AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2015JC011577

Key Points:
- Model intercomparison of three hypoxia models of the northern Gulf of Mexico is presented
- Bottom water temperature and bottom boundary layer thickness are important for hypoxia simulation
- Overall stratification strength does not explain model-to-model differences in hypoxic conditions

Supporting Information:
- Supporting Information S1

Correspondence to:
K. Fennel,
Katja.Fennel@dal.ca

Citation:
Fennel, K., A. Laurent, R. Hetland, D. Justić, D. S. Ko, J. Lehrter, M. Murrell, L. Wang, L. Yu, and W. Zhang (2016), Effects of model physics on hypoxia simulations for the northern Gulf of Mexico: A model intercomparison, *J. Geophys. Res. Oceans*, *121*, 5731–5750, doi:10.1002/2015JC011577.

Received 16 DEC 2015
Accepted 11 JUL 2016
Accepted article online 14 JUL 2016
Published online 11 AUG 2016

Effects of model physics on hypoxia simulations for the northern Gulf of Mexico: A model intercomparison

Katja Fennel[1], Arnaud Laurent[1], Robert Hetland[2], Dubravko Justić[3], Dong S. Ko[4], John Lehrter[5], Michael Murrell[5], Lixia Wang[3], Liuqian Yu[1], and Wenxia Zhang[1,2]

[1]Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada, [2]Department of Oceanography, Texas A&M University, College Station, Texas, USA, [3]Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, Louisiana, USA, [4]Oceanography Division, Naval Research Laboratory, Stennis Space Centre, Hancock County, Mississippi, USA, [5]Gulf Ecology Division, U.S. Environmental Protection Agency, Gulf Breeze, Florida, USA

Abstract A large hypoxic zone forms every summer on the Texas-Louisiana Shelf in the northern Gulf of Mexico due to nutrient and freshwater inputs from the Mississippi/Atchafalaya River System. Efforts are underway to reduce the extent of hypoxic conditions through reductions in river nutrient inputs, but the response of hypoxia to such nutrient load reductions is difficult to predict because biological responses are confounded by variability in physical processes. The objective of this study is to identify the major physical model aspects that matter for hypoxia simulation and prediction. In order to do so, we compare three different circulation models (ROMS, FVCOM, and NCOM) implemented for the northern Gulf of Mexico, all coupled to the same simple oxygen model, with observations and against each other. By using a highly simplified oxygen model, we eliminate the potentially confounding effects of a full biogeochemical model and can isolate the effects of physical features. In a systematic assessment, we found that (1) model-to-model differences in bottom water temperatures result in differences in simulated hypoxia because temperature influences the uptake rate of oxygen by the sediments (an important oxygen sink in this system), (2) vertical stratification does not explain model-to-model differences in hypoxic conditions in a straightforward way, and (3) the thickness of the bottom boundary layer, which sets the thickness of the hypoxic layer in all three models, is key to determining the likelihood of a model to generate hypoxic conditions. These results imply that hypoxic area, the commonly used metric in the northern Gulf which ignores hypoxic layer thickness, is insufficient for assessing a model's ability to accurately simulate hypoxia, and that hypoxic volume needs to be considered as well.

1. Introduction

The occurrence of coastal hypoxia has risen dramatically over the past 50 years due to increases in anthropogenic nutrient loading to coastal waters [*Diaz and Rosenberg*, 2008]. Initiatives are underway to reduce nutrient inputs in order to alleviate the negative effects of hypoxia [e.g., *Hypoxia Task Force*, 2008], but the response of hypoxia to nutrient load reductions is not straightforward to predict. This is, in large part, due to the pronounced variability of physical factors that influence hypoxia generation in coastal systems. Separating the relative importance of physical and biological processes in influencing the development of hypoxia is important and can be difficult. Realistic, 3-dimensional circulation-hypoxia models are useful tools for this purpose because they can elucidate the mechanisms underlying hypoxia generation; they allow one to distinguish between physical and biological influences; they can be used to undertake scenario simulations in order to assess the likely effects of various nutrient reductions, and can be used as prediction tools.

The largest hypoxic zone in U.S. coastal waters forms every summer on the Texas-Louisiana shelf in the northern Gulf of Mexico (average area: $15{,}000 \pm 5000$ km^2) due to nutrient and freshwater inputs from the Mississippi/Atchafalaya River System [*Rabalais et al.*, 2002]. The Mississippi is the fifth largest river in the world supplying about 5×10^{11} m^3 yr^{-1} of freshwater and 5×10^{10} mol N yr^{-1} of nitrogen [*Aulenbach et al.*, 2007]. Regression analyses have shown that variations in atmospheric forcing, circulation patterns, and freshwater discharge in addition to spring nutrient load are important in determining the extent of hypoxic conditions on the shelf [*Scavia et al.*, 2003; *Turner et al.*, 2006; *Greene et al.*, 2009; *Forrest et al.*, 2011; *Feng et al.*, 2012]. According to *Forrest et al.* [2011], variations in spring nutrient load, although significantly

correlated with summer hypoxic area, explain only 24% of the interannual variability in hypoxic area. When other factors like directional wind strength and freshwater discharge are incorporated as independent variables, the correlation improves markedly [*Forrest et al*., 2011]. This illustrates the importance of variations in atmospheric forcing and circulation patterns in determining duration and spatial extent of hypoxic conditions on the shelf.

A number of numerical models have been developed for the region that aim to realistically represent coastal circulation coupled with the biogeochemical processes influencing dissolved oxygen. *Hetland and DiMarco* [2008] presented a circulation model coupled with relatively simple parameterizations of oxygen sources and sinks in order to investigate the spatial differences in relative importance of water column and sediment oxygen sinks. *Fennel et al*. [2011, 2013] coupled a nitrogen-based biogeochemical model to the circulation model of *Hetland and DiMarco* [2008] and analyzed patterns of phytoplankton variability, and hypoxia sensitivity to lateral and bottom boundary condition choices. *Justić and Wang* [2014] presented a biogeochemical model based on the circulation model described in *Wang and Justić* [2009] and examined patterns of temporal and spatial variability in hypoxia. *Lehrter et al*. [2013] used a circulation model in combination with measurements of nutrients and organic matter to derive nitrogen and phosphorus budgets for the shelf. *Laurent et al*. [2012] expanded the model of *Fennel et al*. [2011] by explicitly including phosphorus as a nutrient and investigated patterns of nitrogen versus phosphorus limitation. In *Laurent and Fennel* [2014], this expanded model was used to assess the effect of P-limitation on hypoxia generation and to conduct nutrient reduction scenario simulations. *Mattern et al*. [2013] performed an uncertainty analysis showing that uncertainty in certain model inputs, particularly in physical forcing, results in large uncertainties in the simulated hypoxia extent using the model of *Fennel et al*. [2013]. Applying the same model, *Feng et al*. [2014] documented mechanistically how the duration of upwelling-favorable wind influences hypoxia generation, and *Yu et al*. [2015a] developed an oxygen budget for the hypoxic region. *Yu et al*. [2015b] presented highly simplified parameterizations of dissolved oxygen dynamics coupled to the same circulation model in order to investigate the relative importance of several physical factors on hypoxia. Collectively these studies have greatly advanced our understanding of the hypoxia generation in the northern Gulf of Mexico.

In addition to increased mechanistic understanding of the factors and mechanisms determining hypoxia generation, various stakeholders, including the U.S. National Oceanic and Atmospheric Administration (NOAA) and the U.S. Environmental Protection Agency (EPA), are interested in developing a predictive capability for the northern Gulf of Mexico that can be used in operational mode. Questions arise about which model specifications and what complexity best fit the desired use and which aspects have to be scrutinized most closely in such an operational system. It is useful to compare the models that are being developed for this system in order to understand which aspects are most important for accurate hypoxia prediction. This is the purpose of the Shelf Hypoxia project within the Coastal Ocean Modeling Testbed (COMT), an initiative that aims to facilitate the transition of models, tools, and expertise from academic scientists to operational centers [*Luettich et al*., 2013]. Here we report on results from the Shelf Hypoxia project with focus on comparing the physical model aspects that affect a model's ability to accurately simulate hypoxia. To this end, we have coupled the same dynamical equations for dissolved oxygen dynamics into three different circulation models (in a total of four different implementations). These include two implementations of the Regional Ocean Modeling System (ROMS) [*Haidvogel et al*., 2008], an implementation of the Finite Volume Coast Ocean Model (FVCOM) [*Chen et al*., 2006], and an implementation of the Navy Coastal Ocean Model (NCOM) [*Martin*, 2000].

We compare dissolved oxygen as simulated by the different models against observations, and analyze why the models differ in their simulated hypoxic extent. The principal oxygen supply processes in these models are air-sea gas exchange and vertical mixing; oxygen sinks are respiration in the water column and uptake of oxygen by the sediment. Since the same water column respiration rate is used in all models and the same temperature-dependent parameterization of oxygen uptake by the sediment, differences in the simulated evolution of oxygen can only arise from model-to-model differences in oxygen supply and in bottom water temperature. We separate and analyze the relative importance of these two effects with the help of diagnostic simulations. Our main conclusions are that the temperature of bottom water and the thickness of the bottom boundary layer are the most important physical determinants for how likely a model is to develop hypoxia in this system. Thus these two factors have to be faithfully represented by the circulation model underlying an operational hypoxia prediction system.

2. Model Description

Our intercomparison includes four implementations of three different hydrodynamic models: a large and a small-domain implementation of the Regional Ocean Modelling System (ROMS; http://myroms.org) [*Haidvogel et al*., 2008], an implementation of the Finite Volume Coast Ocean Model (FVCOM) [*Chen et al*., 2006], and an implementation of the Navy Coastal Ocean Model (NCOM) [*Martin*, 2000]. The model bathymetries and domains are shown in Figure 1.

ROMS is a finite difference model with terrain-following s-coordinates in the vertical and curvilinear coordinates in the horizontal direction. It uses a fourth-order horizontal advection scheme for tracers and a third-order upwind scheme for momentum [*Haidvogel et al*., 2008]. Both of our ROMS implementations have 30 terrain-following vertical layers, with higher resolution near the surface and bottom. The small-domain model has 1 km resolution near the Mississippi Delta increasing to ~20 km at the southwest corner of the domain; it has been described in more detail by *Hetland and DiMarco* [2008] and *Marta-Almeida et al*. [2013], and has been used in a number of biological modeling studies including in *Fennel et al*. [2011, 2013], *Laurent et al*. [2012], *Feng et al*. [2014], *Laurent and Fennel* [2014], and *Yu et al*. [2015a, 2015b]. The large-domain ROMS model has 0.5 km horizontal resolution near the coast increasing to 1–2 km at the outer slope and has been described in more detail by *Zhang et al*. [2012a,2012b,2014]. Both ROMS domains are nested within the larger-scale, data-assimilative Hybrid Coordinate Ocean Model (HYCOM) for the whole Gulf of Mexico described by *Wallcraft et al*. [2009]. The HYCOM parent model uses isopycnal coordinates in the open ocean and terrain-following coordinates on the shelf, has 20 vertical layers and a horizontal resolution of 4 km. Physical properties of the parent model influence the ROMS child models through radiation conditions for temperature, salinity, and baroclinic velocities along the boundaries, and by imposing SSH and barotropic velocities from the parent model along the boundaries. In addition, a 6 grid-cell wide nudging zone is used along the boundaries in which temperature, salinity, and baroclinic velocities are nudged

