Evaluating Confidence in the Impact of Regulatory Nutrient Reduction on Chesapeake Bay Water Quality

Isaac D. Irby
Virginia Institute of Marine Science

Marjorie A.M. Friedrichs
Virginia Institute of Marine Science

CrossMark

Evaluating Confidence in the Impact of Regulatory Nutrient Reduction on Chesapeake Bay Water Quality

Isaac D. Irby[1] · Marjorie A. M. Friedrichs[1]

Abstract

Excess nutrients derived from anthropogenic activity have resulted in the degradation of coastal water quality and an increase in low-oxygen and hypoxic events worldwide. In an effort to curb these impacts and restore water quality in the Chesapeake Bay, a maximum load of nutrients has been established based on a framework of regulatory standards and models. This research aims to evaluate the projected changes in water quality resulting from the implementation of these nutrient reductions by applying the regulatory methodology to two different models that have been previously shown to have similar model skill. Results demonstrate that although the two models differ structurally and produce a different degree of absolute change, they project a similar relative improvement in water quality along the main stem of the Chesapeake Bay and the lower reaches of the tributaries. Furthermore, the models largely agree on the attainment of regulatory water quality standards as a result of nutrient reduction, while also establishing that meeting water quality standards is relatively independent of hydrologic (wet/dry) conditions. By developing a Similarity Index that compares model results across habitat, time, and methodology, this research identifies the locations and causes of greatest uncertainty in modeled projections of water quality. Although there are specific locations and times where the models disagree, overall this research lends support and increased confidence to the appropriateness of the nutrient reduction levels and in the general impact of nutrient reduction on Chesapeake Bay water quality under current environmental conditions.

Keywords Water quality · Management · Modeling · Chesapeake Bay

Introduction

As the largest estuary in the continental USA with a watershed supporting a growing population of over 18 million people, the Chesapeake Bay ($\sim 11{,}000$ km^2) is particularly prone to water quality degradation as a result of human activity. With decreases in dissolved oxygen (DO) concentrations amplifying since the 1950s (Cooper and Brush 1993; Curtin et al.

2001; Hagy et al. 2004; Bever et al. 2013), fish habitat has been compressed, catch per unit effort has decreased (Buchheister et al. 2013), and harvest of the region's iconic Blue crabs has been negatively affected (Mistiaen et al. 2003).

Multiple efforts by state, federal, and private partners to improve the Bay's water quality eventually resulted in the 2010 Chesapeake Bay Total Maximum Daily Load (TMDL), which established location-specific mandated pollutant reductions throughout the six states and Washington, D.C. that make up the Chesapeake Bay watershed (Fig. 1 and Table 1). The specific level of reduction was set to ensure that minimum water quality standards would be eventually met.

The Chesapeake Bay Program (CBP) utilized a complex modeling system, including an airshed model, watershed model, and water quality model to inform the level of nutrient reduction in the TMDL (USEPA 2010a). The nutrient reductions required to improve the water quality of the Chesapeake Bay are estimated to cost in the tens of billions of dollars (Nelson 2014). With such astounding potential costs, it is

✉ Marjorie A. M. Friedrichs
marjy@vims.edu

[1] Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, VA 23062, USA
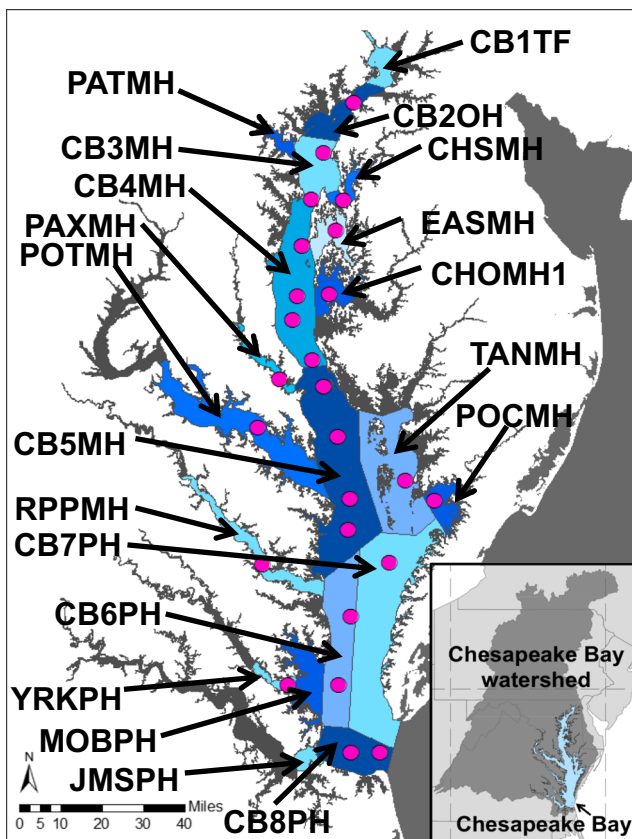
Fig. 1 Map of the Chesapeake Bay showing CBP observation stations (pink circles) and Bay segments (blue regions) used in this analysis. (See Table 1) For those segments that were split in half for regulatory purposes by the TMDL along the Virginia/Maryland border, the combined segment was utilized in this analysis

Table 1 Observation stations and segments shown in Fig. 1

Segment	Stations used in SI analysis
CB1TF	CB1.1, CB2.1
CB2OH	CB2.2, CB3.1
PATMH	WT5.1
CB3MH	CB3.2, CB3.3C
CHSMH	ET4.2
CB4MH	CB4.1C, CB4.2C, CB4.3C, CB4.4
EASMH	EE1.1
CHOMH1	EE2.1
PAXMH	LE1.1, LE1.2, LE1.3
CB5MH	CB5.1, CB5.2, CB5.3, CB5.4
TANMH	EE3.1, EE3.2
POTMH	RET2.4, LE2.2, LE2.3
POCMH	EE3.4
RPPMH	LE3.1, LE3.2, LE3.4
CB6PH	CB6.1, CB6.2, CB6.3, CB6.4
CB7PH	CB7.1, CB7.2, CB7.3, CB7.4
YRKPH	LE4.2, LE4.3
MOBPH	WE4.1
JMSPH	LE5.4, LE5.5
CB8PH	CB8.1, CB8.1E

crucial for regulatory efforts to be successful and that the modeling used to inform the regulation is sound.

Previous research, incorporating the water quality model used by the CBP, has demonstrated that utilizing multiple models can better inform our understanding of the Chesapeake Bay (Irby et al. 2016; Irby 2017). The research described here extends this earlier model comparison effort by comparing how different models respond to potential nutrient reductions, with the goal of enhancing our understanding of the likely success of management decisions. The current research effort builds on this idea by exploring how to compare models beyond simply evaluating model skill and contrasting model structure.

Traditional model comparisons (e.g., Friedrichs et al. 2006, 2007, 2009) focus strictly on evaluating model output through the conventional types of skill metrics used to compare and assess model skill (e.g., Stow et al. 2009; Jolliff et al. 2009). However, when assessing the potential impact of management decisions, it is important to extend the comparison beyond a set of skill metrics and examine the potential achievement of

management goals. As regulations like the nutrient reductions for the Chesapeake Bay and comparisons of complex models are both relatively new, an established framework for assessing how similar models are in terms of their simulation of water quality standard attainment does not yet exist. In the case of the Chesapeake Bay, this type of comparison requires an assessment of whether or not water quality standards will be met across multiple habitats assuming a required reduction in nutrient loading is achieved. Interannual variability and the complexity of the regulatory methodology further complicate the issue. As a result, the comparison effort described here assesses how similar the models are in terms of the binary evaluation of whether or not water quality standards are attained across habitats, years, and the methodology used to go from model output to the assessment of water quality standard attainment.

