2017

Ensemble modeling informs hypoxia management in the northern Gulf of Mexico

D Scavia
I Bertani
DR Obenour
RE Turner
DR Forrest

Virginia Institute of Marine Science

See next page for additional authors

Follow this and additional works at: https://scholarworks.wm.edu/vimsarticles

Part of the Aquaculture and Fisheries Commons

Recommended Citation

Scavia, D; Bertani, I; Obenour, DR; Turner, RE; Forrest, DR; and Katin, A, "Ensemble modeling informs hypoxia management in the northern Gulf of Mexico" (2017). VIMS Articles. 764.
https://scholarworks.wm.edu/vimsarticles/764

This Article is brought to you for free and open access by W&M ScholarWorks. It has been accepted for inclusion in VIMS Articles by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Authors
D Scavia, I Bertani, DR Obenour, RE Turner, DR Forrest, and A Katin
A large region of low-dissolved-oxygen (DO) bottom waters (hypoxia; DO < 2 mg L\(^{-1}\)) forms nearly every summer in the northern Gulf of Mexico because of nutrient inputs from the Mississippi River Basin and water column stratification. Policymakers developed goals to reduce the area of hypoxic extent because of its ecological, economic, and commercial fisheries impacts. However, the goals remain elusive after 30 y of research and monitoring and 15 y of goal-setting and assessment because there has been little change in river nitrogen concentrations. An intergovernmental Task Force recently extended to 2035 the deadline for achieving the goal of a 5,000-km\(^2\) 5-y average hypoxic zone and set an interim load target of a 20% reduction of the spring nitrogen loading from the Mississippi River by 2025 as part of their adaptive management process. The Task Force has asked modelers to reassess the loading reduction required to achieve the 2035 goal and to determine the effect of the 20% interim load reduction. Here, we address both questions using a probabilistic ensemble of four substantially different hypoxia models. Our results indicate that, under typical weather conditions, a 59% reduction in Mississippi River nitrogen load is required to reduce hypoxic area to 5,000 km\(^2\). The interim goal of a 20% load reduction is expected to produce an 18% reduction in hypoxic area over the long term. However, due to substantial interannual variability, a 25% load reduction is required before there is 95% certainty of observing any hypoxic area reduction between consecutive 5-y assessment periods.

Significance

The number of coastal hypoxia areas is spreading worldwide, with severe environmental and societal impacts. The second-largest hypoxic zone occurs in the northern Gulf of Mexico, where anthropogenic nutrient load is a key driving factor, as in many coastal waters. We address policy-relevant questions raised by Gulf stakeholders and decision-makers using an ensemble approach that integrates results from multiple models. Through development of a rigorous framework to propagate intramodel and intermodel uncertainty into the ensemble, we provide policymakers with the response of hypoxic area to a range of different nitrogen load reduction scenarios, with corresponding probabilistic statements that allow for quantitative risk assessment of alternative policy strategies.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission. Freely available online through the PNAS open access option.

To whom correspondence should be addressed. Email: scavia@umich.edu.
probabilistic ensembles to help bridge the gap between research and policymaking (25–29). Our work advances that goal by formally propagating model uncertainty into our ensemble predictions to allow for quantitative risk assessment and policies that are robust within realistic uncertainty ranges.

Results

The U-M model (30) explained 69% of the variability in hypoxic area extent over 1985–2011. This model (and all others) was subjected to a leave-one-year-out cross-validation, in which it explained 45% of hypoxic area variability. This model’s annual hypoxia forecasts have compared well with measurements, especially for years without storms or high winds ($R^2 = 70\%$; Fig. 2). Its load–response curve indicates that a reduction in hypoxia to 5,000 km$^2$ requires a 58% [95% credible interval (CI): 49–70%] decrease in May total nitrogen (TN) load relative to the 1980–1996 average (Table 1). A 20% reduction would produce a hypoxic extent of 15,000 km$^2$ (95% CI: 13,500–16,500 km$^2$). The NCSU model, calibrated for 1985–2011, explains 75% of the variability in BWDO and 70% of the variability in hypoxic area. The cross-validation performed similarly to the full calibration ($R^2 = 72\%$ for BWDO) (31). This model was first used for hypoxia forecasts in 2015 (Fig. 2) and estimates that reaching the hypoxia goal requires a 56% (95% CI: 50–62%) decrease in spring bioavailable N load (Table 1). A 20% reduction would produce a hypoxic extent of 12,400 km$^2$ (95% CI: 10,800–14,000 km$^2$). The LSU/LUMCON model explains 92% of the variation in hypoxic extent since 2000, after removing 5 y with high wind/storms. The cross-validation on the same calibration dataset showed similar model performance ($R^2 = 87\%$). Its annual forecasts have been fairly accurate since 2002 (Fig. 2).