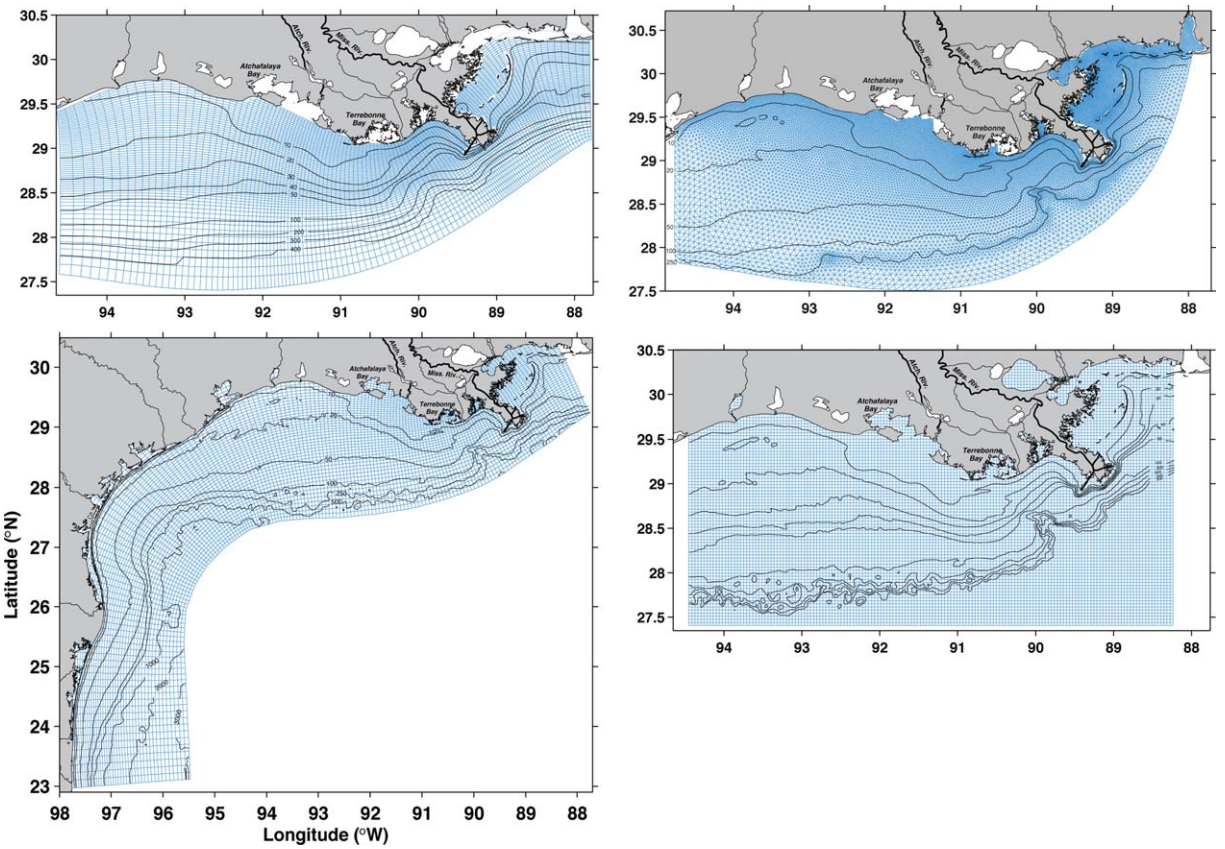

Figure 1. Horizontal model grids of the four circulation models: (top left) small ROMS domain, (top right) FVCOM, (bottom left) large ROMS domain, and (bottom right) NCOM. Only every second gridline is shown for the NCOM domain and only every fourth for the large ROMS domain.

to those of the parent model with a time scale of 8 h at the boundary decaying to zero at the inner edge of the nudging zone. Both ROMS implementations are initialized with physical fields from the parent model and forced with 3-hourly NCEP NARR winds and climatological heat and freshwater fluxes from *da Silva* [1994a,1994b].

NCOM is a finite difference model with hybrid sigma/z-level vertical coordinates and a third-order upwind scheme for advection of tracers and momentum [*Martin*, 2000]. The regional NCOM implementation used here has a rectangular grid with a horizontal resolution of ~1.9 km and 35 vertical levels [*Lehrter et al*., 2013]. Of the 35 vertical layers, 20 are equally spaced sigma layers from the surface down to 100 m, used in order to better resolve topography and processes on the shelf, followed by z-level coordinates below 100 m. Data assimilation of satellite altimeter sea surface height and multichannel sea surface temperature was applied [*Ko et al*., 2008], however, the data assimilation has minimum impact on the shelf due to application of a vertical weighting function with almost no weight in the upper 100 m. The regional NCOM is nested within the data-assimilative Intra-Americas Sea Nowcast/Forecast System (IASNFS) [*Ko et al*., 2003; *Ko and Wang*, 2014], which is a regional implementation of NCOM with 19 sigma layers from the surface down to a depth of 138 m, followed by z-levels below with a total of 41 vertical levels, and has a horizontal grid resolution of about 6 km in the Gulf of Mexico. Along the open boundaries of the child NCOM, upwind advection is imposed for temperature, salinity, and baroclinic velocities normal to the boundary from the parent IASNFS. SSH and barotropic velocities from IASNFS are imposed with a forced radiation condition. A 15 grid-cell wide nudging zone along boundaries is implemented in the child model in which temperature and salinity are nudged to those of IASNFS with a time scale of 10 days at the boundary decaying to zero at the inner edge of the nudging zone. The child NCOM is initialized with physical fields from the parent IASNFS and forced with three hourly wind and air pressure from the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) [*Hodur*, 1997] and heat fluxes from the Navy Operational Global Atmospheric Prediction System (NOGAPS) [*Rosmond*, 1992].

FVCOM is a finite volume model with sigma coordinates in the vertical direction and an unstructured triangular grid in the horizontal direction. It uses a second-order upwind scheme for tracer advection and momentum [*Kobayashi et al*., 1999; *Hubbard*, 1999] and a second-order Runge-Kutta time-stepping scheme for time integration [*Chen et al*., 2006]. The FVCOM implementation, described in more detail in *Wang and Justić* [2009] and *Justić and Wang* [2014], uses 30 vertical sigma layers with uniform resolution from surface to bottom. Its unstructured horizontal grid has a resolution of 1 km near the coast and ~10 km at southern open boundary of the model domain. The FVCOM model is nested within the same IASNFS (described above) as the regional NCOM implementation. The FVCOM child model uses radiation conditions for temperature, salinity, and baroclinic velocities along the boundaries. A 3 grid-cell wide nudging zone is implemented along the open boundaries in which temperature, salinity, and baroclinic velocities are nudged to those of the IASNFS parent model with a nudging time-scale of 3.8 h at the boundary decaying to zero at the inner edge of the nudging zone. The FVCOM child model is initialized with physical fields from the IASNFS parent model, and forced with 6 hourly winds and heat fluxes from the Navy's Operational Global Atmospheric Prediction System Model (NOGAPS; http://www.srh.noaa.gov/ssd/nwpmodel/html/nogaps.htm).

All models use a Mellor-Yamada turbulence closure scheme, ROMS and FVCOM the level 2.5 scheme [*Mellor and Yamada*, 1982] and NCOM the level 2 scheme [*Mellor and Yamada*, 1974]. None of the model implementations includes tides, as tides are very small in the northern Gulf of Mexico [*DiMarco and Reid*, 1998]. In NCOM, the vertical attenuation of shortwave radiation, k_d, is parameterized using temporally and spatially varying attenuation coefficients derived from satellite observations of ocean color [*Schaeffer et al*., 2011]. Both ROMS implementations initially used a constant k_d representative of clear water. In order to assess the effect of these different treatments of shortwave radiation on hypoxia generation, we also performed simulations with the small-domain ROMS model using the satellite-derived k_d of *Schaeffer et al*. [2011]. River freshwater inputs are prescribed using daily transport measurements from the U.S. Army Corps of Engineers from Tarbert Landing for the Mississippi River, and from Simmesport for the Atchafalaya River (http://www.mvn.usace.army.mil/Missions/Engineering/StageandHydrologicData.aspx).

For this study, we implemented the simple oxygen model described by *Yu et al*. [2015b] in all four of our child models. This simple oxygen model consists of a parameterization for air-sea gas exchange of oxygen [*Wanninkhof*, 1992], an empirically derived parameterization of net water column respiration based on

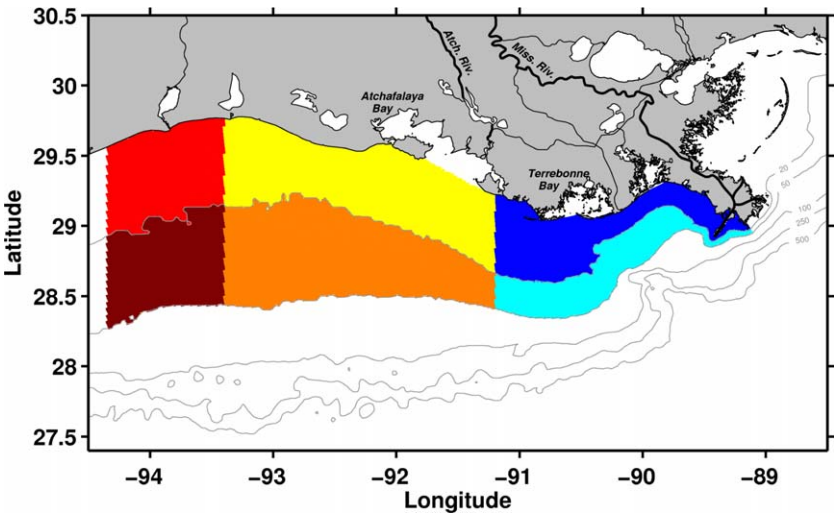

Figure 2. Analysis regions from east to west inshore of the 20 m isobath: Mississippi inshore (blue), Atchafalaya inshore (yellow), Far Field inshore (red), and between the 20 and 50 m isobaths: Mississippi midshelf (light blue), Atchafalaya midshelf (orange), and Far Field midshelf (brown). For all models, hypoxic extent is calculated in the combined region only.

observations by *Murrell et al*. [2013], and an empirical oxygen and temperature-dependent parameterization of sediment oxygen consumption or SOC [*Hetland and DiMarco*, 2008]. The net water column respiration term is constant in time and in the vertical direction, but varies horizontally with water depth. Inshore of the 20 m isobath, net water column respiration is negative (implying an oxygen source); the largest positive net water column respiration (representing an oxygen sink) occurs between the 20 and 30 m isobaths [see *Yu et al*., 2015b, Figure 2]. Oxygen initial conditions are based either on the World Ocean Atlas (for ROMS and NCOM) or on historical observations (for FVCOM). Along the models' open boundaries oxygen concentrations are clamped to monthly fields from the World Ocean Atlas (for ROMS and NCOM) or to saturation concentrations (for FVCOM). Sensitivity studies by *Mattern et al*. [2013] have shown that hypoxia simulations are insensitive to perturbations of the biological boundary conditions. Given this and the short spin-up time for dissolved oxygen on the shelf, we are confident that small model-to-model differences in initial and boundary conditions for dissolved oxygen are inconsequential for the comparisons we present. All four models were run for the years 2005 and 2006. We chose this period because availability of observations is comparably good.

Implementing the same oxygen parameterizations in all our circulation models allows us to focus on the effects of model physics (e.g. density stratification, vertical mixing) on hypoxia simulations. In addition to physical factors affecting oxygen, three oxygen source and sink terms are parameterized (air-sea gas exchange, respiration in the water column, SOC). Of these three terms, only SOC can lead to meaningful model-to-model differences. Since low oxygen concentrations occur near the bottom, model-to-model differences in air-sea gas exchange, which could arise from differences in surface temperature and salinity, are negligible for hypoxia generation. Net water column respiration is directly imposed and identical in all models. In addition to differences in physical oxygen supply, the only other term through which physical properties can affect hypoxia is SOC, because it depends on bottom water temperatures. In order to distinguish the potential effect of bottom water temperatures on SOC from those due to density stratification and vertical mixing, we performed diagnostic simulations where the influence of bottom water temperature on SOC is removed. This was done by imposing a constant temperature of 28°C, which is the average bottom water temperature in summer in the initial ROMS simulation, in the SOC parameterization. Density stratification is not affected by this modification. Output from all model simulations is available from the COMT server at http://comt.sura.org/thredds/catalog/comt2/gom_hypoxia/catalog.html.