The research presented here provides a framework from which to compare multiple models throughout the individual steps it takes to get from model output to assessment of water quality standard attainment. As management decisions are increasingly informed by computer models, it is crucial to repeat the same management procedure with multiple models in an effort to explore the potential success of a management decision. Specifically, this research identifies the locations and causes of greatest uncertainty in modeled projections of water quality.

Modeling Methodology

In this study, nutrient reduction scenarios were applied to two very different water quality models. One model was developed and used in the context of managing the Chesapeake Bay and is hereafter referred to as Curvilinear Hydrodynamics in Three Dimensions–Integrated Compartment Model (CH3D-ICM) (Cerco et al. 2010; Cerco and Noel 2013), while the other was used in a research context and is hereafter referred to as Chesapeake Bay Regional Ocean Modeling System–Estuarine Carbon Biogeochemistry model (ChesROMS-ECB) (Feng et al. 2015). In the subsections below, first the primary characteristics of each model are briefly summarized and then their structural differences are described.

Water Quality Models

CH3D-ICM

The CBP's water quality and sediment transport model, the CH3D-ICM; Cerco et al. 2010; Cerco and Noel 2013), is a coupled hydrodynamic-biogeochemical model that was used to inform development of the Chesapeake Bay TMDL. The model employs a curvilinear boundary-fitted horizontal grid consisting of 11,604 horizontal cells with an average wet cell resolution of 1 km and a 1.52-m vertical z-grid. The ecological water column component of the model includes 24 state variables that interact with a sediment diagenesis module (DiToro 2001). CH3D-ICM has been extensively calibrated for the Chesapeake Bay and has been in use and development since the 1980s (Johnson et al. 1993). Model output was provided by the CBP and is specifically from the version of the model that was used in the development of the TMDL.

ChesROMS-ECB

The ChesROMS-ECB (Feng et al. 2015) is a coupled hydrodynamic-biogeochemical model based on the Regional Ocean Modeling System (ROMS; Shchepetkin and McWilliams 2005) with the specific model domain and curvilinear horizontal grid based on the Chesapeake Bay modeling community's ChesROMS model (Xu et al. 2012) with an average wet cell resolution of 1.7 km. The vertical framework follows a sigma grid with 20 layers and a stretching parameter that compresses the grid (higher resolution) at the air-water and sediment-water interfaces while expanding it (lower resolution) mid-water column.

The biogeochemical component of ChesROMS-ECB is modified from Fennel et al. (2006), Druon et al. (2010), Hofmann et al. (2008, 2011), and Cahill et al. (2016) to be applicable for estuarine rather than coastal applications (Feng et al. 2015). These modifications include the addition of an inorganic suspended solid state variable, water column denitrification, oxygen limitation of nitrification, and a new parameterization for light attenuation that is a function of suspended particulate matter and salinity (as a proxy for colored dissolved organic matter; Xu et al. 2005). Although the model includes deposition and erosion of suspended solids, it does not include a full biogeochemical sediment module that allows for temporary burial and delayed resuspension due to tide and wave activity (e.g., Moriarty et al. 2017).

While the TMDL mandates the reduction of nitrogen, phosphorus, and sediment, ChesROMS-ECB used in this study does not include phosphorus. While not including phosphorus may be a limitation for realistically simulating marine nutrient cycles, ChesROMS-ECB has been calibrated with nitrogen as the only nutrient available for biologic uptake and is able to successfully simulate DO variability especially in the deep mid-Bay during the hypoxic summer season (Feng et al. 2015; Irby et al. 2016), since this time and region are more limited by nitrogen than by phosphorus (Testa et al. 2014).

To facilitate model comparisons, ChesROMS-ECB output was vertically mapped to the grid of CH3D-ICM. Where the CH3D-ICM grid depths were deeper (shallower) than those of ChesROMS-ECB, the ChesROMS-ECB output profiles were linearly stretched (compressed) to match those of the CH3D-ICM output.

Structural Model Differences

While both models have been used to study the same system, there are important differences between the two models. The horizontal resolution for ChesROMS-ECB (1.7 km) is lower than that of CH3D-ICM, resulting in a more strongly smoothed bathymetry. These bathymetric differences are most apparent on the shoals and mouth of the Bay where ChesROMS-ECB is generally shallower than the true bathymetry. In contrast, the vertical resolution is much greater in ChesROMS-ECB than CH3D-ICM, especially in the shallow regions where there are 20 depth levels for ChesROMS-ECB and only one for CH3D-ICM. Another difference between the models is the locations at which watershed loads are delivered to the Bay. In CH3D-ICM, watershed loads are input along the entire land-water interface and can thus include shoreline erosion, whereas in ChesROMS-ECB watershed loads enter only via ten major rivers (Feng et al. 2015). In addition, in the case of CH3D-ICM, the open boundary condition is located across the mouth of the Chesapeake Bay. In contrast, the ChesROMS grid has the open boundary located far offshore (Xu et al. 2012), so that the conditions used along this boundary (in this case radiation conditions for biogeochemical variables; see Feng et al. 2015) do not dramatically impact the estuarine hydrodynamics and biogeochemistry.

The differences in biogeochemical complexity between the two models are even more significant and largely arise from the differences in the original purposes of the two models. While CH3D-ICM was developed to address water quality management efforts, ChesROMS-ECB was developed to quantify the Chesapeake Bay nitrogen cycle (Feng et al. 2015). As a result, CH3D-ICM contains many additional modules that are absent from ChesROMS-ECB, including a wetlands module, an SAV model, and an oyster model. In addition, the biogeochemical component of ChesROMS-ECB contains only one phytoplankton and one zooplankton state variable, which is much simpler than many more complex models (Xiao and Friedrichs 2014). In contrast, CH3D-ICM contains multiple plankton classes (including freshwater phytoplankton, diatoms, and green algae) and multiple limiting nutrients as well (Cerco et al. 2010; Irby et al. 2016). However, ChesROMS-ECB is also more complex than empirical models that have been shown to adequately simulate seasonal DO dynamics in the Chesapeake Bay utilizing a parameterized constant oxygen consumption rate (Scully 2010; Bever et al. 2013; Irby et al. 2016). Unlike CH3D-ICM, ChesROMS-ECB also contains a full inorganic/organic carbon cycle, with required riverine inputs derived from Tian et al. (2015).

The two models also underwent considerably different calibration processes. With respect to calibration years, CH3D-ICM primarily focused on 1991–2000 (Cerco et al. 2010), while ChesROMS-ECB was calibrated to 2001–2005 (Feng et al. 2015). CH3D-ICM also examined a broad set of observations that included cruise stations in the tributaries, as well as flow-through and moored data collected through the CBP Shallow Water Monitoring Program. In contrast, ChesROMS-ECB focused on nitrogen main stem cruise station observations (dissolved organic nitrogen, particulate organic nitrogen, nitrate, and ammonium), since this model was developed with the aim of quantifying the nitrogen budget of the Chesapeake Bay. In addition, ChesROMS-ECB was calibrated using satellite-derived information (e.g., chlorophyll and primary production). Overall, however, CH3D-ICM has undergone a much more extensive and iterative calibration process since its initial development in the 1980s, compared to the more recently developed ChesROMS-ECB.