Without the five high-wind/storm years, its forecasts accounted for 56% of the observed variability, and the model indicates that a 56% (95% CI: 50–64%) reduction in the May nitrite+nitrate (NO) load is needed to meet the hypoxia goal (Table 1). A 20% reduction would produce a hypoxic extent of 15,600 km$^2$ (95% CI: 14,400–16,800 km$^2$). The VIMS model explained 64% of the variability in hypoxic area between 1985 and 2015, and the cross-validation explained 52% of the interannual variation. A simplified version of the model has been used to produce blind forecasts since 2014 (Fig. 2). This model indicates that an 80% (95% CI: 50%) reduction in the May NO load is needed to meet the hypoxia goal (Table 1), while a 20% load reduction would result in a hypoxic area of 12,900 km$^2$ (95% CI: 10,700–15,000 km$^2$).

The ensemble modeling results indicate that, under typical weather conditions, a 59% reduction in N load from the 1980–1996 baseline would be required to reach the goal. The 20% reduction interim goal would result in 13,900 km$^2$ (95% CI: 11,100–16,400 km$^2$), corresponding to an 18% reduction in hypoxic extent over the long-term, and a 1% reduction compared with the most recent 5-y period (Fig. 3 and Table 1).

Discussion

While multiple models have been used to inform hypoxia management in the past, in this work, multiple models are synthesized within a probabilistic framework to develop a “consensus” estimate of how the system will respond and to quantify the uncertainty in that estimate. By developing the ensemble, we explore and quantify uncertainty due to differences in model structure and inputs (32), which is not captured by the individual models themselves. Our four models differ substantially in mechanistic form, type of N load driver, internal sources of DO demand, and criteria for selecting calibration datasets. For example, the VIMS and LSU/LUMCON models are linear regressions with different assumptions on the form of the relationship between nutrients, stratification, and hypoxia, whereas the U-M and NCSU models are based on different mechanistically derived relationships. In addition, while the VIMS and LSU/LUMCON models use NO as the primary nutrient driver, the U-M model uses TN, and the NCSU model uses an estimation of bioavailable N (Fig. 1). The U-M and LSU/LUMCON models use nutrient load as the primary driver, and, as a result, their response curves are steeper than the VIMS curve, which attributes more variation in hypoxia to freshwater discharge and wind. The NCSU model provides a compromise in that load, freshwater discharge, and wind are incorporated into its mechanistic framework. The U-M, NCSU, and VIMS models use the period of record from 1985 for calibration, whereas the LSU/LUMCON model is calibrated from 2000 on, under the hypothesis that the nutrient–hypoxia relationship has changed over time. Finally, the U-M, NCSU, and VIMS models define outlier years according to quantitative meteorological criteria, whereas the LSU/LUMCON model defines outliers based on sea conditions during the midsummer sampling cruise (Methods).

Despite these differences, the individual model results show similar responses to load reductions (Fig. 3), and the mean ensemble results indicate that a 59% reduction from the 1980–1996 average N load is needed to meet the Task Force goal (Table 1). The worst-case scenario in our ensemble results (upper bound of predictive intervals) indicates that even an 80% load reduction (which is likely infeasible) may not meet the goal.
Table 1. Model estimates of the load reduction required to meet the 5,000-km² Task Force goal and estimates of the hypoxic area expected in response to a 20% load reduction

<table>
<thead>
<tr>
<th>Model</th>
<th>Load reduction needed for 5000 km²</th>
<th>Hypoxia area expected for a 20% load reduction,</th>
<th>Hypoxia area, % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-M</td>
<td>58 (49–70)</td>
<td>15.0 (13.5–16.5)</td>
<td>1,000 km² (95% CI)</td>
</tr>
<tr>
<td>NCSU</td>
<td>56 (50–62)</td>
<td>12.4 (10.8–14.0)</td>
<td></td>
</tr>
<tr>
<td>LSU/LUMCON</td>
<td>56 (50–64)</td>
<td>15.6 (14.4–16.8)</td>
<td></td>
</tr>
<tr>
<td>VIMS</td>
<td>80 (50–)</td>
<td>12.9 (10.7–15.0)</td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>59 (50–)</td>
<td>13.9 (11.1–16.4)</td>
<td></td>
</tr>
</tbody>
</table>

(Table 1 and Fig. 3). However, this scenario represents the tail of the probability distribution and is unlikely. We estimate that the probability that hypoxia would be <5,000 km² under an 80% load reduction is 87%. The 59% reduction is larger than the 45% reduction called for in the most recent Action Plan (17), but within the range of previous individual models (5, 30, 31, 33–36). This percentage is also higher than recommendations made for other eutrophic systems. For example, a multimodel effort (20) recommended a 40% phosphorus (P) load reduction for Lake Erie under the Great Lakes Water Quality Agreement (37); the Chesapeake Bay agreement calls for a 25% reduction in N load and a 24% reduction in P (38); and the Neuse River Strategy calls for a 30% reduction in N load (23).

