though such exclusions can only be identified *post facto*, this further supports use of shorter calibration windows to minimize including multiple system states.

Initial comparisons, adding years to the calibration dataset, starting from the oldest (figure 1) or most recent (data not shown) data, indicated that 5 or more years

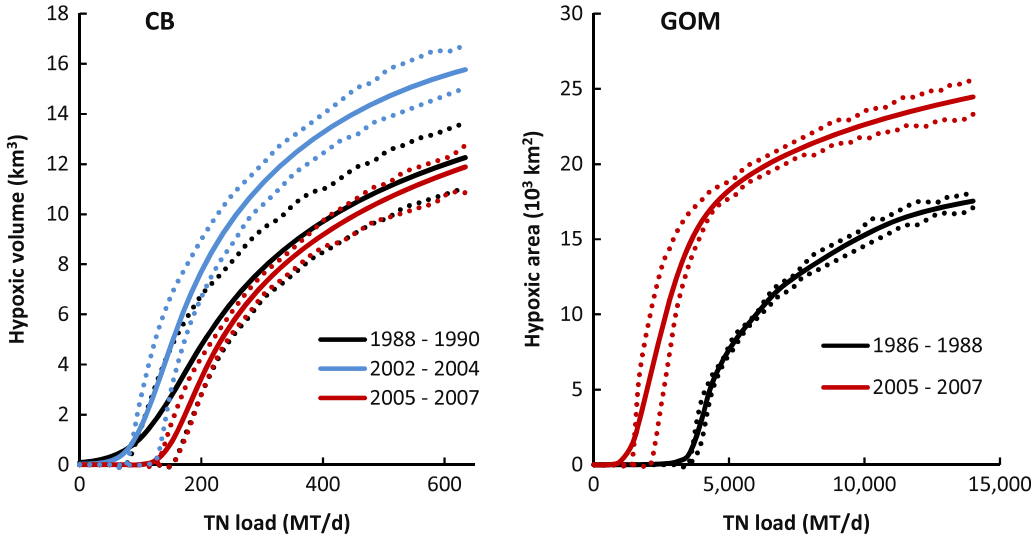


**Figure 6.** Hypoxia response curves for CB and the GOM based on years early in the historic record and recent years. The year ranges cited in each legend are the years in the respective calibration datasets. Solid lines show the average forecast response and dotted lines show the 50% CI of forecast response.

of calibration data were needed to optimize precision and accuracy. However, more extensive comparisons using 3, 5, and 7 year calibration datasets and employing more robust accuracy measures (by forecasting only the year following a calibration window) show that model performance is optimized with 3 year calibrations. These tests showed less of a tradeoff between model precision and accuracy than was expected. Model precision did not differ among window lengths in any comparison and in fitting either year-specific or constant $v$. Model accuracy was either higher in the shortest window or did not change with window size. As expected, the shortest window calibrations were the most adaptive to system change. This responsiveness is seen in both the improved model accuracy in the changing GOM and in increased parameter variability in both systems.

The tradeoff between precision and accuracy existed for both models calibrated using year-specific or constant $v$. We believe that fitting and averaging year-specific $v_i$ introduces an artifact that improves model accuracy by arbitrarily reducing precision. Thus, we propose that the optimal calibration for annual forecasting is to use a rather short (3 year), recent dataset treating all parameters as constants.

### *4.1. Forecasts*

Using the 3 year window calibrations, we developed load–hypoxia response curves for CB and the GOM for different periods in the historical record (figure 6). The GOM has undergone two shifts in sensitivity between the earliest (1986–8) and most recent (2005–7) calibration periods, and the resulting increase in sensitivity can be seen across the TN range.

Though the primary CB regime shift appears to have occurred before the start of our dataset (Hagy *et al* 2004, Kemp *et al* 2005, Scavia *et al* 2006, Conley *et al* 2009), CB appears to be undergoing a gradual increase in sensitivity to nutrient loads from 1983 through at least 2005 (Liu and Scavia 2010, Scully 2010). This increasing sensitivity is reflected in our parameter estimates (figure 4(b)) and in response curves using 1988–90 and 2002–4 calibration datasets (figure 6). Between these periods, hypoxia sensitivity increased, especially at high TN loads. Parameter values for recent years trend back toward, and even beyond, those earlier in the dataset. Accordingly, the most recent (2005–2007) response curve shows decreased sensitivity. At high TN loads, the curve resembles the 1988–90 'low sensitivity' case, and it appears to have even lower sensitivity at lower TN loads.