The difference in skill between the two models has been examined in detail by Irby et al. (2016). This earlier work demonstrated that although these two models, and the other six models participating in the model comparison analysis, had difficulty resolving the variables that are generally considered to be the main drivers of oxygen variability (e.g., stratification, nutrients, chlorophyll), all eight models had similar skill in reproducing the mean and seasonal variability in DO along the main stem of the Bay in 2004 and 2005. Somewhat surprisingly this was the case even though some of the models, e.g., ChesROMS-ECB, were not calibrated

using any oxygen data and other models, e.g., CH3D-ICM, were not calibrated using data from these specific years. An extension of this skill assessment for the two models used here, including more years, more observation stations, and an additional variable (light attenuation), demonstrates that both models again exhibit a similar level of skill despite the many differences between model structure and calibration methodology described above (see Online Resource A.)

Nutrient Forcing

CBP Watershed Model Inputs

To ensure consistency between model runs, both models were forced at the land-water interface by the CBP regulatory watershed model version 5.3.2 (WSM; Shenk and Linker 2013). This is the version used in the development of the TMDL. The WSM, based on Hydrologic Simulation Program - Fortran (HSPF), utilizes multiple components to simulate land use, river flow, and the loading of nutrients and sediment to the Bay and has been in continual development with the input of multiple stakeholders since 1982 (Linker et al. 2002; Shenk and Linker 2013). The WSM was used to allocate the requisite location-specific reductions required by each jurisdiction throughout the watershed (Shenk and Linker 2013). Because the watershed and estuarine models express organic nutrient constituents differently, the watershed constituents had to be partitioned into the estuarine constituents. For ChesROMS-ECB, specifically, to match the organic nitrogen constituents in to those of the watershed model: (1) the estuarine refractory dissolved organic nitrogen was set to be 20% of the watershed total refractory organic nitrogen, (2) the estuarine semi-labile dissolved organic nitrogen included 100% of the biological oxygen demand from organic nitrogen and 80% of the phytoplankton nitrogen, and (3) the estuarine particulate organic nitrogen included 80% of the refractory organic nitrogen and 20% of the phytoplankton nitrogen. These assumptions are consistent with the partitioning used in the development of ChesROMS-ECB (Feng et al. 2015).

Standard Run and TMDL-WIP Scenario

Two WSM nutrient load scenarios were applied to both estuarine models used in this study. The first was the calibration scenario for the WSM, which the CBP used as a baseline for the TMDL. This scenario provides an estimate of the observed watershed loads for 1991–2000. Estuarine model results utilizing the WSM calibration scenario will hereafter be referred to as the "standard run." The second scenario applied was the TMDL Watershed Implementation Plan nutrient reduction scenario, hereafter referred to as the "TMDL-WIP scenario." This scenario assumes all nutrient reduction strategies have been successfully implemented and is very similar to the
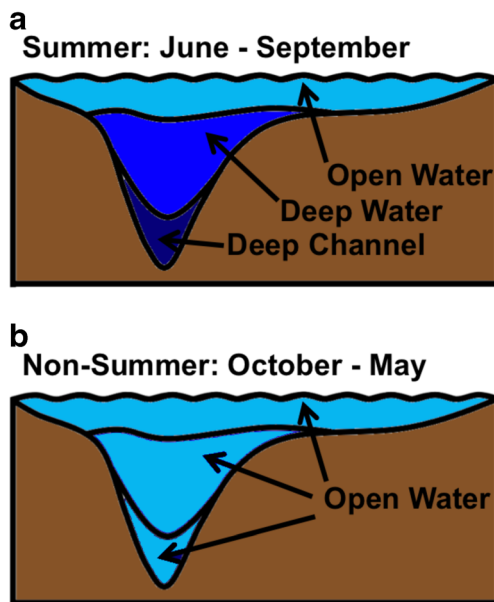
Fig. 2 Schematic of designated uses throughout the water column for **a** summer and **b** non-summer seasons

actual scenario used to establish the TMDL. Specific information on the nutrient reduction loads can be found in Shenk and Linker (2013) and include a percent reduction from the study period average of 33% for total nitrogen, 27% for total phosphorus, and 27% for sediment across the entire watershed.

Direct atmospheric nitrogen deposition for both models (Da et al. 2018) is derived from the CBP airshed model, which combines output from the Community Multiscale Air Quality model (CMAQ) for dry deposition and a regression model for wet deposition (Linker et al. 2013a). The TMDL allocates a maximum annual direct deposition of 7.1 million kg N to the tidal waters of the Bay (USEPA 2010a; Linker et al. 2013a). This represents a 30% reduction of direct atmospheric nitrogen deposition from the study period average.

The impact of the TMDL-WIP reduced nutrient scenario is compared for the estuarine models by examining the absolute as well as relative changes in DO concentrations, both spatially and temporally. The relative change in DO is calculated by dividing the absolute change by the concentration from the standard run.

CBP Procedure for Assessing Attainment of Water Quality Standards

Designated Uses

The CBP evaluates water quality for multiple habitats across the Bay and its tributaries (Tango and Batiuk 2013). These habitats are termed designated uses and are characterized in the Chesapeake Bay by ecological use (USEPA 2003, 2010b). Each designated use (Fig. 2) has a specific mandated minimum DO criterion, otherwise known as a DO Water Quality Standard (Table 2). Because of the importance of seasonal differences in DO, the TMDL specifies different Water Quality Standards (WQS) for the summer (June–September) and non-summer (October–May) seasons. During the non-summer, Open Water encompasses the entire water column. During the summer, the Open Water designation incorporates all surface water environments across the Bay and extends down to the bottom of the mixed surface layer, if there is one. The Deep Water designated use represents the summer transitional zone of the water column that is influenced by the pycnocline, incorporating all water below the well mixed surface layer and above the well mixed deep layer. The Deep Channel designated use encompasses the deep summer waters of the main stem trench and deep tributaries where physical characteristics limit the elevation of DO concentrations regardless of controls on water quality during the summer months. The depths delineating each designated use are

Table 2 Dissolved oxygen Water Quality Standards (WQS) by designated use (adapted from Tango and Batiuk 2013)

Designated use	Dissolved oxygen Water quality standard	Rationale	Timeframe
Open water	30-day mean ≥ 5.0 mg/L (tidal habitats with salinity ≥ 0.5 PSU)	Protects growth of larval, juvenile, and adult fish and shellfish as well as threatened/endangered species	Year-round
	7-day mean ≥ 4.0 mg/L	Protects survival of open water fish larvae	
	Instantaneous minimum ≥ 3.2 mg/L	Protects survival of threatened/ endangered sturgeon species	
Deep water	30-day mean ≥ 3.0 mg/L	Protects survival and recruitment of Bay anchovy eggs and larvae	June 1–September 30
	1-day mean ≥ 2.3 mg/L	Protects survival of open water juvenile and adult fish	
	Instantaneous minimum ≥ 1.7 mg/L	Protects survival of Bay anchovy eggs and larvae	
Deep channel	Instantaneous minimum ≥ 1.0 mg/L	Protects survival of bottom-dwelling worms and clams	June 1–September 30

defined by the observed physical characteristics for each month and are therefore non-uniform across time and space. The migratory fish spawning and nursery designated use and the shallow water Bay grass designated use were not individually evaluated in this study because they follow the Open Water DO criteria for the summer months.

Protocol for Determining Attainment of Water Quality Standards

To examine whether nutrient reductions will result in DO WQS being met, i.e., attained, the output from the standard run and TMDL-WIP scenario of both estuarine models underwent the published process for identifying attainment of WQS as established by the EPA and CBP (USEPA 2010a). While a brief synopsis of the methodology is below, a more complete documentation of this process can be found in the Online Resource B and throughout the Chesapeake Bay TMDL literature (USEPA 2010a; Linker et al. 2013b; Tango and Batiuk 2013).