The Task Force also set an interim N load reduction target of 20% by 2025 (17). Our analysis suggests that a 20% reduction may not have a measurable effect within the next 5–10 y. According to the ensemble modeling results, a 20% load reduction will result in a 13,900-km² mean hypoxic region (95% CI: 11,100–16,400 km²), which is not significantly different from the current 5-y average of 14,024 km² (Table 1). However, over the long term (as opposed to a 5-y average, which is subject to annual anomalies), a 20% reduction would produce an 18% reduction in hypoxic area, relative to model-estimated current conditions (Fig. 3).

An important contribution of this study, relative to previous Gulf hypoxia modeling efforts, is the characterization of model structure and input uncertainty, in addition to model parameter and predictive error uncertainty (32). We developed two types of predictive intervals that distinguish different sources of uncertainty and which are relevant to different types of management questions (Fig. 3). The credible interval for the mean of the four models quantifies the deterministic uncertainty associated with the response curve (Fig. 3, shaded area), which can be partitioned between within-model uncertainty (i.e., parameter uncertainty) and across-model uncertainty (i.e., input and structural uncertainty). If hypoxia observations are taken over a large number of years (i.e., n > 30), we expect the observed mean will fall within this credible interval. For load reductions of 0–50%, the across-model uncertainty accounts for 60% of the total uncertainty on average. For an 80% reduction, the across-model uncertainty accounts for 70%, reflecting the increasing influence of structural uncertainty. At large load reductions, the VIMS predictions are substantially larger than the other models (Fig. 3) because the model attributes more variation in hypoxia to river discharge and wind forcing, and less to N load. Greater ensemble uncertainty is expected at large load reductions where observational data are scarce and the response of the system is less certain. This is especially true for empirical models, where extrapolating predictions near or outside of the calibration range requires caution. Across-model structural uncertainties underscore the need for continued research to address mechanistic uncertainties (discussed below). In general, our results show that across-model uncertainty represents a considerable portion of the overall deterministic uncertainty, and accounting for this uncertainty will lead to more dependable, albeit wider, credible intervals.

The 5-y predictive interval (Fig. 3, gray lines) addresses the additional uncertainty that arises when estimating a mean value from a limited sample size (e.g., the 5-y period used to assess the Hypoxia Task Force goal) (17). The portion of overall uncertainty due to model prediction error can be used to estimate the minimum load reduction required to achieve a statistically significant decrease in hypoxia across 5-y periods. Our results suggest that a 25% load reduction would be required to be 95% certain to observe a reduction in hypoxic area when comparing any two 5-y periods. This supports our finding that the 20% interim load reduction may not produce a measurable effect over 5- to 10-y time scales.

All models are simplified representations of the real world, and the models used here are no exception. Although N has historically been considered the main nutrient driver of hypoxia in the Gulf, the relative roles of N vs. P limitation of primary production, and how that ratio may change seasonally and spatially, remains unclear (39). Based on studies suggesting P limitation on the Gulf shelf at critical time periods, the Task Force adopted a dual-nutrient strategy, with the same percent reduction goals for N and P (5, 17). While our models are not designed to answer this question, other modeling supports the dual strategy, with percentage reductions for N and P loads in line with our N load recommendations (40, 41). Expanding the models to incorporate both N and P loads, coupled with better understanding of the stoichiometry regulating primary productivity, has the potential to significantly improve our ability to assess different N and P load-reduction scenarios. Until then, we believe that current evidence points toward a dual-nutrient strategy as the most prudent management approach (39).

There is uncertainty in how internal nutrient loads and sediment oxygen demand (SOD) will modulate the system’s response to external load reductions, which presents a challenge for long-term...
forecasting. Currently, only the NCSU model accounts for this process by predicting changes in SOD as a function of long-term average nutrient loads. There is also uncertainty in predicting the impacts of climate change on meteorological and hydrological patterns and how these impacts may affect the system’s susceptibility to hypoxia. For example, the frequency, intensity, and timing of droughts and storms are predicted to shift as a result of global climate change (42), with potential interacting effects on the timing and amount of nutrient delivery, the intensity and duration of stratification, the solubility of oxygen, and biogeochemical cycling (43, 44). More research is needed to enhance our predictive understanding of how these and other processes may cause shifts in internal feedbacks and complex nonlinearities in response to management actions (45–47). However, until these structural uncertainties are decreased, we think that it is critical to develop probabilistic models and probabilistic ensembles of models whose results are communicated to decision-makers to support adaptive strategies that consider uncertainties in system response.