These changes are driven mostly by changes in the parameters $F$, $v$, $K$, and $D_i$. $F$ and $v$ increase between the first two periods and then decrease again before the final period. Sensitivity analyses (not shown) indicate that increases in $F$ tend to increase sensitivity at all TN loads while increases in $v$ increases sensitivity more at high TN loads. At the same time $K$ decreased between the first two time periods while remaining relatively unchanged between the second and third. Like increases in $v$, decreases in $K$ tend to increase sensitivity at high TN loads. Finally, the measured parameter $D_i$ remained constant between the first two periods but decreased between the second and third. Decreases in $D_i$ lead to decreases in sensitivity at low TN loads. This measured decrease in DO deficit in recent years may indicate a release from oxygen stress further down bay.

Using models calibrated with the three most recent years, or in the case of the GOM the three most recent years that were not impacted by severe tropical storms, provides a consistent method for annual forecasts that is relatively robust to regime shifts and changes in system sensitivity. However, changes in system sensitivity still pose a significant challenge for developing long-term scenarios—that is, in setting nutrient load targets, which response curve is most appropriate? Such long-term forecasts require assumptions about the future system sensitivity to hypoxia formation. Will

the system continue to follow the most recent curve, will it revert to a former sensitivity (as may be happening in CB), or will it become even more sensitive? We suggest this public policy challenge is best met with ensemble modeling using the family of response curves with curve selection weighted based on expert judgment and acceptable risk. For example, if a given hypoxia level were deemed ecologically or socially unacceptable, any response curve that predicted hypoxia above this level at certain nutrient loadings could be weighted higher over that loading range based on the precautionary principle. Alternately, evidence of system recovery to a lower sensitivity state could shift the weighting of curves toward those with lower sensitivity while still maintaining some weight on other observed curves.

The model presented here is primarily focused on forecasts of interest for hypoxia management and has also been used to explore system level trends in hypoxia sensitivity (Liu *et al* 2010). Both simpler and more complex models have also been applied to the GOM and CB systems and each model type yields different insights into the physical, watershed, and biological controls of hypoxia as well as its impacts on individual organisms, food-webs, and biogeochemistry (Peña *et al* 2010).

## 5. Conclusions

The forecasting ability of a simple hypoxia model with Bayesian incorporation of parameter uncertainty and variability for GOM and CB was optimized by calibration to short (3 year), recent datasets. This calibration window approach was used to assess the tradeoff between incorporating adequate system variability into model parameterization and the ability to track gradual (in CB) and abrupt (in the GOM) ecosystem changes in hypoxia sensitivity to nutrient loads. We propose use of this moving window calibration method for future short-term (annual) forecasts. The underlying changes in system sensitivity pose a great challenge to the long-term forecasting and additional work, using Bayesian weighting among families of models or incorporation of more complex model features, coupled with climate models, is likely needed.

## Acknowledgments

This work is contribution number 136 of the Coastal Hypoxia Research Program and was supported in part by grant NA05NOS4781204 from NOAA's Center for Sponsored Coastal Ocean Research and by the University of Michigan Graham Sustainability Institute. We appreciate the insight and advice provided by Yong Liu.

## References

Bianchi T S, DiMarco S F, Cowan 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 *Sci. Total Environ.* **408** 1471–84

Bierman V J, Hinz S C, Zhu D W, Wiseman W J, Rabalais N N and Turner R E 1994 A preliminary mass-balance model of primary productivity and dissolved-oxygen in the Mississippi river plume inner Gulf shelf region *Estuaries* **17** 886–99

Boesch D F, Brinsfield R B and Magnien R E 2001 Chesapeake Bay eutrophication: scientific understanding, ecosystem restoration, and challenges for agriculture *J. Environ. Quality* **30** 303–20

Bronmark C, Brodersen J, Chapman B B, Nicolle A, Nilsson P A, Skov C and Hansson L A 2010 Regime shifts in shallow lakes: the importance of seasonal fish migration *Hydrobiologia* **646** 91–100

Cerco C F 1995 Response of Chesapeake Bay to nutrient load reductions *J. Environ. Eng.* **121** 298–310

Cerco C F and Cole T 1993 3-dimensional eutrophication model of Chesapeake Bay *ASCE J. Environ. Eng.* **119** 1006–25

Chapra S C 1997 *Surface Water-Quality Modeling* (New York: McGraw-Hill)

Chesapeake Bay Program (CBP) 2008 unpublished data

Childs C R, Rabalais N N, Turner R E and Proctor L M 2002 Sediment denitrification in the Gulf of Mexico zone of hypoxia *Mar. Ecol. Prog. Ser.* **240** 285–90

Childs C R, Rabalais N N, Turner R E and Proctor L M 2003 Sediment denitrification in the Gulf of Mexico zone of hypoxia *Mar. Ecol. Prog. Ser.* **247** 310 (erratum)