The nutrient reduction levels were not computed as a function of the absolute DO concentrations simulated by the models; rather, the regulation uses the change in DO between the standard run and reduced nutrient scenario. To quantitatively estimate the change in DO due to the nutrient reduction at a specific location and specific time, the hourly output for each month at each vertical cell at each station (Table 1) for the standard and scenario runs were regressed against each other using a simple linear regression (Fig. 3). (Also see USEPA Appendix H 2010). The resulting linear regression was used to create a scenario-modified dataset by inputting an actual observed DO concentration from 1991 to 2000 into the regression equation as the independent variable and obtaining a projected DO concentration as a result of the nutrient

reduction scenario. By utilizing this regression method, the errors in the true predictive capabilities of the models are minimized since the models are not being used to predict an explicit future DO concentration, but rather they are predicting the relative change in DO concentrations that can be expected as a result of decreased nutrient availability.

Once the regressions were applied and the set of future "observations" were generated, a stoplight analysis was used to identify the percent time and space that the volume of water in question met or exceeded the mandated WQS. This process uses the CBP Interpolator (USEPA 2012) to interpolate scenario-modified DO concentrations throughout the Bay. The regulations were designed with the flexibility to allow water quality levels to exceed minimum standards, i.e., fail, for a specific percent time and space while still being granted an overall passing grade based on a cumulative distribution function reference curve (USEPA 2010b; Tango and Batiuk 2013). In the stoplight analysis, green represents the percent time and space that WQS are met, yellow represents the percent time and space that WQS are not met but are still within the buffer allowed by the cumulative distribution function reference curve, and red represents the percent time and space that WQS are not met and are beyond the buffer. Therefore, while all three colors are utilized in the stoplight analysis, only red signifies a segment that has failed, or exceeded, the regulatory standards. All percentages are rounded to the nearest whole percent and an exceedance of 1% red is deemed allowable due to impacts from rounding and computational uncertainty. Furthermore, to account for the fact that some locations in the Bay might exhibit low DO concentrations even under pristine conditions, an extra allowance was given in specific cases (Table 3). These extra allowances, defined as "variances," are allocated at the state level and therefore the stoplight analyses presented here do not include the variances that have been granted; however, their significance will be considered in the Discussion.
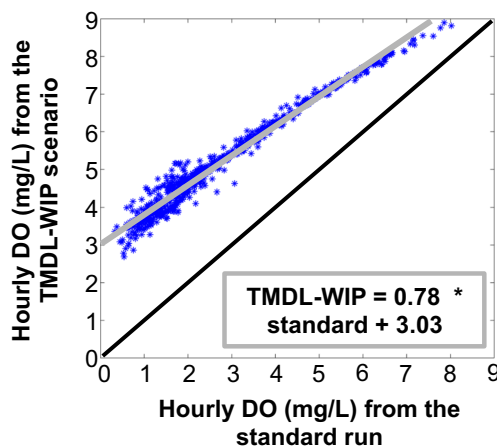
Study Period (1991–2000) and Critical Period (1993–1995)

The Chesapeake Bay TMDL was established using a 10-year hydrologic study period from 1991 to 2000 (USEPA 2010a; Linker et al. 2013b). This study period was chosen because it characterized a representative 10-year variability of freshwater

Fig. 3 Example regression between hourly DO from the standard run and hourly DO from the TMDL-WIP scenario for a single depth at a single observation station in a single month. Black line represents the 1:1 line; gray line represents best-fit regression

Table 3 Variances allowed in certain Maryland segments

Designated use	Segment	Variance
Deep water	CB4MH	7%
	PATMH	7%
Deep channel	CB4MH	2%
	EASMH	2%
	CHSMH	16%

flow and was fully encompassed within the 1985–2005 period for which model results were available. Within the 10-year study period, the 3-year critical period of 1993–1995 was used as the basis for the TMDL assessment. These 3 years were selected based on being representative of a relatively high flow period, because higher stream flow has been found to result in larger nutrient fluxes from the watershed and ultimately worse water quality conditions (Murphy et al. 2011). While the study period encompasses the entire 10 years of 1991–2000, the majority of the research presented in this study is focused on the 3-year critical period of 1993–1995 that was used in the TMDL regulations. To explore the impact of the choice of 3-year period on projected water quality, the results of every 3-year period between 1991 and 2000 (eight total periods) were put through the regulatory WQS protocol.

New Methodology for Assessing Similarity in WQS Attainment: the Similarity Index

To evaluate the degree of similarity in the projected impact of nutrient reductions and go beyond simple comparisons of model skill, a similarity index (SI) is introduced that incorporates multiple forms of information regarding how similarly the models respond to nutrient reduction. Assessing similarity at the segment level (SI_{Seg}) allows for an easily digestible framework for visualizing the metric. Specifically, the similarity index is computed for a given segment (SI_{Seg}) as the average of three metrics that examine the models across habitat (H_{Seg}), time (T_{Seg}), and methodology (M_{Seg}):

$$SI_{Seg} = \left(H_{Seg} + T_{Seg} + M_{Seg} \right)/3 \tag{1}$$

where:

$$H_{Seg} = \sqrt{(DCH_a - DCH_b)^2 + (DWH_a - DWH_b)^2 + (OWH_a - OWH_b)^2 + (OWSH_a - OWSH_b)^2} \tag{2}$$

$$T_{Seg} = \sqrt{(DCT_a - DCT_b)^2 + (DWT_a - DWT_b)^2 + (OWT_a - OWT_b)^2 + (OWST_a - OWST_b)^2} \tag{3}$$

$$M_{Seg} = \sqrt{(m_a - m_b)^2 + (b_a - b_b)^2 + (r_a - r_b)^2} \tag{4}$$

The terms in Eqs. 2, 3, and 4 are defined similarly for each of the habitats or seasons and are explained below.

For the habitat metric (Eq. 2), the sub-metrics include the Deep Channel (DCH), Deep Water (DWH), Open Water (OWH), and Open Water Summer (OWSH) and are computed for both models (i.e., models a and b), for example as:

$$DCH_a = \sqrt{(\%Gn_a)^2 + (\%Ye_a)^2 + (\%Rd_a)^2} \tag{5}$$

for the DCH of model a where %Gn, %Ye, and %Rd represent the green, yellow, and red percentages in the stoplight analysis (see section "CBP procedure for assessing attainment of water quality standards") for each particular habitat over the 1993–1995 time period. (Note that not all segments contain all four designated uses; in these cases, only the existing designated uses are included in the calculation). Thus, the H_{Seg} metric represents the similarity of the stoplight analysis results for each of the designated uses in a given segment.

For the time metric (Eq. 3), the sub-metrics include the Deep Channel (DCT), Deep Water (DWT), Open Water (OWT), and Open Water Summer (OWST). These are computed as the standard deviation of the percentage red (%Rd) for each segment across the eight 3-year periods from 1991 to 2000, for example the DCT for model a is represented as:

$$DCT_a = \sqrt{\frac{\sum_{i=1}^{8} \left(\%Rd_{a_i} - \overline{\%Rd_a} \right)^2}{7}} \tag{6}$$

The time metric thus represents the similarity in the impact of the choice of 3-year critical period, i.e., how similarly the models perform in wet vs. dry years, across all designated uses.