Despite these uncertainties, our results show that the hypoxia response to N load reductions is robust across substantially different and independent models, providing increased confidence that the load reduction proposed will achieve management goals. The strong relationship between nutrient loading and hypoxia illustrated here is also consistent with the results from sediment cores studies that indicate little to no hypoxia on the Gulf shelf before the escalation of nitrogen fertilizer use in the mid-1900s (48). However, it matters little whether the load reduction target is 30%, 45%, or 59% if insufficient resources are in place to make even modest reductions. Recent analyses by the Department of Agriculture (49, 50) comparing modeled nutrient losses between the present state and a hypothetical past without conservation practices indicate that there is some level of conservation effectiveness. However, while there are undoubtedly significant lag times between action on the land and changes in loads (51, 52), river nitrate concentrations have not declined since the 1980s (53, 54), and the current 5-y running average nitrate load to the Gulf is not significantly different from the 1980–1996 baseline (ref. 17; https://toxics.usgs.gov/hypoxia/mississippi/oct_jun/index.html) after US Farm Bill conservation programs have spent more than $28 billion in the 20 Mississippi Basin states since 1995 (55).

Most large-scale environmental restoration efforts are structured within an adaptive management framework (56)—one that sets goals, takes action, measures progress, and adjusts actions if needed. Models documented in loads of subpycnocline DO concentration downstream from the outflows of the Mississippi and Atchafalaya Rivers. Organic matter loads are calculated from the May TN loads by converting TN to algal carbon and the associated DO demand based on Redfield and respiratory ratios (75). The length of the hypoxic zone is the sum of all locations along the longitudinal profile where DO is below the hypoxic threshold. Hypoxic length is converted to area through empirical relationship (30). The model is calibrated through Bayesian inference over the period 1985–2011 (30). The model has been compared with others (18) and was used to explore N vs. P control (76), provide guidance for the Gulf Action Plans (13, 14), and explore impacts of climate change (77). The load-response curve was developed with parameter estimates for normal weather years.

U-M Model. This is an adaptation of a river model (73), predicting DO concentrations downstream from point sources of organic matter loads. It uses mass balance equations to estimate the DO consumed during organic matter decomposition and to predict the DO deficit (74) in longitudinal profiles of subpycnocline DO concentration downstream from the outflows of the Mississippi and Atchafalaya Rivers. Organic matter loads are calculated from the May TN loads by converting TN to algal carbon and the associated DO demand based on Redfield and respiratory ratios (75). The length of the hypoxic zone is the sum of all locations along the longitudinal profile where DO is below the hypoxic threshold. Hypoxic length is converted to area through empirical relationship (30). The model is calibrated through Bayesian inference over the period 1985–2011 (30). The model has been compared with others (18) and was used to explore N vs. P control (76), provide guidance for the Gulf Action Plans (13, 14), and explore impacts of climate change (77). The load-response curve was developed with parameter estimates for normal weather years.

NCSU Model. This Bayesian biophysical model predicts BWDO concentrations in western and eastern segments of the Louisiana-Texas shelf that are separated by the Atchafalaya River outfall (31). BWDO predictions are converted to hypoxic area by using empirical relationships between mean BWDO and hypoxic area, both of which are determined from a geostatistical model (70). The model is a steady-state solution to mechanistic, mass-balance equations (31), and is calibrated within a Bayesian framework that accounts for prior information on model parameters. It systematically characterizes parameter and prediction uncertainty (78). Important prior information includes SOD (79) and vertical carbon flux (80) measurements. The east and west shelf sections are segmented into upper and lower layers, where upper layers accommodate transport of freshwater and nutrients across the shelf. The transport of flow and loads along the shelf are regulated by long-term, along-shore wind velocities obtained from NOAA weather stations. Monthly water flow and load estimates are linearly interpolated to determine bioavailable N loads for consecutive 30-d averaging periods leading up to the beginning of the annual shelf-wide cruises. Surface-layer nutrients are subject to an effective settling rate that accounts for organic matter production and sinking. The DO in the bottom layer is controlled by water column DO demand, SOD, and reaeration which is a function of the surface layer. Hypoxia is defined by the SOD, which is a function of the surface layer. Hypoxia is defined by the SOD, which is a function of the subpycnocline DO concentration downstream from the outflows of the Mississippi and Atchafalaya Rivers. Organic matter loads are calculated from the May TN loads by converting TN to algal carbon and the associated DO demand based on Redfield and respiratory ratios (75). The length of the hypoxic zone is the sum of all locations along the longitudinal profile where DO is below the hypoxic threshold. Hypoxic length is converted to area through empirical relationship (30). The model is calibrated through Bayesian inference over the period 1985–2011 (30). The model has been compared with others (18) and was used to explore N vs. P control (76), provide guidance for the Gulf Action Plans (13, 14), and explore impacts of climate change (77). The load-response curve was developed with parameter estimates for normal weather years.
fluxes to the sediment over the long-term (74, 81, 82). To quantify this change in SOD, we use a relationship developed for coastal estuaries (83):