For model analysis, we defined six regions in which we compare the simulated average oxygen concentration to observations (Figure 2). We also calculated correlation coefficients, *R*, root-mean-square errors, *rmse*, and biases, *b*, between observed values y_i $(i=1,\ldots n)$ and their simulated counterparts x_i $(i=1,\ldots n)$ as follows:

$$R=\frac{\sqrt{\sum_{i=1}^{n}(\bar{x}-x_i)(\bar{y}-y_i)}}{\sqrt{\sum_{i=1}^{n}(\bar{x}-x_i)^2}\sqrt{\sum_{i=1}^{n}(\bar{y}-y_i)^2}},$$

$$rmse=\sqrt{\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{n}},$$

and

$$b=\frac{\sum_{i=1}^{n}(x_i-y_i)}{n}.$$

The observational data set includes profiles of temperature, salinity, and dissolved oxygen and was compiled from the following sources: the Louisiana Universities Marine Consortium hypoxia cruises (see http://www.gulfhypoxia.net/), the Environmental Protection Agency's cruises described in *Lehrter et al*. [2009] and *Murrell et al*. [2013], the Mechanisms Controlling Hypoxia project (see http://hypoxia.tamu.edu/), and the summer groundfish survey's by the National Ocean and Atmosphere Administration's Southeast Area Monitoring and Assessment Program (http://www.ncddc.noaa.gov/hypoxia/products/). The data set is available in NetCDF format at http://comt.sura.org/thredds/catalog/comt_1_archive_full/shelf_hypoxia/Observations/catalog.html.

3. Results

3.1. Dissolved Oxygen in Bottom Waters and Hypoxic Area

Time series of simulated bottom water oxygen in comparison to observed values in our six analysis regions are shown in Figure 3 for 2005 and 2006. (Note: Since the small and large-domain ROMS models yield very similar results with respect to bottom water oxygen concentrations and hypoxic area, we only show the large-domain results in the figures in this section, but report quantitative metrics for all four models in Table 1.) The simulated temporal evolution is coherent between all models and generally agrees with the observations. In both years, all models are overestimating bottom oxygen from March to August in the Mississippi region; this overestimation is more pronounced on the inner shelf. There is also a less pronounced overestimation of bottom oxygen in the inner Atchafalaya region in 2005. There are no noticeable biases in the outer Atchafalaya region in 2005 and the inner Atchafalaya region in 2006, while in the outer Atchafalaya region in 2006 all models underestimate bottom oxygen in June and July. In the Far Field regions, all models agree well with the available observations, which are more limited. The models are more coherent on the inner shelf (inside the 20 m isobath); model-to-model differences are more noticeable on the outer shelf. The most obvious model-to-model differences are deviations between NCOM and the other two models in the midshelf regions in late summer of 2006.

Point-to-point comparisons of observed bottom water oxygen concentrations with the corresponding simulated values are shown in Figure 4 and statistics reported in Table 1. The correlation between observed and simulated oxygen concentrations is similar in all three models ranging from 0.27 to 0.37. Root-mean-square errors (RMSEs) are also similar ranging from 55 to 67 mmol O_2 m^{-3}, but the models differ in terms of their biases (calculated as model minus observation). FVCOM and the small ROMS with satellite-based k_d have the smallest biases of 1 and 2 mmol O_2 m^{-3}, respectively. NCOM and the large ROMS have the largest biases with 19 and 23 mmol O_2 m^{-3}, and the initial small ROMS is in-between with 12 mmol O_2 m^{-3}.

The temporal evolution of simulated hypoxic area for NCOM, FVCOM, and the initial small ROMS is shown in Figures 5a and 5b for 2005 and 2006. There are marked differences between all three models with deviations frequently exceeding 5000 km^2. The observation-based estimate of hypoxic area from an annually recurring midsummer mapping cruise [*Rabalais et al*., 2002; *Obenour et al*., 2013] is shown for reference. None of the models match the observed extent in late July of 2005 although all models exceed the simulated extent in August. In 2006 all models reproduce the observed extent at the right time. Obviously one

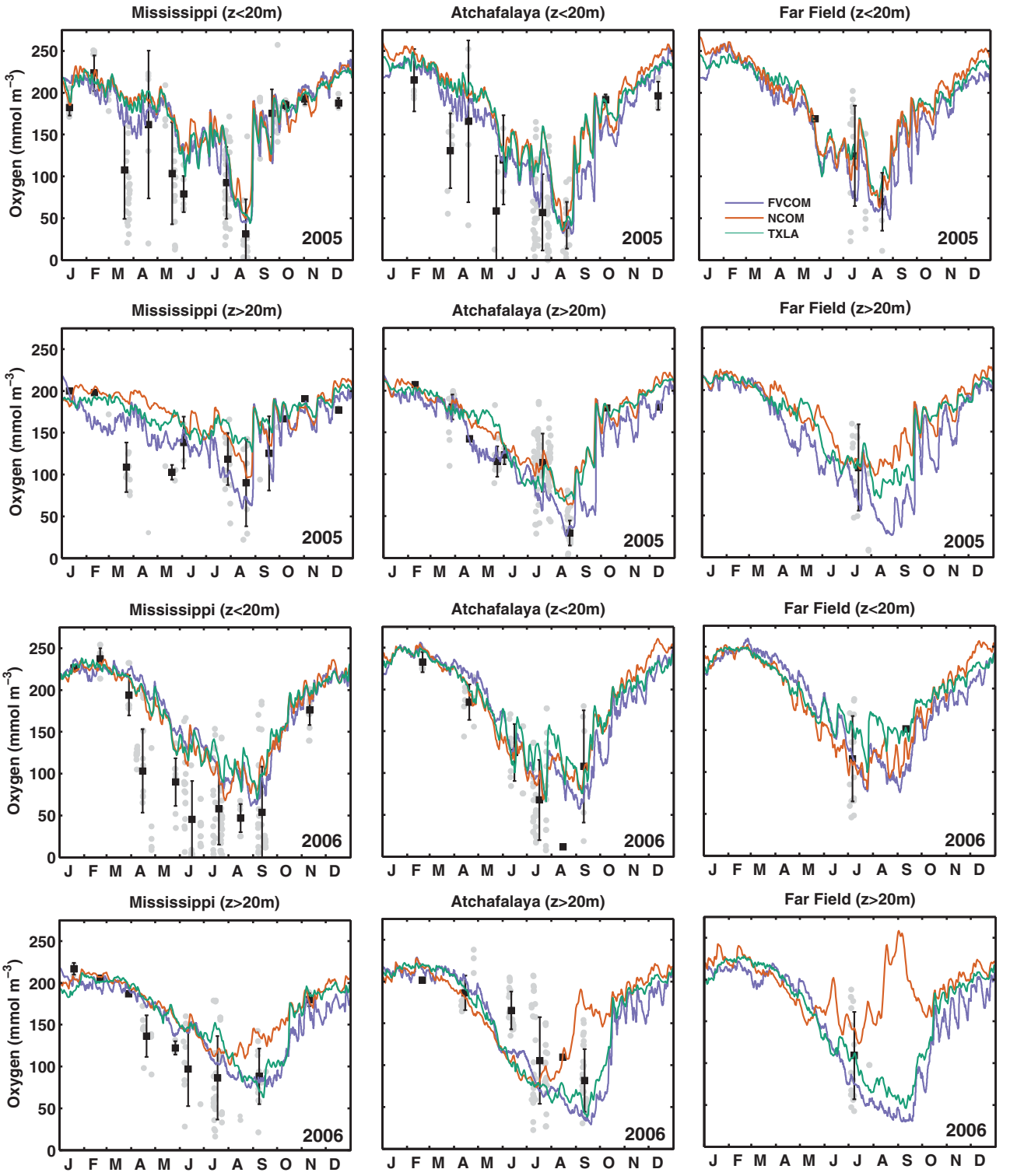

Figure 3. Simulated and observed oxygen concentration in bottom waters in the six analysis regions defined in Figure 2. The top six plots are for 2005, the bottom six for 2006. Evolution of average simulated oxygen is shown by the colored lines. Individual oxygen observations are shown by gray dots and their monthly means by black squares with error bars indicating one standard deviation.

data point per year is not sufficient to verify the simulated temporal evolution of hypoxic area. Hence, we focus our description and later discussion of hypoxic area magnitude and evolution on model-to-model differences.

Table 1. Metrics of Hypoxic Extent and Model-Data Differences

	Large ROMS	Small ROMS	Small ROMS w/Satellite k_d	FVCOM	NCOM
Time-integrated hypoxic area (HA) and hypoxic volume (HV)					
HA ($\times 10^3$ km^2 yr)					
Simple oxygen model 2005	1176	843	1555	1243	682
Simple oxygen model 2006	1625	925	1666	1737	925
Diagnostic run 2005	1184	918	2027	1745	1315
Diagnostic run 2006	1757	1104	1984	1580	1610
HV ($\times 10^{10}$ m^3 yr)					
Simple oxygen model 2005	586	378	603	409	102
Simple oxygen model 2006	949	463	754	606	118
Diagnostic run 2005	625	454	901	599	273
Diagnostic run 2006	1012	594	988	560	339
Point-by-point model-data comparison measures for bottom water properties					
Temperature					
R^2	0.88	0.88	0.88	0.86	0.92
RMSE (°C)	3.14	2.67	1.33	1.73	1.61
Bias (°C)	2.60	2.09	0.14	0.37	−0.68
Salinity					
R^2	0.56	0.56	0.55	0.59	0.66
RMSE	2.03	1.75	1.78	1.88	1.43
Bias	−1.21	−0.62	−0.66	−1.07	−0.15
Oxygen					
R^2	0.37	0.31	0.27	0.37	0.30
RMSE (mmol O_2 m^{-3})	58.51	64.07	66.61	55.49	61.85
Bias (mmol O_2 m^{-3})	11.58	22.68	2.02	0.97	19.45

In 2005 from mid-July to mid-September all models simulate a very similar magnitude and temporal evolution of hypoxic area, while in June and early July ROMS simulates a much larger hypoxic area than the other two, and in late September FVCOM simulates a much larger hypoxic area than the other two (Figure 5a). In 2006 all models simulate a relatively similar hypoxic area in July; however, in August, September, and early October NCOM differs markedly from the other two in its simulated hypoxic area, which is much smaller (Figure 5b). Overall, NCOM tends to simulate smaller hypoxic areas than the other two. This can also be seen in the time-integrated hypoxic area values (Table 1), which are similar between ROMS and FVCOM in both years, but smaller in NCOM.

We deem the simple oxygen models to be representative enough of the observed oxygen dynamics in our study system for our model intercomparison to be meaningful. While correlation coefficients and RMSEs between simulated and observed bottom water oxygen concentrations are similar between all models, they have different biases indicating that systematic model-to-model differences exist. The presence of systematic model-to-model differences is confirmed by the discrepancies in simulated hypoxic area in 2005 and 2006, which we will now analyze further.