Finally, the sub-metrics for the methodology metric (Eq. 4) are computed from the regressions used to determine attainment of WQS, i.e., the average hourly slope (m), y-intercept (b), and correlation coefficient (r) for all observation depths at the stations in Table 1. This metric compares the intermediate methodology that is used for the stoplight analysis and is based only on the summer regressions, as that is when low-oxygen concentrations are of greatest management concern.

Finally, the full SI is then computed as the average of the three metrics (Eq. 1) for each segment. As part of the analyses conducted for this paper, a variety of different methods for computing the full SI were tested. For example, each metric was normalized by its largest value, before computing the average. In general, although the different methods for computing the full SI altered the specific values of the SI, these alternate methods did not substantially change the ranking of the most and least similar segments and thus did not change the overall results and conclusions of this analysis. Furthermore, by not normalizing the SI, it is possible to directly compare multiple iterations of this analysis as the models continue to be improved. It is important to note that the utility and generality of the SI methodology are derived from comparing model results across multiple metrics; the definition and calculation of each individual metric are a function of the details of the specific system to which the SI is applied.

Results

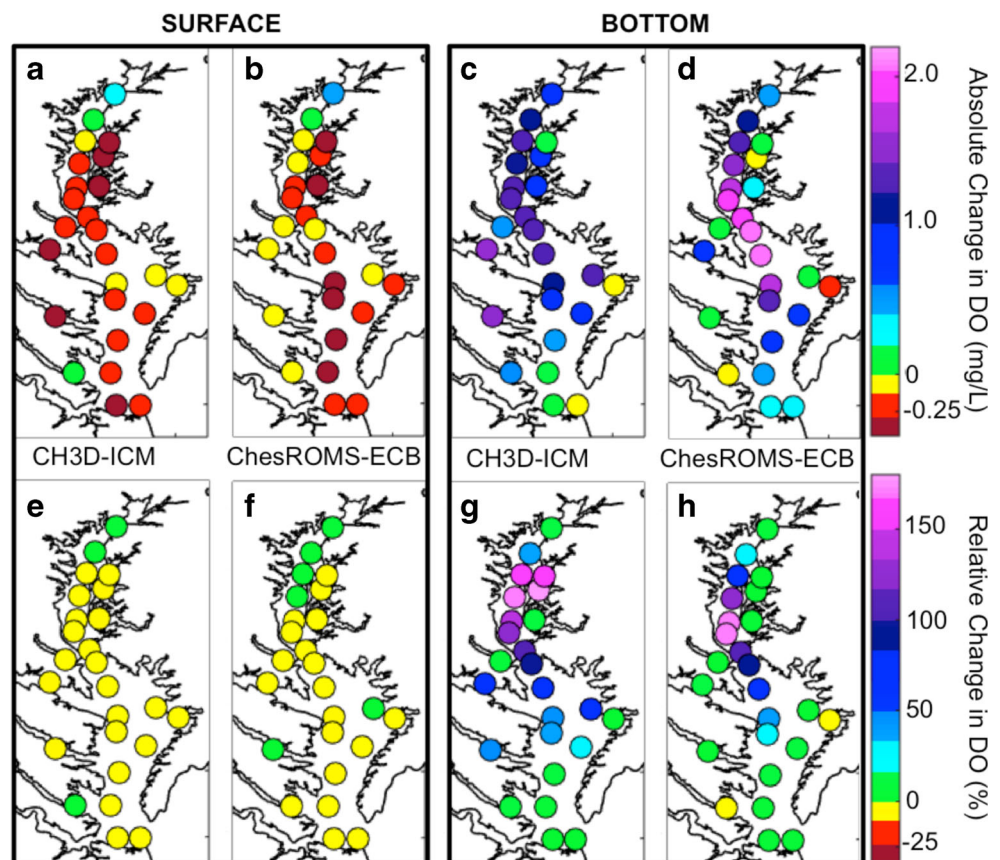## Comparison of the Standard Runs Versus the TMDL-WIP Scenarios

When the TMDL-WIP nutrient reduction scenario is applied to both models for 1993–1995, the models produce similar changes in summer DO concentrations; however, especially at the bottom, the relative changes are more similar than the absolute changes (Fig. 4). Although throughout most of the Bay the nutrient reductions result in decreases in surface DO for both models, at the northernmost stations, the models simulate an increase in surface DO. Whereas ChesROMS-ECB simulates slightly larger decreases in the southern half of the Bay compared to CH3D-ICM (~ 0.25 mg L^{-1}), the relative change in DO between the models (Fig. 4e, f) is similar across the entire Bay surface (~ 0–5%).

At the bottom, the TMDL-WIP nutrient reduction scenarios result in the absolute increase in summer DO in ChesROMS-ECB being roughly 1 mg L^{-1} higher than CH3D-ICM along the central main stem (Fig. 4c, d); however, again the relative increase is remarkably similar in magnitude between the two models with increases in DO between 100 and 175% (Fig. 4g, h). The differences in relative impact between the models are accentuated in some of the tributaries where ChesROMS-ECB generally produces small relative changes in summer bottom DO while CH3D-ICM simulates relatively large increases. This is most evident in the Chester River and Eastern Bay (CHSMH and EASMH; Fig. 1) where CH3D-ICM simulates a large relative increase in DO of ~ 160% while ChesROMS-ECB simulates a modest relative increase of ~ 5%.

The nutrient reduction scenarios cause both models to exhibit a larger increase in DO concentrations during the summer than in the winter (Fig. 5). During the summer in the mesohaline main stem (CB3MH and CB4MH), ChesROMS-ECB simulates a slightly larger increase in bottom DO (Fig. 5b) than CH3D-ICM (Fig. 5a), particularly in the dry year of 1995. However, the difference between the two standard runs (Fig. 5c; black line 5c = black line 5a–black line 5b) is larger than the difference between their changes in DO (Fig. 5c; orange line 5c = (blue line 5a–black line 5a)–(blue line 5b–black line 5b)). In other words, the models' simulated change in DO is more similar to each other than their simulation of the absolute DO concentration.

Fig. 4 Model results at 25 observation stations illustrating the absolute difference in summer DO concentration (Scenario Run − Standard Run) at the surface (**a**, **b**) and bottom (**c**, **b**), and the relative change in DO concentration ((Scenario Run − Standard Run) / Standard Run) at the surface (**e**, **f**) and bottom (**g**, **h**). CH3D-ICM results are displayed in **a**, **c**, **e**, and **g**. ChesROMS-ECB results are displayed in **b**, **d**, **f**, and **h**
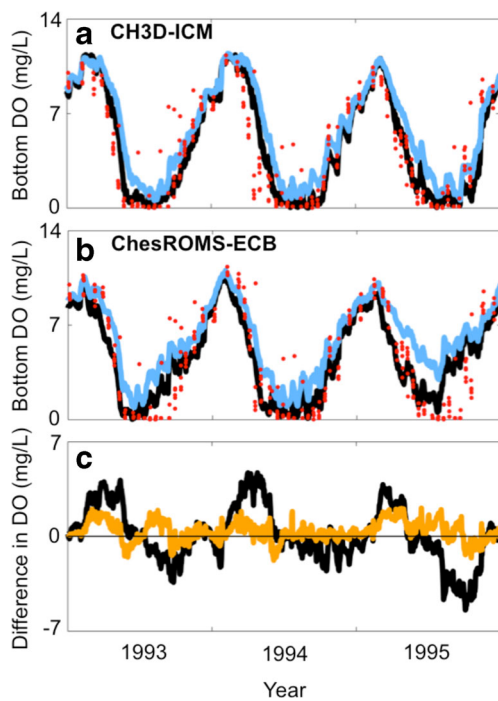
Fig. 5 Times series of average modeled bottom DO concentrations across main stem stations in CB3MH and CB4MH for **a** CH3D-ICM and **b** ChesROMS-ECB at a representative mid-Bay deep channel station. In **a** and **b**, the black line represents the standard run, the blue line represents the TMDL-WIP scenario, and the red dots represent the observed bottom oxygen at stations CB3.2, CB3.3C, CB4.1C, CB4.2C, CB4.3C, and CB4.4. In **c**, the black line is the standard run of CH3D-ICM minus the standard of ChesROMS-ECB; the orange line is the difference between the standard run and TMDL-WIP scenario for CH3D-ICM minus the difference between the standard run and TMDL-WIP scenario for ChesROMS-ECB