\[ SOD = a \left( \frac{L_c}{1 + KL_h} \right)^b \]

where SOD is sediment oxygen demand (mol O_2 m^{-2} y^{-1}), \( L_c \) is the organic carbon deposition rate (mol C m^{-2} y^{-1}), \( h \) is the thickness of the lower layer (m; \( \sim 20 \) m for shelf sediments), and \( a, b, \) and \( k \) are parameters with mean values of 0.76, 0.79, and 0.000079, respectively. The average vertical carbon flux in the original model (31) was 5.48 mol C m^{-2} y^{-1} and SOD was 3.76 mol C m^{-2} y^{-1}.

Based on Eq. 1, the modeled carbon flux produces an SOD, which is 29.8% lower than the SOD determined by the model. Therefore, Eq. 1 is multiplied by 1.425 for our load-reduction scenarios. Some adjustment is to be expected, given the uncertainty in Eq. 1 and the modeled benthic fluxes, and the differences between the Gulf shelf and the estuaries used to establish Eq. 1 (31, 82). To create the load–response curve, the model was run for 21 normal weather years with actual wind and flow conditions, and a range of N load reductions. Thus, the response curves are based on average hydrological and meteorological conditions under various load reductions.

**LSULUMCON Model.** This model is a regression of summer hypoxic zone size as a function of May nitrate load, similar to what has been done for lakes (84, 85). The model assumes that the hypoxic zone is driven mostly by the spring N load and that other influences cause variations around a relatively stable baseline. Previous studies (36, 86) suggested that the relationship between load and hypoxic extent changed over time due to the system becoming more sensitive to N loading as the result of incremental changes in organic matter accumulated in the sediments (87), increases in the nitrate fraction of the total N pool, and long-term climate change. As a result, the model coefficients varied over time, but stabilized after 15 yr (36). Therefore, the model is calibrated for years 2000–2015. The N loading data are log10-transformed to linearize the curvilinear relationship observed between load and the estimates of hypoxic extent.


**VIMS Model.** This regression model uses river NO concentration, river discharge, and wind to estimate hypoxic area (35). To create load–response curves for the present analysis, the river NO concentration was varied to achieve load reductions, and river discharge and wind were held constant at the 2000–2015 averages.

**Creating an Ensemble.** The ensemble load–response curve uses individual models’ curves and their parameter uncertainty. For each load reduction, we determined each model’s predictive distribution based on the model’s mean and 95% CIs. The ensemble predictive distribution for each load was based on Monte Carlo (MC) sampling from the predictive distributions of the individual models (10,000 samples from each model), thereby obtaining an estimate of the deterministic uncertainty (i.e., including input, structural, and parameter uncertainty) associated with mean ensemble predictions. Negative predictions obtained through MC sampling were set to zero. We compared ensemble deterministic uncertainty with that obtained by accounting for both deterministic uncertainty and the uncertainty associated with model prediction error. We determined each model’s prediction error variance from the last 10y (2005–2015; excluding 2009 as unusual weather) of blind-forecast errors, using a maximum-likelihood estimation. Whenever blind forecasts were not available, then the forecast error was set to the model’s cross-validation error, or to the mean absolute error of the U-M and LSULUMCON blind forecasts for that year, whichever was greater. Because the Task Force goal is based on a 5y running average, we divided the models’ prediction error variance by 5 to estimate the uncertainty associated with a 5y mean prediction. We then added each model’s prediction error variance to the variance associated with deterministic uncertainty to get the overall predictive variance. The ensemble distribution was generated through MC sampling from the overall predictive distributions of the individual models, as described above.

**ACKNOWLEDGMENTS.** This work was supported in part by NOAA Grants NA12OAR4320071 and NA09NOS782004 and the University of Michigan Graham Sustainability Institute.


73. Scavia et al.