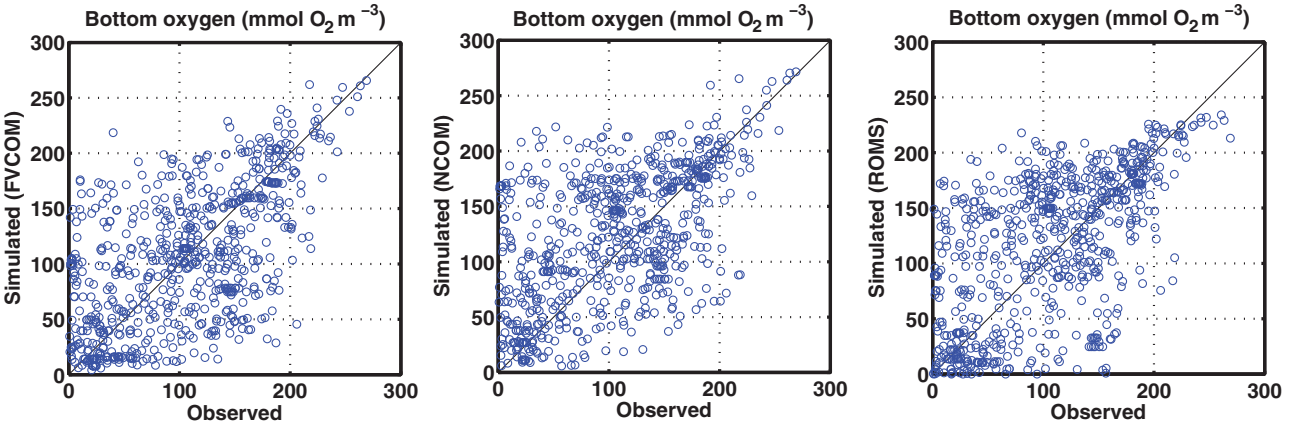

Figure 4. Point-by-point comparison of simulated and observed bottom water oxygen using all available observations in 2005 and 2006. FVCOM, NCOM, and large-domain ROMS are shown from left to right. Correlation coefficients, RMSE, and bias are given in Table 1.

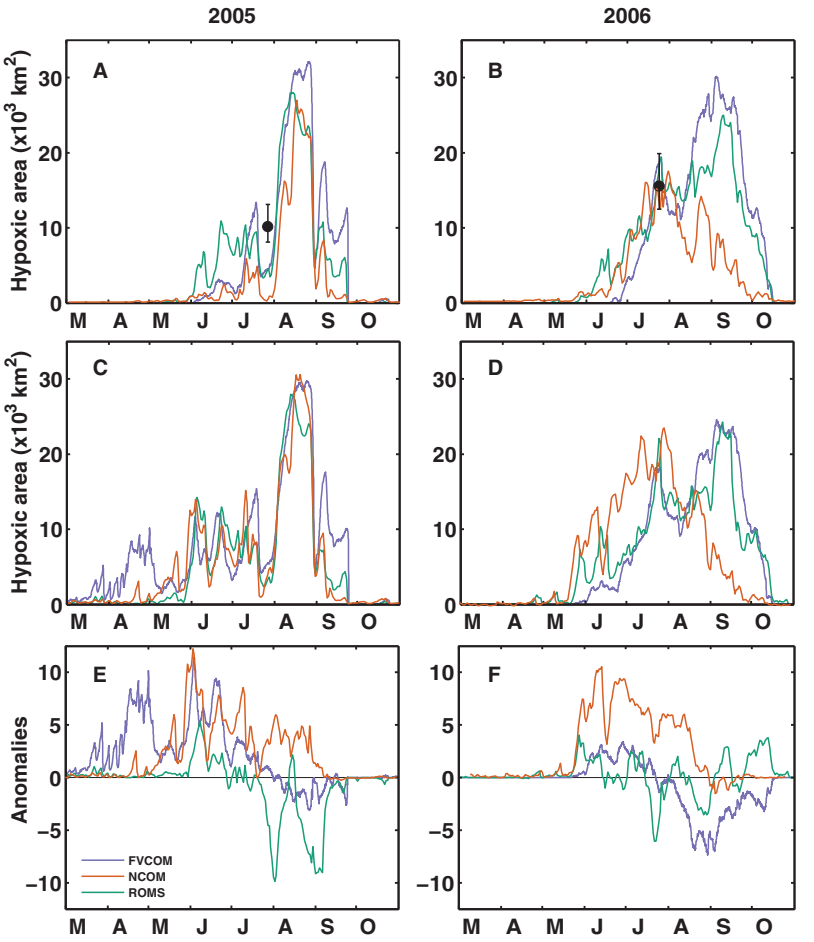

Figure 5. (a and b) Simulated hypoxic extent within the combined analysis region (defined in Figure 2) for 2005 and 2006. Observed hypoxic extent and standard deviation are shown by the black dot with error bar. (c and d) Simulated hypoxic extent from diagnostic runs where the temperature dependence of sediment oxygen consumption (SOC) was removed by imposing a constant bottom water temperature of 28°C within the SOC parameterization. (e and f) Anomalies in hypoxic extent calculated as diagnostic run (c and d) minus simple model (a and b).

3.2. Diagnostic Runs

The generation of hypoxia in these simple oxygen model simulations is principally determined by the imbalance between oxygen supply and oxygen consumption. Oxygen supply is controlled by physical factors like water column stratification, the vertical mixing regime and to a lesser extent advective processes. Net respiratory oxygen consumption in the water column and at the sediment-water interface is directly prescribed. Water column oxygen consumption is identical in all models, thus differences in hypoxic extent can only arise from differences in physical oxygen supply, and differences in the magnitude of sediment oxygen consumption (SOC), which is dependent on bottom water temperature. In order to distinguish the effects of bottom water temperature (affecting a key oxygen sink) from those of oxygen supply, we conducted diagnostic simulations where the temperature dependence of SOC was removed by imposing a constant 28°C in the SOC parameterization (Figures 5c and 5d). This means that oxygen sinks are identical between all diagnostic simulations; differences in hypoxic extent can only arise from differences in oxygen supply. Whenever a diagnostic run differs from its corresponding simple oxygen model run, this must be because bottom water temperature diverges sufficiently from the imposed 28°C in the diagnostic run to affect the simulated hypoxic area.

In 2005 all three diagnostic runs produce a relatively similar hypoxic area from June to August (Figure 5c) indicating that the models' constraints on oxygen supply are similar during this time. This is in contrast to the results from the simple oxygen model runs (Figure 5a) where NCOM tends to produce a smaller area than the other two models. The smaller hypoxic area in NCOM must be due to its colder bottom water

temperatures. For September 2005, the diagnostic FVCOM simulates a larger hypoxic area than the other two models (similar to the results from the simple model runs), and in April 2005 the diagnostic FVCOM produces a large hypoxic area (>5000 km^2), while neither the other two diagnostic models nor any of the simple oxygen model simulations do. This suggests that oxygen supply to bottom waters is smaller in FVCOM than in the other two models in April and September.

In 2006 the differences between simulated hypoxic area in the diagnostic runs versus the simple oxygen model runs are relatively small for ROMS and FVCOM, with a small increase in the former and a small decrease in the latter, while the hypoxic area in the diagnostic NCOM increases notably in June and July (Figure 5d). In June and July, the diagnostic NCOM simulates a much larger hypoxic area than the other two models but produces a much smaller hypoxic area in August and September. It appears that constraints on oxygen supply are different between NCOM and the other two models in 2006. It is also worthwhile pointing out that from November to March no hypoxia forms in the diagnostic runs (with the exception of FVCOM in March 2005); hence oxygen supply must be large enough during that time to counteract the relatively large oxygen sink via SOC that is imposed in the diagnostic runs.

The time-integrated hypoxic area values (Table 1) show that in the initial ROMS simulations the overall hypoxic area changes little (<10%) between the simple model and the diagnostic run, while in FVCOM the hypoxic area increases in the diagnostic run in 2005 (by 40%) and decreases in 2006 (by 9%). In NCOM the hypoxic area in the diagnostic increases by 93% in 2005 and 74% in 2006. In the ROMS simulation with satellite-based k_d, the hypoxic area increases by 30% in 2005 and 19% in 2006 in the diagnostic run.

The differences or anomalies between diagnostic and simple model runs are shown in Figures 5e and 5f and illustrate how deviations in bottom water temperature from 28°C affect hypoxic extent. For NCOM the anomalies are almost always positive indicating that, on average, bottom water temperatures are lower than 28°C and that a temperature increase would significantly increase hypoxic extent in June, July, and August of both years. For FVCOM the anomalies are positive from March to July 2005 and in June and July 2006, but turn negative in both years at the beginning of August; especially in August and September of 2006 there are large negative anomalies (>5000 km^2) in FVCOM indicating that the average bottom water temperatures are above 28°C. ROMS has large negative anomalies in August and September of 2005 indicating average bottom water temperatures are above 28°C, and less pronounced anomalies in 2006. Whenever the anomalies between models differ significantly, model-to-model differences in bottom water temperatures contribute to the differences in simulated hypoxic area between models.

The results of the diagnostic simulations show that hypoxic area is sensitive to bottom water temperature, which we will investigate further in the next subsection, and that there are marked model-to-model differences in oxygen supply or availability, which we will analyze in subsections 3.4 and 3.5.

3.3. Bottom Water Temperature and SOC

Results reported in the previous section indicate that differences in bottom water temperatures, through their effect on SOC, can explain a significant fraction of the model-to-model differences in the estimates of hypoxic area. For example, we can infer that in June, July, and August of 2005 the smaller hypoxic area simulated by NCOM is largely due to its colder bottom water temperatures, while in August and September of 2006 FVCOM simulates the largest hypoxic area because its bottom water temperatures are warmer than those of the other two models.

Time series of average bottom water temperatures from all three models (Figure 6) confirm these inferences. In both years there is an almost constant offset between NCOM and ROMS, the latter being about 2 to 3°C warmer. In 2005 and during the first 5 months of 2006, FVCOM's bottom water temperatures are relatively similar to NCOM, but at the beginning of June FVCOM starts to warm notably and exceeds ROMS temperatures in August, September, and October 2006.

Figure 7 shows how bottom water temperatures affect SOC in the simple oxygen model. In 2005 the warmer bottom waters in ROMS lead to larger SOC (by at least 5 mmol O_2 m^{-2} d^{-1} in the oxygen range between 50 and 100 mmol O_2 m^{-3}) than in the other two models, which have a very similar SOC. This is in contrast to 2006 when FVCOM's SOC is larger, because of the above mentioned warming of its bottom waters in the second half of the year, and very similar to ROMS.

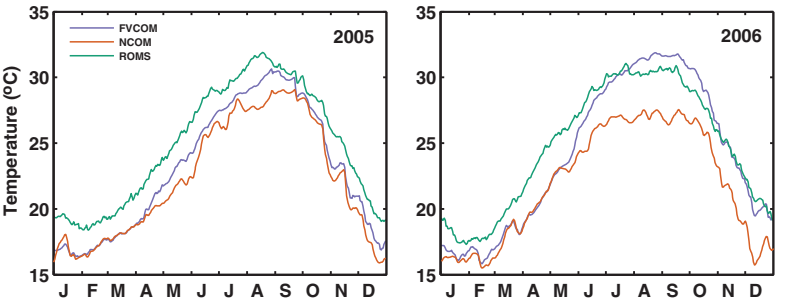

Figure 6. Simulated average bottom water temperature within the combined analysis region (defined in Figure 2) in 2005 and 2006.

Point-to-point comparisons of simulated and observed bottom water salinities and temperatures are given in Figure 8 and statistics in Table 1. For salinity ROMS and FVCOM have similar correlation coefficients (0.56 and 0.59, respectively) and RMSEs (between 1.8 and 2). Both models have a negative (fresh) bias ranging from 0.7 to 1.2 in magnitude. NCOM has a smaller salinity bias (−0.15), a slightly larger correlation coefficient (0.66), and its RMSE (1.4) is similar to that of the other two models. For temperature all models have a similar correlation coefficient (∼0.9). The initial ROMS simulations have the largest RMSE (3.1°C) while FVCOM, NCOM, and ROMS with satellite-derived k_d are similar with RMSEs of 1.7°C, 1.6°C, and 1.3°C. The models differ in their temperature biases. The initial ROMS simulations have the largest warm bias of ∼2.6°C followed by FVCOM with a small warm bias of 0.4°C and NCOM with a cold bias of −0.7°C. In the ROMS simulation with satellite-derived k_d the bias is smallest with 0.14°C.