Conley D J, Carstensen J, Vaquer-Sunyer R and Duarte C M 2009 Ecosystem thresholds with hypoxia *Hydrobiologia* **629** 21–9

Diaz R J and Rosenberg R 2008 Spreading dead zones and consequences for marine ecosystems *Science* **321** 926–9

Environmental Protection Agency (EPA) Science Advisory Board (SAB) 2007 *Hypoxia in the Gulf of Mexico* (available online at: http://www.epa.gov/sab/panels/hypoxia_adv_panel.htm) (accessed June 2010)

Gelman A and Hill J 2007 *Data Analysis Using Regression and Multilevel/Hierarchical Models* (New York: Cambridge University Press)

Gill J 2002 *Bayesian Methods: A Social and Behavioral Sciences Approach* (Boca Raton, FL: Chapman and Hall/CRC)

Goolsby D A, Battaglin W A, Aulenbach B T and Hooper R P 2001 Nitrogen input to the Gulf of Mexico *J. Environ. Quality* **30** 329–36

Greene R M, Lehrter J C and Hagy J D 2009 Multiple regression models for hindcasting and forecasting midsummer hypoxia in the Gulf of Mexico *Ecol. Appl.* **19** 1161–75

Hagy J D 2002 Eutrophication, hypoxia and trophic transfer efficiency in Chesapeake Bay *PhD* University of Maryland

Hagy J D, Boynton W R, Keefe C W and Wood K V 2004 Hypoxia in Chesapeake Bay, 1950–2001: long-term change in relation to nutrient loading and river flow *Estuaries* **27** 634–58

Higgins S N and Zanden M J V 2010 What a difference a species makes: a meta-analysis of dreissenid mussel impacts on freshwater ecosystems *Ecol. Monographs* **80** 179–96

Justić D, Bierman V J, Scavia D and Hetland R D 2007 Forecasting Gulf's hypoxia: the next 50 years? *Estuaries Coasts* **30** 791–801

Justić D, Rabalais N N and Turner R E 2003 Simulated responses of the Gulf of Mexico hypoxia to variations in climate and anthropogenic nutrient loading *J. Mar. Syst.* **42** 115–26

Justić D, Rabalais N N, Turner R E and Wiseman W J 1993 Seasonal coupling between riverborne nutrients, net productivity and hypoxia *Mar. Pollut. Bull.* **26** 184–9

Kemp W M *et al* 2005 Eutrophication of Chesapeake Bay: historical trends and ecological interactions *Mar. Ecol. Prog. Ser.* **303** 1–29

Liu Y, Evans M A and Scavia D 2010 Gulf of Mexico hypoxia: exploring increasing sensitivity to nitrogen loads *Environ. Sci. Technol.* **44** 5836–41

Liu Y and Scavia D 2010 Analysis of the Chesapeake Bay hypoxia regime shift: insights from two simple mechanistic models *Estuaries Coasts* **33** 629–39

Lunn D J, Thomas A, Best N and Spiegelhalter D 2000 WinBUGS—a Bayesian modeling framework: concepts, structure, and extensibility *Stat. Comput.* **10** 325–37

Mississippi River/Gulf of Mexico Watershed Nutrient Task Force 2008 *Gulf Hypoxia Action Plan 2008 for Reducing, Mitigating, and Controlling Hypoxia in the Northern Gulf of Mexico and Improving Water Quality in the Mississippi River Basin* (Washington, DC: USEPA Office of Wetlands, Oceans, and Watersheds) (Action Plan available at: www.epa.gov/owow_keep/msbasin/actionplan.htm)

Peña M A, Katsev S, Oguz T and Gilbert D 2010 Modeling dissolved oxygen dynamics and hypoxia *Biogeosciences* **7** 933–57

Penta B *et al* 2009 Using coupled models to study the effects of river discharge on biogeochemical cycling and hypoxia in the northern Gulf of Mexico *OCEANS 2009 MTS/IEEE Biloxi-Marine Technology for our Future: Global and Local Challenges* (New York, NY: IEEE) pp 1–7 (available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5422347&isnumber=5422059)

Rabalais N N 2009 personal communication

Rabalais N N 2006 Oxygen depletion in the Gulf of Mexico adjacent to the Mississippi river *Gayana (Concepc.)* **70** 73–8

Rabalais N N, Atilla N, Normandeau C and Turner R E 2004 Ecosystem history of Mississippi river-influenced continental shelf revealed through preserved phytoplankton pigments *Mar. Pollut. Bull.* **49** 537–47

Rabalais N N and Turner R E 2001 *Coastal Hypoxia: Consequences for Living Resources and Ecosystems* (Washington, DC: American Geophysical Union)