Comparison of Water Quality Standard Attainment for Both Models

When the results of the model simulations are put through the CBP protocol for determining whether or not WQS would be met with the TMDL-WIP nutrient reduction, the two models predict surprisingly similar results (Table 4) relative to the difference in their raw model output (Fig. 4c, d). With 0% red in the stoplight analysis for Open Water Summer and Open Water Non-Summer, both models simulate a complete attainment of WQS in these habitats. While the Open Water Non-Summer was widely in attainment before nutrient reduction, the Open Water Summer designated use was not. The models begin to diverge in Deep Water where the percent agreement in the stoplight analysis falls below 95% for four of the segments. In only two of these segments, do the models disagree on percent non-attainment (red) by > 1%. Differences are larger in the Deep Channel waters where four of the eight segments disagree by > 1%. In general, the greatest differences occur in the mid-Bay main stem and mid-Bay tributaries.

While the model simulations of WQS attainment differ most in the Deep Water and Deep Channel portions of the water column, the increase in WQS attainment compared to the 1993–1995 levels is still quite similar between the models (Fig. 6). Based on observations from 1993 to 1995, all Deep Channel and the majority of Deep Water volumes were out of attainment with many segments exhibiting substantial percentages of red (Fig. 6a, d). In the Deep Channel, both models simulate a considerable reduction in non-attainment with nearly all segments reducing non-attainment to < 1% (Fig. 6e, f). In the Deep Water, the pattern of non-attainment diverges between the models particularly in the tributaries with ChesROMS-ECB generating non-attainment in the Rappahannock and Patuxent Rivers and CH3D-ICM generating non-attainment in the Patapsco and Chester Rivers (Fig. 6b, c). Both models identify issues in the Eastern Bay. Along the main stem, the non-attainment simulated in ChesROMS-ECB is isolated to CB4MH. In CH3D-ICM, the non-attainment spans the entire mid-Bay from CB3MH to CB5MH. Overall, however, both models simulate a dramatic improvement in WQS attainment compared to the 1993–1995 levels (Fig. 6b, c compared to Fig. 6a and Fig. 6e, f compared to Fig. 6d) and many of the differences between the models are due to non-attainments of << 1%.

Comparison of the Impact of 3-Year Period

The two models behaved similarly across all eight 3-year periods examined, with both models exhibiting higher non-attainment of WQS during 3-year periods that encompassed multiple wet years (1993, 1994, 1996, 1998; Fig. 7). As was seen for the 1993–1995 period (Table 4), the Deep Water generally exhibits the largest percent non-attainment for both models regardless of which years are examined. Deep Water non-attainment in ChesROMS-ECB is below ~ 0.4% for each time period examined except for 1996–1998. This corresponds to a 3-year period of prolonged high flows. While CH3D-ICM also simulates the highest percent Deep Water non-attainment for 1996–1998 (~ 1.6%), it is not as large of a difference between that time period and the other wet periods (1993–1995, 1994–1996) as is seen for ChesROMS-ECB, which is more than twice as large as any other 3-year period. Non-attainment in the Deep Channel for ChesROMS-ECB is also particularly high in 1996–1998, resulting in one of the few instances where ChesROMS-ECB exhibits greater non-attainment than CH3D-ICM. However, even considering the variability of non-attainment, the non-attainment for any given designated use across all eight potential time