Bottom water temperature, through its influence on SOC, is a key contributor to model-to-model differences in hypoxic area. While correlations and RMSEs between observed and simulated bottom water temperatures are not effective indicators of model-to-model differences in our case, biases are useful in indicating systematic differences between models.

3.4. Stratification Strength

As described in section 3.2, the results of the diagnostic runs show that model-to-model differences in oxygen supply can produce notably different estimates of hypoxic area. For example, in September 2005 the diagnostic FVCOM has a much larger hypoxic area than the other two models implying a smaller oxygen supply in FVCOM, and from mid-August to mid-October 2006 the diagnostic NCOM has a much smaller hypoxic area than the other two models indicating more effective oxygen supply in NCOM. An important determinant of oxygen supply is likely overall stratification strength in the models.

In June and July 2005 overall stratification strength (shown as histograms of the squared Brunt-Väisälä frequency, N^2, by the shaded areas in Figure 9) is similar in all models in that the modes of the distributions coincide, although stratification in ROMS is slightly stronger and in FVCOM slightly weaker than the other two models. In August 2005, ROMS is more strongly stratified than the other two models, which have very similar N^2 distributions. In September 2005, NCOM is most weakly stratified while the other two models are

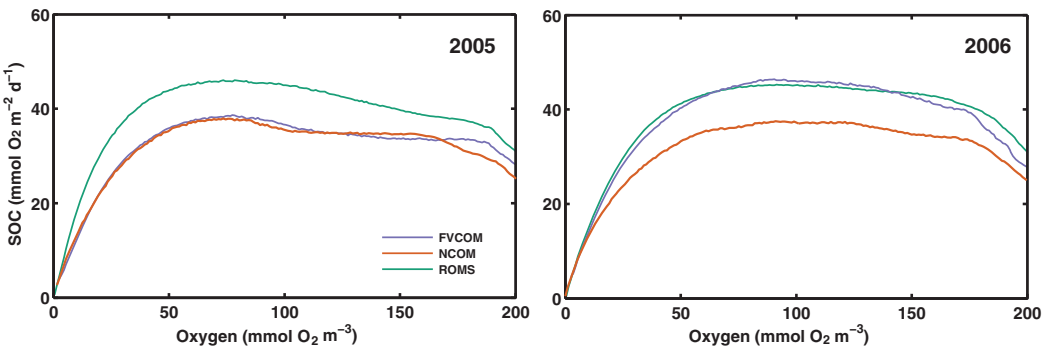

Figure 7. Simulated average sediment oxygen consumption (SOC) over bottom water oxygen concentration within the combined analysis region (defined in Figure 2) in 2005 and 2006.

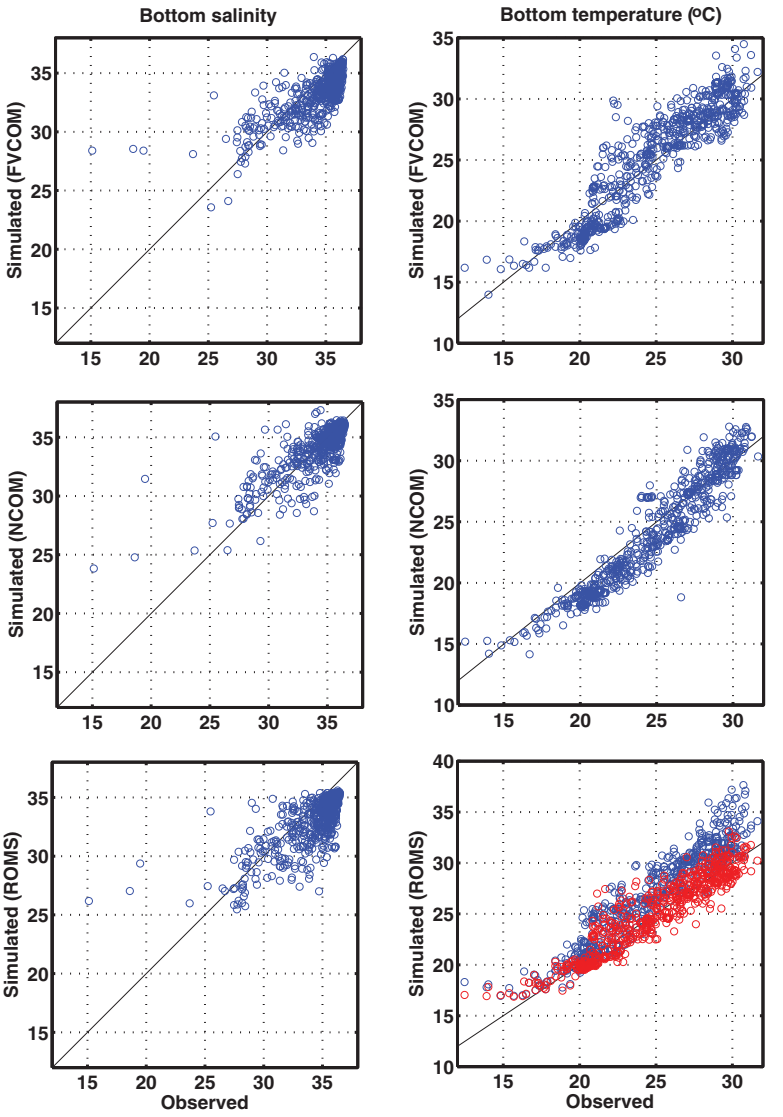

Figure 8. Point-by-point comparison of simulated and observed bottom water temperature and salinity using all available observations in 2005 and 2006. FVCOM, NCOM, and the large-domain ROMS are shown from top to bottom by blue circles. The bottom-right plot also shows the small ROMS with satellite-based k_d (red circles). Correlation coefficients, RMSE, and bias are given in Table 1.

similar. From June to September 2006, FVCOM is consistently the most weakly stratified model. NCOM and ROMS are relatively similar in June, July, and August, although the distribution in ROMS is slightly skewed toward stronger stratification. In September 2006, ROMS is more strongly stratified than NCOM.

These stratification differences alone do not explain the model-to-model differences in simulated hypoxic extent in the diagnostic runs; in fact, in some months differences in stratification strength are at odds with the results of the diagnostic simulations. One indication comes from N^2 histograms that only include those grid cells with hypoxic bottom waters in the diagnostic runs (Figure 9, solid lines). In June 2005, when overall stratification is similar between all models, the mode of the distribution for hypoxic grid cells in FVCOM is noticeably smaller than in the other two models illustrating that stronger stratification is needed in ROMS and NCOM than in FVCOM for hypoxic bottom waters to develop. In August 2005, when all models have similar hypoxic areas in the diagnostic runs, ROMS is more strongly stratified than the other two models. In September 2006, when FVCOM and ROMS have similar hypoxic areas in the diagnostic runs and NCOM a much smaller area, FVCOM has the weakest stratification. It appears that, given equal stratification, FVCOM is more prone to developing hypoxia than the other two models, while ROMS appears to require stronger stratification than the other two in order to generate hypoxia.

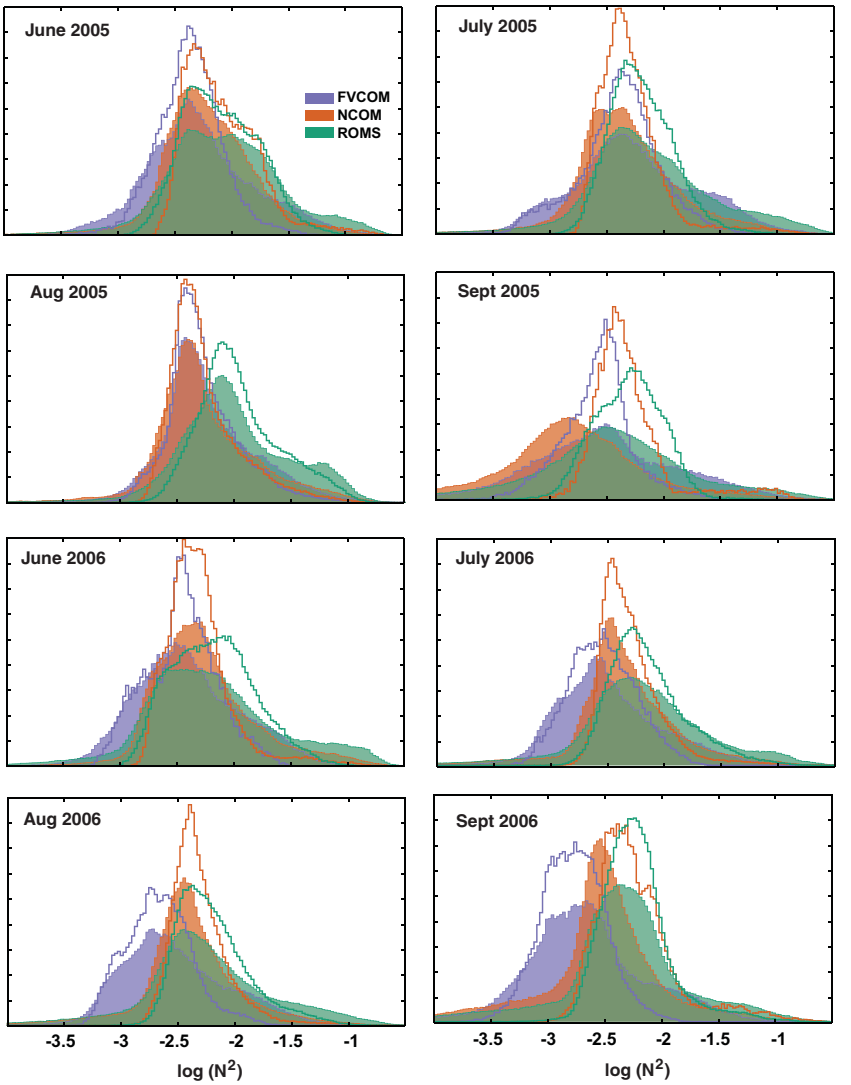

Figure 9. Histograms of squared Brunt-Väisälä frequency given as $\log(N^2)$ are shown for all grid cells in the combined analysis region (defined in Figure 2) as transparent-filled colors. Histograms of $\log(N^2)$ for those grid cells with hypoxic bottom waters in the diagnostic run are given by the colored solid lines.

3.5. Hypoxic Layer Thickness and Hypoxic Volume

Results in subsection 3.4 show that stratification strength is, somewhat surprisingly, a poor descriptor of model-to-model differences in hypoxic conditions. Oxygen supply/availability in bottom waters, which is the only factor in which the diagnostic models differ from each other, must thus be dependent on another property. In order to elucidate this further, we now examine the vertical structure of hypoxic waters.