Rabalais N N, Turner R E, Dortch Q, Justić D, Bierman V J and Wiseman W J 2002a Nutrient-enhanced productivity in the northern Gulf of Mexico: past, present and future *Hydrobiologia* **475** 39–63

Rabalais N N, Turner R E and Scavia D 2002b Beyond science into policy: Gulf of Mexico hypoxia and the Mississippi river *Bioscience* **52** 129–42

Rabalais N N, Turner R E, Sen Gupta B K, Boesch D F, Chapman P and Murrell M C 2007 Hypoxia in the northern Gulf of Mexico: does the science support the plan to reduce, mitigate, and control hypoxia? *Estuaries Coasts* **30** 753–72

Rabalais N N, Turner R E and Wiseman W J 2002c Gulf of Mexico hypoxia, aka 'The dead zone' *Annu. Rev. Ecol. Syst.* **33** 235–63

Rabalais N N, Turner R E, Wiseman W J and Dortch Q 1998 Consequences of the 1993 Mississippi river flood in the Gulf of Mexico *Regul. Rivers-Res. Manag.* **14** 161–77

Rabalais N N, Wiseman W J and Turner R E 1994 Comparison of continuous records of near-bottom dissolved-oxygen from the hypoxia zone along the Louisiana coast *Estuaries* **17** 850–61

Renaud M L 1986 Hypoxia in Louisiana coastal waters during 1983—implications for fisheries *Fishery Bull.* **84** 19–26 (available at: http://fishbull.noaa.gov/841/renaud.pdf)

Scavia D and Donnelly K A 2007 Reassessing hypoxia forecasts for the Gulf of Mexico *Environ. Sci. Technol.* **41** 8111–7

Scavia D, Justić D and Bierman V J 2004 Reducing hypoxia in the Gulf of Mexico: advice from three models *Estuaries* **27** 419–25

Scavia D, Kelly E L A and Hagy J D 2006 A simple model for forecasting the effects of nitrogen loads on Chesapeake Bay hypoxia *Estuaries Coasts* **29** 674–84

Scavia D, Rabalais N N, Turner R E, Justić D and Wiseman W J 2003 Predicting the response of Gulf of Mexico hypoxia to variations in Mississippi river nitrogen load *Limnol. Oceanogr.* **48** 951–6

Scheffer M and van Nes E H 2007 Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size *Hydrobiologia* **584** 455–66

Scully M E 2010 The importance of climate variability to wind-driven modulation of hypoxia in Chesapeake Bay *J. Phys. Oceanogr.* **40** 1435–40

Stow C A and Scavia D 2009 Modeling hypoxia in the Chesapeake Bay: ensemble estimation using a Bayesian hierarchical model *J. Mar. Syst.* **76** 244–50

Turner R E, Rabalais N N and Justić D 2006 Predicting summer hypoxia in the northern Gulf of Mexico: riverine N, P, and Si loading *Mar. Pollut. Bull.* **52** 139–48

Turner R E, Rabalais N N and Justić D 2008 Gulf of Mexico hypoxia: alternate states and a legacy *Environ. Sci. Technol.* **42** 2323–7

Turner R E, Rabalais N N, Swenson E M, Kasprzak M and Romaire T 2005 Summer hypoxia in the northern Gulf of Mexico and its prediction from 1978 to 1995 *Mar. Environ. Res.* **59** 65–77

USGS 2007 *Chesapeake Bay: River Input Monitoring Program: Loads* (available online at: http://va.water.usgs.gov/chesbay/RIMP/loads.html) (accessed 10 May 2010)

USGS 2009 *Streamflow and Nutrient Delivery to the Gulf of Mexico* (available online at: http://toxics.usgs.gov/hypoxia/mississippi/flux_ests/delivery/index.html) (accessed 14 June 2010)

USGS 2010 *Streamflow and Nutrient Delivery to the Gulf of Mexico for October 2009 to May 2010 (Preliminary)* (available online at: http:/toxics.usgs.gov/hypoxia/mississippi/oct_jun/index.html) (accessed 20 June 2010)

Walker N D and Rabalais N N 2006 Relationships among satellite chlorophyll a, river inputs, and hypoxia on the Louisiana continental shelf, Gulf of Mexico *Estuaries Coasts* **29** 1081–93

Wang L X and Justić D 2009 A modeling study of the physical processes affecting the development of seasonal hypoxia over the inner Louisiana–Texas shelf: circulation and stratification *Cont. Shelf Res.* **29** 1464–76

Zhang J *et al* 2010 Natural and human-induced hypoxia and consequences for coastal areas: synthesis and future development *Biogeosciences* **7** 1443–67