Table 4 Percent similarities (%Sim) and stoplight analysis results of green (%Gn), yellow (%Ye), and red (%Rd) percentages of CH3D-ICM (top, gray shading) and ChesROMS-ECB (bottom) for 1993–1995 water quality standard assessment
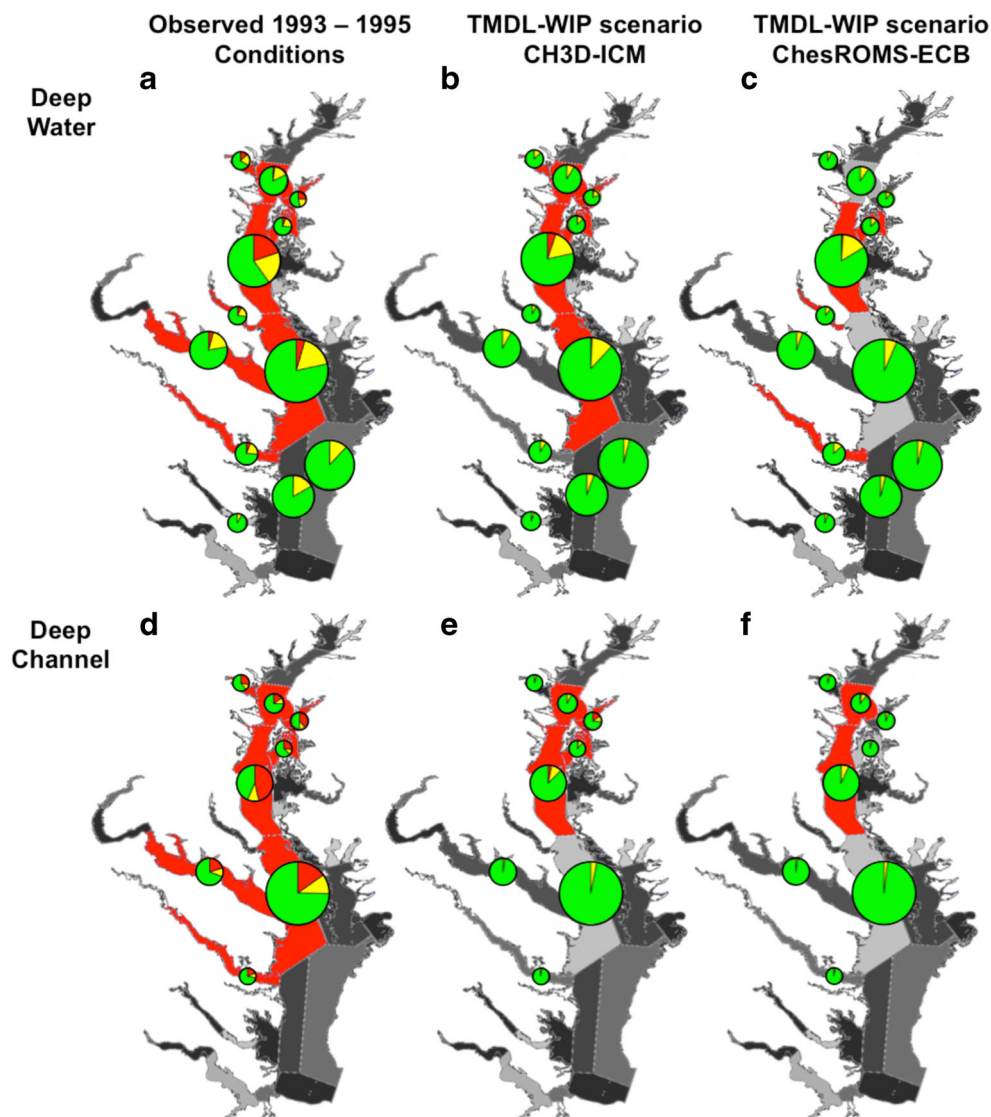
Segment	Model	Deep Channel %Sim	%Gn	%Ye	%Rd	Deep Water %Sim	%Gn	%Ye	%Rd	Open Water Summer %Sim	%Gn	%Ye	%Rd	Open Water Non-Summer %Sim	%Gn	%Ye	%Rd
CB1TF	CH3D-ICM									100	96	4	0	100	98	2	0
	ChesROMS-ECB										96	4	0		98	2	0
CB2OH	CH3D-ICM									99	94	6	0	100	98	2	0
	ChesROMS-ECB										95	5	0		98	2	0
PATMH	CH3D-ICM	100	95	5	0	94	87	13	0	100	95	5	0	100	97	3	0
	ChesROMS-ECB		95	5	0		93	7	0		95	5	0		97	3	0
CB3MH	CH3D-ICM	96	93	7	0	99	91	9	0	100	96	4	0	100	97	3	0
	ChesROMS-ECB		89	9	2		90	10	0		96	4	0		97	3	0
CHSMH	CH3D-ICM	79	73	11	16	91	79	18	3	100	96	4	0	100	98	2	0
	ChesROMS-ECB		94	6	0		88	12	0		96	4	0		98	2	0
CB4MH	CH3D-ICM	95	88	9	3	94	79	16	5	99	95	5	0	99	95	5	0
	ChesROMS-ECB		93	7	0		84	15	1		96	4	0		96	4	0
EASMH	CH3D-ICM	92	88	10	2	99	87	12	1	100	96	4	0	100	97	3	0
	ChesROMS-ECB		96	4	0		88	12	0		96	4	0		97	3	0
CHOMH1	CH3D-ICM									99	94	6	0	100	98	2	0
	ChesROMS-ECB										93	7	0		98	2	0
PAXMH	CH3D-ICM					98	92	8	0	97	95	5	0	100	98	2	0
	ChesROMS-ECB						90	10	0		92	8	0		98	2	0
CB5MH	CH3D-ICM	99	97	3	0	94	88	11	1	100	96	4	0	99	97	3	0
	ChesROMS-ECB		98	2	0		94	6	0		96	4	0		98	2	0
TANMH	CH3D-ICM									99	93	7	0	100	98	2	0
	ChesROMS-ECB										94	6	0		98	2	0
POTMH	CH3D-ICM	100	98	2	0	97	92	8	0	100	96	4	0	100	98	2	0
	ChesROMS-ECB		98	2	0		95	5	0		96	4	0		98	2	0
POCMH	CH3D-ICM									100	96	4	0	100	98	2	0
	ChesROMS-ECB										96	4	0		98	2	0
RPPMH	CH3D-ICM	100	97	3	0	96	91	9	0	100	96	4	0	100	98	2	0
	ChesROMS-ECB		97	3	0		87	13	0		96	4	0		98	2	0
CB6PH	CH3D-ICM					98	94	6	0	99	93	7	0	100	98	2	0
	ChesROMS-ECB						96	4	0		94	6	0		98	2	0
CB7PH	CH3D-ICM					100	96	4	0	97	90	10	0	100	98	2	0
	ChesROMS-ECB						96	4	0		93	7	0		98	2	0
YRKPH	CH3D-ICM					98	96	4	0	97	96	4	0	100	98	2	0
	ChesROMS-ECB						94	6	0		93	7	0		98	2	0
MOBPH	CH3D-ICM									99	93	7	0	100	98	2	0
	ChesROMS-ECB										92	8	0		98	2	0
JMSPH	CH3D-ICM									100	95	5	0	100	98	2	0
	ChesROMS-ECB										95	5	0		98	2	0
CB8PH	CH3D-ICM									99	96	4	0	100	98	2	0
	ChesROMS-ECB										95	5	0		98	2	0

periods for both models does not go above 1.6% and averages much less than 1.0%. As a result, while there is variability between the 3-year periods, the specific time period chosen does not have a major impact on the total non-attainment.

Examination of Similarity Index

Calculation of the SI, based on the three similarity metrics described in the previous sections, reveals a high degree of similarity between the models for the

Fig. 6 Pie charts showing attainment (green), attainment with buffer (yellow), and non-attainment (red) for the 1993–1995 observations (**a**, **d**), CH3D-ICM (**b**, **e**), and ChesROMS-ECB (**c**, **f**) for the Deep Water (**a**, **b**, **c**) and Deep Channel (**d**, **e**, **f**) designated uses. Size of the pies is relative to the volume of applicable water for that given segment. Segments coded in red exhibit a stoplight analysis of red that is greater than 0%

majority of Bay segments (Table 5 and Fig. 8). Exceptions include the Chester River (CHSMH) and Eastern Bay (EASMH), which exhibited the lowest degrees of similarity, and the central mid-Bay (CB4MH and CB5MH), Patapsco River (PATMH), and the York River (YRKPH). Across the rest of the Bay, the two models are very similar for the three metrics examined lending a degree of confidence in their projections of whether or not WQS will be attained with the requisite nutrient reduction. The SI for Mobjack Bay (MOBPH), lower Choptank River (CHOMH1), Pocomoke Sound (POCMH), lower James River (JMSPH), oligohaline main stem (CB2OH), lower polyhaline main stem (CB8PH), and Tangier Sound (TANMH) are particularly high, leading to a high confidence in these WQS attainment projections.
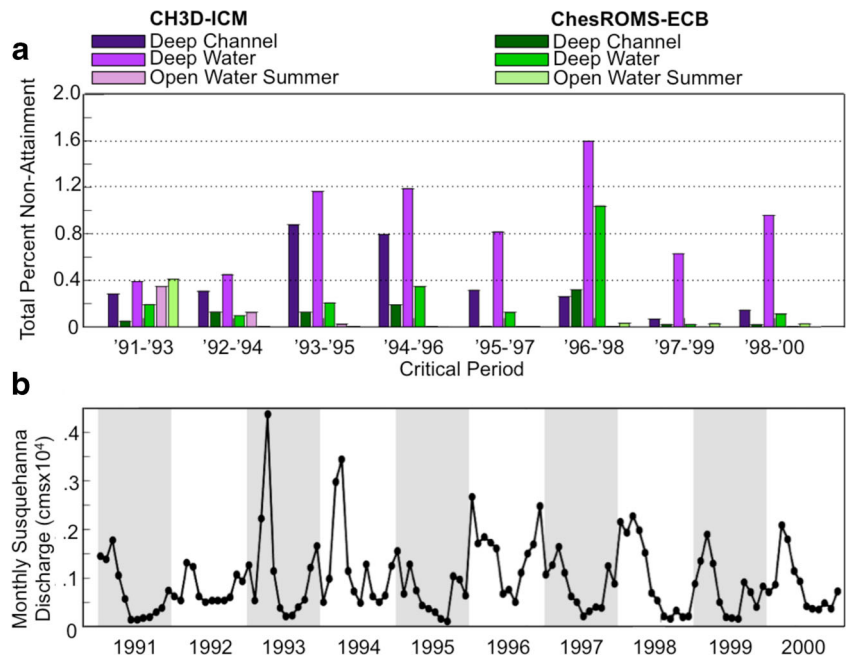
Discussion

How Do Chesapeake Bay Models Compare in Terms of How Nutrient Reduction Impacts DO Concentrations?