In Figure 10 the thickness of the hypoxic layer is shown for the diagnostic models. The hypoxic layer is thinnest in FVCOM where hypoxia is typically constrained to within the bottommost two meters. In NCOM the hypoxic layer is slightly thicker but typically restricted to within 3 m of the bottom. The initial ROMS has the thickest hypoxic layers (the mode of the distribution at between 3 and 4 m above the bottom). Also shown in Figure 10 are the thicknesses of the bottom boundary layer, which closely match those of the hypoxic layer in all three models. This indicates that the thickness of the bottom boundary layer essentially sets the thickness of the hypoxic layer. The model-to-model differences in bottom boundary layer thickness determine how quickly hypoxic conditions can be generated because SOC is an important oxygen sink in this system. Given the same initial oxygen concentration, the timescale for oxygen drawdown to hypoxic concentrations is longer for a thicker bottom boundary layer (because a larger volume and thus a larger oxygen reservoir has to be depleted). This explains why, on the one hand, the hypoxic area in the diagnostic FVCOM

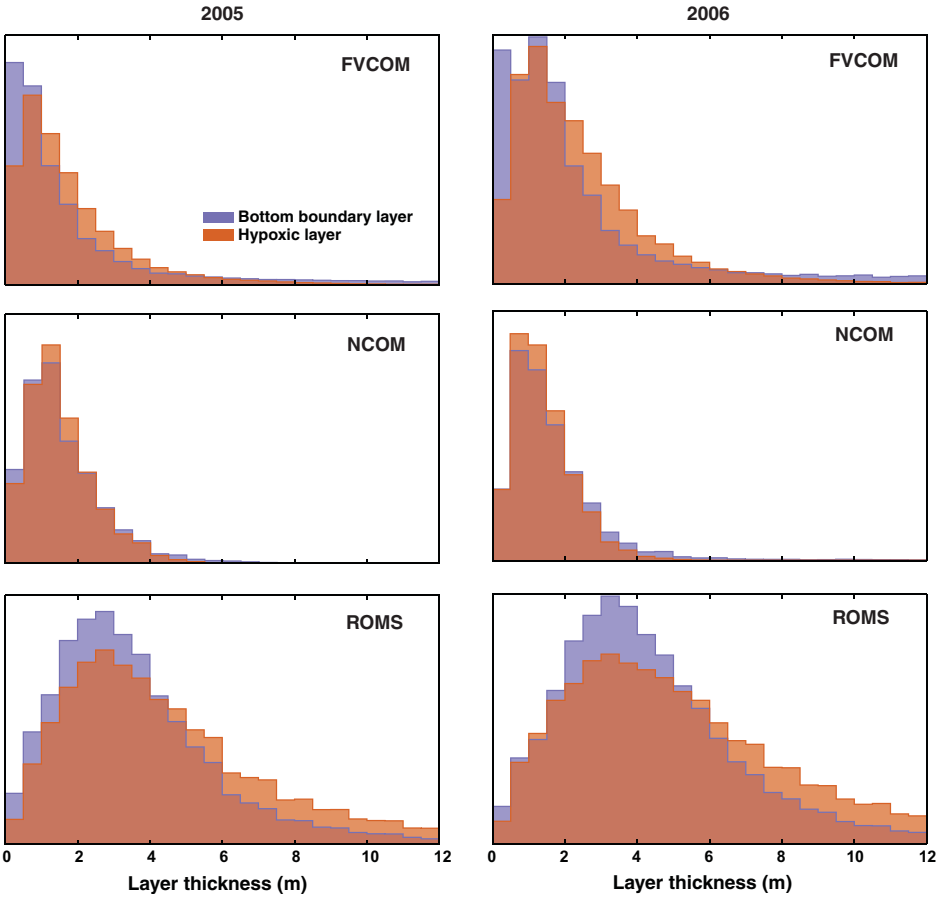

Figure 10. Histograms of the thickness of the hypoxic layer (orange) and the bottom boundary layer thickness (purple) for June, July, August, and September combined. The top of the hypoxic layer is defined as the layer where dissolved oxygen concentration is <63 mmol m^{-3}. The top of the bottom boundary layer is defined here as the maximum depth for which the vertical density gradient $d\rho/dz \geq 0.1$ kg m^{-4}.

is often similar to one or both of the other models despite weaker overall stratification. On the other hand, in the initial ROMS simulations, which have the thickest bottom boundary layers, stronger stratification is required to generate the same hypoxic area as one of the other models.

A comparison of simulated and observed thicknesses of the bottom boundary layer is given in Figure 11a using all available CTD profiles in the combined analysis region in 2005 and 2006. It is important to note that the observed thicknesses (shown by the black line) are underestimates of the true thicknesses because measured profiles never reach all the way to the bottom of the water column. Depending on CTD operator and sea state, profiles typically extend to only within 0.5–3 m above the bottom. Taking this into account it appears that FVCOM's and NCOM's bottom boundary layer thicknesses agree reasonably well with the observations, while bottom boundary layers in the initial ROMS simulation are too thick by about 2–3 m. We identified the treatment of vertical heat penetration in the initial ROMS simulations as the primary reason for their thicker and warmer bottom boundary layers. Due to the assumed clarity of water in these simulations heat penetrates deeper, overestimates warming in bottom waters and creates a thicker bottom boundary layer. By using the more realistic satellite-derived values of k_d, the temperature bias in ROMS is essentially removed (as stated above in section 3.3) and the bottom boundary layer thickness is in better agreement with the observations (Figure 11b). The effect of the more realistic bottom boundary layer thickness on simulated hypoxia in ROMS is illustrated in the supporting information Figures S1 and S2.

The model-to-model differences in the thickness of the hypoxic layer suggest that hypoxic area is an incomplete metric for model-to-data and model-to-model comparisons and that hypoxic volume should be considered as well. The evolution of hypoxic volume for the simple oxygen models is shown in Figure 12 for

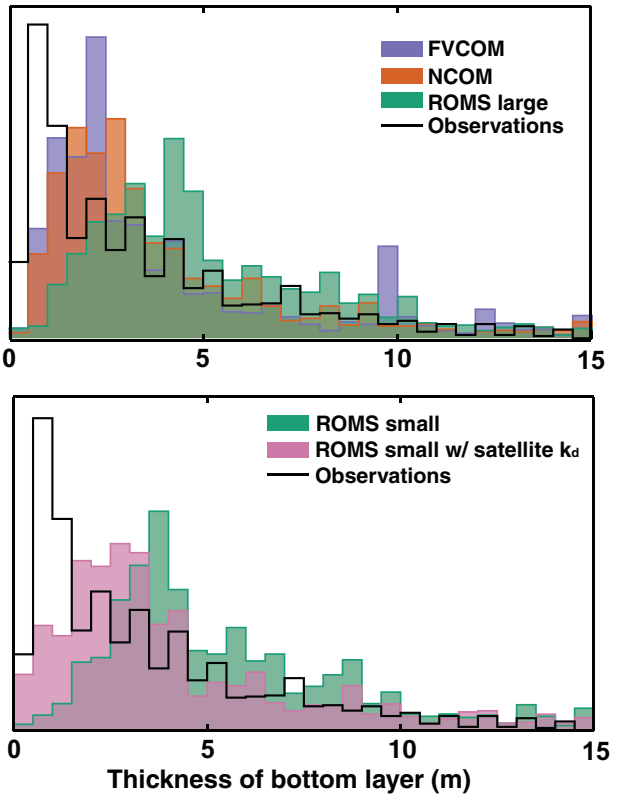

Figure 11. Histogram of observed thicknesses of the bottom boundary layer using all available observations for 2005 and 2006 (black line) and histograms of the corresponding simulated thicknesses (shaded areas). See caption of Figure 10 for the bottom boundary layer criterion. (a) Initial large-domain ROMS, FVCOM, NCOM, and observations. (b) Initial small-domain ROMS, small-domain ROMS with satellite-based k_d and observations.

2005 and 2006. ROMS, the model with the thickest bottom boundary layer and strongest stratification, consistently simulates the largest hypoxic volume (except for September 2005 when FVCOM's is slightly larger). This is in contrast to the hypoxic area estimates where FVCOM consistently has the largest or one of the largest hypoxic extents in July, August, and September (Figures 5a and 5b).

4. Discussion

We implemented a simple oxygen parameterization in three different circulation models for the continental shelf in the northern Gulf of Mexico in order to assess and compare the models' ability to simulate recurring summer hypoxia in the region. The simple oxygen parameterization allows us to focus on the effects that model-to-model differences in model physics (including differences in temperature, salinity, vertical stratification, mixing and boundary layer structure) have on hypoxia generation. The oxygen parameterization is simple in that it only includes air-sea gas exchange, a vertically uniform term for net respiration in the water column and a term for sediment oxygen consumption (SOC) [*Yu et al*., 2015b]. The net water column respiration is constant in time and in the vertical direction, but varies horizontally with bathymetry. The SOC term is dependent on bottom water oxygen concentration and bottom water temperature.

Despite its simplicity the simple oxygen model performs well. *Yu et al*. [2015b] demonstrated that the simple model, when coupled with a ROMS circulation model, produces oxygen distributions and hypoxic conditions that are in close agreement with those from a full biogeochemical model and also agree with observations. Here we show that the simple model also works well in two other circulation models: FVCOM and NCOM. A comparison of simulated and observed bottom water oxygen concentrations across six analysis regions on the shelf shows that the simulated evolution of oxygen is coherent in all models and agrees well with many of the average observed values, although the models do not accurately reproduce observed

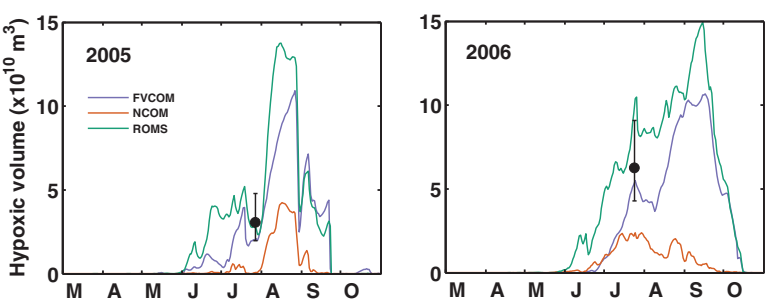

Figure 12. Simulated hypoxic volume within the combined analysis region (defined in Figure 2) for 2005 and 2006. Observed hypoxic volume and standard deviation are shown by the black dot with error bar.

oxygen in all regions. For example, the models overestimate observed oxygen in the Mississippi regions, which one might expect given that the same constant water column respiration term is used within bathymetric depth ranges regardless of distance to river mouth, and that the SOC term also does not take into account the distance from the river sources. Also, in reality oxygen sinks vary interannually because annual nutrient loads vary, but the oxygen consumption terms do not take these variations into account. This may explain why, for example, in the outer Atchafalaya region the models accurately reflect observed oxygen levels in June of 2005, but underestimate them in June of 2006.

Other investigators had similar successes in simulating dissolved oxygen and hypoxia in Chesapeake Bay with highly simplified oxygen models. *Scully* [2010, 2013] presented a realistic circulation model of Chesapeake Bay coupled to a highly simplified oxygen model that only includes a constant water column respiration term and a parameterization of air-sea gas exchange. The model was shown to reproduce the overall distribution of dissolved oxygen consistent with observed typical summertime conditions [*Scully*, 2010] and reproduces bottom water oxygen concentrations observed at monitoring stations with reasonable skill [*Scully*, 2013]. *Li et al*. [2015] formulated a slightly more complex oxygen model for coupling with their realistic circulation model for Chesapeake Bay; it includes air-sea gas exchange, photosynthetic oxygen production, respiration in the water column, and sediment oxygen consumption. Their model also reproduces the observed oxygen dynamics skillfully. These models were used to investigate the role of physical processes in hypoxia generation in Chesapeake Bay including wind-driven ventilation [*Scully*, 2010], and the influences of wind direction and speed, river discharge, and water temperature on oxygen dynamics [*Scully*, 2013; *Li et al*., 2015].

We calculated correlation coefficients and RMSEs as statistical comparison measures between observations and corresponding model values for bottom water temperature, salinity, and oxygen. Both measures are similar for the three models and thus not useful in discriminating model skill between models, even though there are large differences between the models' simulated hypoxic extent. These differences are not reflected in the correlation coefficients or RMSEs. The similarity in correlation coefficients and RMSEs, which are point-by-point comparison metrics, is likely due to the low signal-to-noise ratio in the system. During the hypoxic season, the northern Gulf of Mexico shelf is characterized by pronounced small-scale variability in surface salinity distributions and in bottom water oxygen. *Marta-Almeida et al*. [2013] demonstrated how this variability, which largely results from instabilities along many dynamic plume fronts, leads to a low signal-to-noise ratio in surface salinity. Since surface salinity strongly affects stratification strength at any given location, the uncertainty in this physical surface property likely also leads to uncertainty in bottom water oxygen concentrations and thus hypoxia. Indeed, *Mattern et al*. [2013] showed that this is the case by quantifying the uncertainty in hypoxia predictions that results from uncertainty in various input parameters. The fact that in this system the signal-to-noise ratio is large in summer for physical properties and bottom oxygen concentrations limits the usefulness of point-to-point comparison metrics like the correlation coefficient and the RMSE for the evaluation of model-data differences as well as model-to-model comparisons. In order for circulation models to match salinity distributions better in a point-by-point sense one would need a data-assimilative model that ingests a large number of detailed observations; it is doubtful that such a model system is attainable for the study region in the near term.

We also calculated biases between models and observations, which are a cumulative measure and less dependent on point-to-point agreement. Biases for bottom water temperature, salinity, and oxygen concentration were different between models and useful in assessing their skill and comparing them. All models have a fresh bias in bottom water salinity, which is largest in ROMS and smallest in NCOM. In terms of bottom water temperature, NCOM has a cold bias, FVCOM is slightly too warm and ROMS has a significant warm bias of 2.6°C if clear water is assumed for vertical heat penetration. The warm bias in ROMS bottom waters had not been noticed in previous validation studies [e.g., *Marta-Almeida et al*., 2013]. Here we showed that using the more realistic attenuation coefficients for shortwave radiation based on satellite observations [*Schaeffer et al*., 2011] that are used in NCOM corrects the temperature bias in ROMS (to 0.14°C). With regard to the relatively small bias in FVCOM, it should be noted that FVCOM is significantly warmer in 2006 than in 2005 and positively biased in 2006. Biases in bottom oxygen are positive in all models (they all overestimate oxygen). The oxygen bias is small in FVCOM and ROMS with satellite-based k_d, largest in NCOM and between the other two models for the initial ROMS simulations. The overestimation of oxygen by NCOM is consistent with NCOM underestimating the observed hypoxic area in 2005, while the

difference in bias between FVCOM and initial ROMS is not reflected in their ability to match the observed hypoxic extent, which is similar.

We now discuss the reasons behind model-to-model differences in hypoxia simulations trying to answer the following two questions. When simulated hypoxic extent is different between models, what is the reason for this difference? When models simulate a similar hypoxic extent, do they do so for different reasons? NCOM tends to simulate a smaller hypoxic area than the other two models. We are interested in understanding why. ROMS and FVCOM tend to simulate similar hypoxic areas, but are they similar for different reasons? Given the simplicity of our oxygen parameterization there are principally only two processes due to which models can differ in their simulated oxygen: one is physical oxygen supply and the other is oxygen consumption in the sediment (SOC). The diagnostic runs we performed allow us to separate between these two possibilities.

Comparison between the simple oxygen model simulations and the corresponding diagnostic runs illustrates the effect of bottom water temperature. By forcing SOC in all models to experience the same bottom water temperature, the simulated hypoxic area can be affected significantly (by 40% in FVCOM and by almost a factor of 2 in NCOM in 2005). In 2005 the warmer average bottom water temperature in ROMS leads to an increase in SOC by about 12% compared the other two models. In 2006, SOC in ROMS and FVCOM is about 12% larger than in NCOM. Since hypoxia generation in this system is very sensitive to SOC [*Fennel et al*., 2013; *Yu et al*., 2015b], these relatively modest changes are enough to significantly affect the simulated hypoxic area. Bottom water temperature is known to be an important control on microbial respiration in coastal sediments [e.g., *Wilson et al*., 2013, their Figure 5]. In hypoxia models for this system that include an explicit temperature-dependence of SOC, bottom water temperature is an important physical property that needs to be validated.

Comparisons of the diagnostic runs tell us about model-to-model differences in oxygen supply, which should primarily be determined by a model's stratification/mixing regime and possibly by advective transport. The most notable differences in hypoxic area in the diagnostic runs are in September of 2005 and August to October of 2006. Model-to-model differences in vertical stratification strength seem to be an obvious factor to explain these differences. Within a given model, stratification strength is highly correlated with bottom water oxygen concentration, see, e.g., *Fennel et al*. [2013, their Figure 9] for an example from ROMS. However, overall stratification strength does not explain the apparent model-to-model differences here. For example, in September of 2006 the diagnostic ROMS and FVCOM simulate a large hypoxic area in excess of 20,000 km^2 while the hypoxic area in NCOM is only 5000 km^2. At the same time FVCOM is the most weakly stratified model of the three, while NCOM and ROMS are similar in overall stratification strength. In both years, FVCOM has weaker stratification than the other two models, while ROMS and NCOM are similar in terms of stratification strength. In other words, given equal stratification strength it appears that FVCOM is more likely to develop hypoxia than the other two models.

A key factor to consider in explaining model-to-model differences in hypoxic area in the diagnostic runs is the structure of the bottom boundary layer. It has been pointed out previously [e.g., *Fennel et al*., 2013, *Yu et al*., 2015a] that the stratification structure near the bottom of the water column matters for hypoxia generation in the northern Gulf of Mexico, where hypoxic conditions are typically constrained to the bottom boundary layer [*Wiseman et al*., 1997]. In all three models the bottom boundary layer apparently sets the thickness of the hypoxic layer. Bottom boundary layers and hypoxic layers are thinnest in FVCOM, and similar between NCOM and the initial ROMS simulations, although slightly thicker in the latter. In systems where SOC is an important oxygen sink, a larger reservoir of oxygen has to be drawn down for a thicker hypoxic layer. In a system where water column respiration dominates, the hypoxic layer thickness does not matter because the volume-integrated oxygen uptake scales with volume. FVCOM, with its thin bottom boundary layers, is prone to generating hypoxia quickly despite its overall weaker stratification. NCOM and the initial ROMS, with their thicker boundary layers, are less prone to developing hypoxia. The initial ROMS accomplishes a similar hypoxic area as FVCOM despite its thick bottom boundary layer by having stronger stratification and overly warm bottom waters, which increases SOC. In other words, the initial ROMS generates comparable hypoxic area estimates as FVCOM, but for different reasons. NCOM, with its relatively thick bottom boundary layers but colder bottom water temperatures and slightly weaker overall stratification than ROMS, cannot generate the same hypoxic extent as the other two models. In ROMS with satellite-based k_d, i.e., the simulation with realistic vertical heat penetration and unbiased bottom-water temperature, the

thickness of the bottom boundary layer agrees much better with observations (Figure 11b). In this case the simulated hypoxic area expands significantly compared to the initial ROMS case (see supporting information Figure S1).

At this point we would like to emphasize that the simple hypoxia model was first developed and tested within ROMS [*Yu et al*., 2015b] and thus may inadvertently correct for some shortcomings of the physical ROMS implementation, i.e., the thick bottom boundary layers. Indeed, as shown in *Yu et al*.'s [2015a] validation of the full biogeochemical model, the net water column respiration rates in ROMS agree well with the observed rates by *Murrell and Lehrter* [2011], but SOC is at the upper end of the range of available SOC observations [see *Yu et al*., 2015a, Figure 7]. Our simple oxygen model may thus overestimate SOC. Such an overestimation in combination with the thick bottom boundary layer and overly warm bottom waters in ROMS can lead to accurate hypoxic area estimates, but for the wrong reasons. In ROMS with satellite-based k_d, where the temperature bias is corrected and bottom boundary layer thickness is more realistic, simulated hypoxia expands.

Our results suggest the following steps in order to improve our hypoxia models (i.e., ensuring they reproduce hypoxia observations accurately for the right reasons) and toward a predictive capability. The first is to ensure that several physical model aspects, which we have shown here to be crucially important, are represented accurately. These are overall stratification strength, the thickness of the bottom boundary layer and bottom water temperature. Fortunately observations of the required physical variables are relatively abundant. For validation of the three specific properties we stress here, CTD profiles are sufficient, as long as there is a large number available and they extend close to the bottom of the water column. The next logical step would be to reparameterize the simple oxygen model, specifically the SOC component, which is at the upper range of available measurements and probably overestimates SOC at least slightly. Refinement of the water column net respiration term is also warranted taking into account the vertical structure of net respiration, which should be negative in the upper part of the water column due to photosynthesis. With respect to hypoxia, our results point to the importance of using hypoxic volume as an evaluation metric in addition to hypoxic area. We would like to emphasize one other obvious limitation, namely the relative scarcity of shelf-wide surveys of dissolved oxygen. Availability of more frequent oxygen surveys like, e.g., the monthly sampling of the Chesapeake Bay monitoring program, would allow for a more rigorous assessment of model skill.

At this point in our study we cannot say with confidence whether there are any systematic differences in the models' ability to simulate hypoxia that result from model architecture (i.e., grid construction or numerical schemes like finite volume versus finite difference). Differences are more likely due to the many choices that have to be made when setting up a model ranging from atmospheric forcing and lateral boundary conditions to parameters for turbulent closure parameterizations and depth of solar heat penetration.

5. Conclusions

In order to elucidate how physical processes influence hypoxia generation in the northern Gulf of Mexico, and to assess which aspects of physical model dynamics are most important for skillful hypoxia prediction, we compared three circulation models that were coupled with the same simple oxygen model. This approach allows us to investigate the effects of model physics on dissolved oxygen concentrations without the potentially confounding effects of a full biogeochemical model, which will be the subject of a forthcoming study.

The circulation models differ in their underlying numerical schemes as well as in the details of their initial, boundary and forcing conditions and physical model parameters, thus providing us with three different realizations of the time-evolving physical state. The three coupled models simulate the observed bottom water oxygen concentrations reasonably well, but produce notable discrepancies in their simulated hypoxic area. Of the three statistical measures of model-data agreement that we analyzed, correlation and RMSE are not useful in discriminating between models, while biases contain valuable information. In analyzing the discrepancies in simulated hypoxic area between models, we separated the physical model states' influence on sediment oxygen consumption (SOC), an important oxygen sink, from physical supply processes representing the dominant oxygen source. We found simulated hypoxic area to be sensitive to bottom water temperature through its influence on SOC. Unexpectedly, stratification strength does not explain model-to-

model differences in hypoxic area; however, the thickness of the bottom boundary layer (BBL) is a key factor. We found the hypoxic layer to be constrained to the BBL in all three models. The thickness of the BBL, and thus the hypoxic layer, varies between the three models. Given identical SOC, the time to reach hypoxic bottom waters is longer for a thicker BBL, making models with thicker BBL less prone to developing hypoxia. It follows that hypoxic area is an incomplete metric for hypoxia quantification in this system, and that hypoxic volume needs to be considered as well. The aspects of physical model dynamics that matter most to accurate hypoxia prediction in the northern Gulf of Mexico are the bottom water temperature and the thickness of the BBL.

Acknowledgments
This work was supported by NOAA through the Coastal Ocean Modeling Testbed (COMT) project. All data used in this publication are available via the links provided in the methods section. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. NGOMEX publication number 210.

References

Aulenbach, B., H. Buxton, W. Battaglin, and R. Coupe (2007), Stream flow and nutrient fluxes of the Mississippi-Atchafalaya River Basin and subbasins for the period of record through 2005, *U.S. Geol. Surv. Open-File Rep. 2007-1080*.

Chen, C., R. C. Beardsley, and G. Cowles (2006), An unstructured grid, finite-volume coastal ocean model (FVCOM) system, *Oceanography*, *19*, 78–89.

da Silva, A., C. Young-Molling, and S. Levitus (1994a), *Atlas of Surface Marine Data 1994, Vol. 3, Anomalies of Fluxes of Heat and Momentum*, *NOAA Atlas NESDIS 8*.

da Silva, A., C. Young-Molling, and S. Levitus (1994b), *Atlas of Surface Marine Data 1994, Vol. 4, Anomalies of Fresh Water Fluxes, NOAA Atlas NESDIS 9*.

Diaz, R. J., and R. Rosenberg (2008), Spreading dead zones and consequences for marine ecosystems, *Science*, *321*, 926–929, doi:10.1126/science.1156401.

DiMarco, S. F., and R. O. Reid (1998), Characterization of the principal tidal current constituents on the Texas-Louisiana Shelf, *J. Geophys. Res.*, *103*, 3093–3110.

Feng, Y., S. F. DiMarco, and G. A. Jackson (2012), Relative role of wind forcing and riverine nutrient input on the extent of hypoxia in the northern Gulf of Mexico, *Geophys. Res. Lett.*, *39*, L09601, doi:10.1029/2012GL051192.

Feng, Y., K. Fennel, G. A. Jackson, S. F. DiMarco, and R. D. Hetland (2014), A model study of the response of hypoxia to upwelling-favorable wind on the northern Gulf of Mexico shelf, *J. Mar. Syst.*, *131*, 63–73.

Fennel, K., R. Hetland, Y. Feng, and S. DiMarco (2011), A coupled physical-biological model of the Northern Gulf of Mexico shelf: Model description, validation and analysis of phytoplankton variability, *Biogeosciences*, *8*, 1881–1899.

Fennel, K., J. Hu, A. Laurent, M. Marta-Almeida, and R. Hetland (2013), Sensitivity of hypoxia predictions for the Northern Gulf of Mexico to sediment oxygen consumption and model nesting, *J. Geophys. Res. Oceans*, *118*, 990–1002, doi:10.1002/jgrc.20077.

Forrest, D. R., R. D. Hetland, and S. F. DiMarco (2011), Multivariable statistical regression models of the areal extent of hypoxia over the Texas-Louisiana Shelf, *Environ. Res. Lett.*, *6*, 045002.

Greene, R. M., J. C. Lehrter, J. D. Hagy III (2009), Multiple regression models for hindcasting and forecasting midsummer hypoxia in the Gulf of Mexico, *Ecol. Appl.*, *19*, 1161–1175.

Haidvogel, D. B., et al. (2008) Regional Ocean Forecasting in Terrain-following Coordinates: Model Formulation and Skill Assessment, *J. Comp. Phys.*, *227*, 3595–3624, doi:10.1016/i.jcp.2007.06.016.

Hetland, R. D., and S. F. DiMarco (2008), How does the character of oxygen demand control the structure of hypoxia on the Texas- Louisiana continental shelf?, *J. Mar. Syst.*, *70*, 49–62.

Hodur, R. M. (1997), The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS), *Mon. Weather Rev.*, *135*, 1414–1430.

Hubbard, M. E. (1999), Multidimensional slope limiters for MUSCL type finite volume schemes on unstructured grids, *J. Comput. Phys.*, *155*, 54–74.

Hypoxia Task Force (2008), *Mississippi River/Gulf of Mexico Watershed Nutrient Task Force, 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, D. C.

Justić, D., and L. Wang (2014), Assessing temporal and spatial variability of hypoxia over the inner Louisiana–upper Texas shelf: Application of an unstructured-grid three-dimensional coupled hydrodynamic-water quality model, *Cont. Shelf Res.*, *72*, 163–179.

Ko, D. S., and D. P. Wang (2014), *Intra-Americas Sea Nowcast/Forecast System Ocean Reanalysis to Support Improvement of Oil-Spill Risk Analysis in the Gulf of Mexico by Multi-Model Approach*, *BOEM 2014-1003*, pp. 55, Dep. of the Inter., Bur. of Ocean Energy Manage., Herndon, Va. [Available at http://www.data.boem.gov/PI/PDFImages/ESPIS/5/5447.pdf.]

Ko, D. S., R. H. Preller, and P. J. Martin (2003), An experimental real-time Intra-Americas Sea Ocean Nowcast/Forecast System for coastal prediction, paper presented at the AMS 5th Conference on Coastal Atmospheric and Oceanic Prediction and Processes, Seattle, Washington, pp. 97–100, Am. Math. Soc., Boston, Mass.

Ko, D. S., P. J. Martin, C. D. Rowley, and R. H. Preller (2008), A real-time coastal ocean prediction experiment for MREA04, *J. Mar. Syst.*, *69*, 17–28.

Kobayashi, M. H., J. M.C. Pereira, J. C.F. Pereira, and J. C. F. Pereira (1999), A conservative finite-volume second-order-accurate projection method on hybrid unstructured grids, *J. Comput. Phys.*, *150*, 40–75.

Laurent, A., and K. Fennel (2014), Simulated reduction of hypoxia in the northern Gulf of Mexico due to phosphorus limitation, *Elementa*, *2*, 000022, doi:10.12952/journal.elementa.000022.

Laurent, A., K. Fennel, J. Hu, and R. Hetland (2012), Simulating the effects of phosphorus limitation in the Mississippi and Atchafalaya River plumes, *Biogeosciences*, *9*, 4707–4723.

Lehrter, J. C., M. C. Murrell, and J. C. Kurtz (2009), Interactions between Mississippi River inputs, light, and phytoplankton biomass and phytoplankton production on the Louisiana continental shelf, *Cont. Shelf Res.*, *29*, 1861–1872.

Lehrter, J. C., D. S. Ko, M. C. Murrell, J. D. Hagy, B. A. Schaeffer, R. M. Greene, R. W. Gould, and B. Penta (2013), Nutrient distributions, transports, and budgets on the inner margin of a river-dominated continental shelf, *J. Geophys. Res. Oceans*, *118*, 4822–4838, doi:10.1002/jgrc.20362.

Li, Y., M. Li, and W. M. Kemp (2015), A budget analysis of bottom-water dissolved oxygen in Chesapeake Bay, *Estuaries Coasts*, *38*, 2132–2148.

Luettich, R. A., Jr., et al. (2013), Introduction to special section on The U.S. IOOS Super-regional Coastal Ocean Modeling Testbed, *J. Geophys. Res. Oceans*, *118*, 6319–6328, doi:10.1002/2013JC008939.

Marta-Almeida, M., R. D. Hetland, and X. Zhang (2013), Evaluation of model nesting performance on the Texas-Louisiana continental shelf, *J. Geophys. Res. Oceans*, *118*, 2476–2491, doi:10.1002/jgrc.20163.
Martin, P. J. (2000), A Description of the Navy Coastal Ocean Model Version 1.0, *NRL Rep. NRL/FR/7322-00-9962*, 42 pp., Nav. Res. Lab., Stennis Space Cent., Miss.
Mattern, J. P., K. Fennel, and M. Dowd (2013), Sensitivity and uncertainty analysis of model hypoxia estimates for the Texas-Louisiana shelf, *J. Geophys. Res. Oceans*, *118*, 1316–1332, doi:10.1002/jgrc.20130.
Mellor, G. L., and T. Yamada (1974), A hierarchy of turbulence closure models for planetary boundary layers, *J. Atmos. Sci.*, *31*, 1791–1806.
Mellor, G. L., and T. Yamada (1982), Development of a turbulence closure model for geophysical fluid problems, *Rev. Geophys.*, *20*, 851–875.
Murrell, M. C., and J. C. Lehrter (2011), Sediment and lower water column oxygen consumption in the seasonally hypoxic region of the Louisiana continental shelf, *Estuaries Coasts*, *34*, 912–924.
Murrell, M. C., R. S. Stanley, and J. C. Lehrter (2013), Plankton community respiration, net ecosystem metabolism, and oxygen dynamics on the Louisiana continental shelf: Implications for hypoxia, *Cont. Shelf Res.*, *52*, 27–38.
Obenour, D. R., D. Scavia, N. N. Rabalais, R. E. Turner, and A. M. Michalak (2013), Retrospective analysis of midsummer hypoxic area and volume in the Northern Gulf of Mexico, 1985–2011, *Environ. Sci. Technol.*, *47*, 9808–9815.
Rabalais, N. N., R. E. Turner, and W. J. Wiseman (2002), Gulf of Mexico hypoxia, aka "The dead zone," *Annu. Rev. Ecol. Syst.*, *33*, 235–263.
Rosmond, T. E. (1992), The design and testing of the Navy Operational Global Atmospheric Prediction System, *Weather Forecast*, *7*, 262–272, doi:10.1175/1520-0434.
Scavia, D., N. N. Rabalais, E. R. Turner, D. Justić, and W. J. Wiseman Jr. (2003), Predicting the response of Gulf of Mexico hypoxia to variations in Mississippi River nitrogen load, *Limnol. Oceanogr.*, *48*, 951–956.
Schaeffer, B. A., G. A. Sinclair, J. C. Lehrter, M. C. Murrell, J. C. Kurtz, R. W. Gould, and D. F. Yates (2011), An analysis of diffusive light attenuation in the northern Gulf of Mexico hypoxic zone using the SeaWiFS satellite data record, *Remote Sens. Environ.*, *115*, 3748–3757.
Scully, M. E. (2010), Wind modulation of dissolved oxygen in Chesapeake Bay, *Estuaries Coasts*, *33*, 1164–1175.
Scully, M. E. (2013), Physical controls on hypoxia in Chesapeake Bay: A numerical modeling study, *J. Geophys. Res. Oceans*, *118*, 1239–1256, doi:10.1002/jgrc.20138.
Turner, R. E., N. N. Rabalais, and D. Justić (2006), Predicting summer hypoxia in the northern Gulf of Mexico: Riverine N, P, and Si loading, *Mar. Pollut. Bull.*, *52*, 139–148.
Wang, L., and D. Justić (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*, 1464–1476.
Wallcraft, A. J., E. J. Metzger, and S. N. Carroll (2009), *Software Design Description for the HYbrid Coordinate Ocean Model (HYCOM), Version 2.2*, technical report, Nav. Res. Lab., NRL/MR/7320–09-9166, Stennis Space Cent., Miss.
Wanninkhof, R. (1992), Relationship between wind speed and gas exchange, *J. Geophys. Res.*, *97*, 7373–7382.
Wilson, R. F., K. Fennel, and P. Mattern (2013), Simulating sediment-water exchange of nutrients and oxygen: A comparative assessment of models against mesocosm observations, *Cont. Shelf Res.*, *63*, 69–84.
Wiseman, W., N. Rabalais, R. Turner, S. Dinnel, and A. Mac-Naughton (1997), Seasonal and interannual variability within the Louisiana coastal current: Stratification and hypoxia, *J. Mar. Syst.*, *12*, 237–248.
Yu, L., K. Fennel, A. Laurent, M. C. Murrell, and J. C. Lehrter (2015a), Numerical analysis of the primary processes controlling oxygen dynamics on the Louisiana shelf, *Biogeosciences*, *12*, 2063–2076, doi:10.5194/bg-12-2063-2015.
Yu, L., K. Fennel, and A. Laurent (2015b), A modeling study of physical controls on hypoxia generation in the northern Gulf of Mexico, *J. Geophys. Res. Oceans*, *120*, 5019–5039, doi:10.1002/2014JC010634.
Zhang, X., M. Marta-Almeida, and R. Hetland (2012a), A high-resolution pre-operational forecast model of circulation on the Texas-Louisiana continental shelf and slope, *J. Oper. Oceanogr.*, *5*(1), 19–34.
Zhang, X., R. D. Hetland, M. Marta-Almeida, and S. F. DiMarco (2012b), A numerical investigation of the Mississippi and Atchafalaya freshwater transport, filling and flushing times on the Texas-Louisiana shelf, *J. Geophys. Res.*, *117*, C11009, doi:10.1029/2012JC008108.
Zhang, Z., R. Hetland, and X. Zhang (2014), Wind-modulated buoyancy circulation over the Texas-Louisiana shelf, *J. Geophys. Res. Oceans*, *119*, 5705–5723, doi:10.1002/2013JC009763.