ChesROMS-ECB simulates a larger absolute improvement in DO compared to CH3D-ICM, but both models simulate a similar relative improvement in DO.

Along the main stem of the Chesapeake Bay, ChesROMS-ECB simulates a higher summer absolute increase in bottom DO as a result of nutrient reduction than CH3D-ICM (Fig. 4c, d). This difference continues up the water column, attenuating to the surface where the models perform more similarly (Fig. 4a, b). The difference in the absolute change in bottom

Fig. 7 **a** Total percent non-attainment for CH3D-ICM and ChesROMS-ECB for the Deep Channel, Deep Water, and Open Water Summer. **b** Monthly Susquehanna freshwater discharge from the CBP watershed model. Open Water Non-Summer is in near full attainment and therefore is not shown

DO between the models is potentially due to the positive bias of DO concentrations in ChesROMS-ECB (Irby et al. 2016; Online Resource A). This bias is likely caused by the lack of a full sediment diagenesis model and underestimated bottom depth caused by low horizontal resolution in regions of steep bathymetric gradients. At the surface, the decreased input of

nutrients causes both models to predict a decrease in DO. This decrease in DO is a result of decreased production in the surface layer of the water column and is consistent with other modeling studies exploring the impact of nutrients on water quality (e.g., Testa et al. 2014). Both models predict an increase in DO at the surface for the northernmost stations as a

Table 5 Values of the Similarity Index (SI) metrics for all 20 Bay segments

Segment	Habitat metric (H_{Seg})	Time metric (T_{Seg})	Methodology metric (M_{Seg})	Similarity index (SI_{Seg})
CB1TF	3.83	0.01	1.08	**1.64**
CB2OH	0.94	0.13	0.63	**0.57**
PATMH	5.30	0.32	0.92	**2.18**
CB3MH	3.89	0.21	0.34	**1.48**
CHSMH	20.19	4.02	0.76	**8.32**
CB4MH	6.73	0.75	0.52	**2.67**
EASMH	7.56	1.53	1.59	**3.56**
CHOMH1	0.93	0.16	1.23	**0.77**
PAXMH	3.31	0.60	0.88	**1.60**
CB5MH	5.67	0.28	0.80	**2.25**
TANMH	0.93	0.30	0.50	**0.58**
POTMH	2.78	1.42	0.48	**1.56**
POCMH	0.00	0.00	0.71	**0.24**
RPPMH	3.48	0.81	1.04	**1.77**
CB6PH	2.11	0.89	0.15	**1.05**
CB7PH	2.71	0.36	0.30	**1.12**
YRKPH	3.40	1.93	1.05	**2.13**
MOBPH	0.92	0.12	0.19	**0.41**
JMSPH	0.00	0.51	0.71	**0.41**
CB8PH	0.95	0.12	0.65	**0.58**

result of the decrease in sediment in the TMDL-WIP scenario,
which alleviates light limitation on production. This area has
the highest turbidity and consequently benefits most from the
reduction in sediment delivered to the Bay.

While the models disagree on the absolute change in bot-
tom DO as a result of the nutrient reductions (Fig. 4c, d), they
are surprisingly similar in terms of the relative change in DO
at the bottom and throughout the water column (Fig. 4e, h).
This is possible because ChesROMS-ECB generally simu-
lates higher bottom DO than CH3D-ICM, particularly in dry
years (Fig. 5; Irby et al. 2016; Online Resource A) and
ChesROMS-ECB also generally simulates a larger change in
absolute bottom DO as a result of the nutrient reduction

(~ 1 mg/L; Fig. 4c, d). The only important difference between
their simulated relative changes in DO is in the middle of the
main stem of the Bay at depth, where the magnitude of the
relative change is similar between the models but CH3D-ICM
places the maximum impact further north than does
ChesROMS-ECB. As discussed below, this has important
ramifications for the assessment of Chesapeake Bay water
quality standards.

How Do Chesapeake Bay Models Compare in Terms of Whether Nutrient Reductions Will Lead to the Desired Attainment of Water Quality?

Overall, the models predict very similar levels of water
quality standard attainment throughout most of the Bay.
However, the lowest similarity (least confidence) in the
impact of nutrient reduction on the attainment of water
quality standards is in the Chester River and Eastern
Bay, though substantial differences also exist in the
mid-Bay main stem where DO concentrations are low-
est each summer.

Water quality observations from 1993 to 1995 demonstrate
that there were large areas throughout the Deep Channel and
Deep Water of the Bay where water quality standards were not
being met (Fig. 6). Both models predict that the vast majority
of those exceedances will be alleviated once the nutrient re-
duction is in place, i.e., all but three Deep Channel and two
Deep Water segments for CH3D-ICM and all but one Deep
Channel segment for ChesROMS-ECB (Table 4).

Differences between the models in the Chester River and
Eastern Bay are likely the result of modeling deficiencies. In
the Chester River, CH3D-ICM predicts that even with the
required nutrient load reductions in place, 16% of the Deep
Channel will not meet the required water quality levels,
whereas ChesROMS-ECB is fully in attainment. This large
discrepancy is potentially due to a mischaracterization in
CH3D-ICM of oxygen concentrations in the lateral freshwater
flow entering the Chester River. The CH3D-ICM issue has
been identified during a shallow water study of the Chester
River (Friedrichs et al. 2012) and is currently being remedied
(C. Cerco, personal communication). Differences in the
Eastern Bay are likely attributable to the horizontal grid reso-
lution of ChesROMS-ECB, which leads to bottom depths that
are too shallow, particularly where bathymetric gradients are
strong. Where bottom depths are too shallow, the influence of
surface air-sea oxygen flux is too large, keeping bottom oxy-
gen concentrations higher than they should be. As a result, the
mean y-intercept of the regression differs by 2.16 mg/L, which
is a substantial difference when examining hypoxic waters.

The model differences in the mid-Bay are due to the
models simulating the location of greatest impact from

nutrient reduction in different locations. Since CH3D-ICM simulates its largest impact in CB3MH, CB4MH does not pass the WQS attainment as it does for ChesROMS-ECB, which simulates the largest impact in CB4MH. The impact of the location of maximum impact also affects CH5MH because in particularly poor water quality years (generally wetter years), the hypoxia pushes south into the segment. This raises the question of which model, if either, is correctly simulating the location of greatest impact.

The difference in location of greatest impact (Fig. 4g, h) is likely due to a variety of factors including grid differences and river influence. The z-grid employed by CH3D-ICM and the sigma grid employed by ChesROMS-ECB can cause an under- and over-smoothing of the true bathymetry, respectively. An accurate representation of the bathymetry is particularly important in segments CB3MH and CB4MH because the deep main stem channel ends just north of the border between the segments. Both model grids have the deep channel ending at station CB3.3C, but the bathymetric differences can lead to differences in the distribution of low oxygen. River influences can also impact this difference. Because ChesROMS-ECB only has ten major rivers, whereas CH3D-ICM has freshwater inputs along the entire land-water interface, it is more influenced by Susquehanna River flow. As a result, it is possible that the influence of the Susquehanna River pushes impacts further south. CB3MH bottom DO in CH3D-ICM may also be influenced by the bottom DO issues present in the Chester River, artificially drawing down oxygen levels in the main stem.

For water quality levels to pass the regulatory minima as mandated by the TMDL regulation, all areas must pass the WQS with no exceedances. To account for some numerical errors in calculating the volumes and percent space and time of attainment, the TMDL allows for a 1% buffer for all segments and all designated uses. Therefore, a stoplight analysis that exhibits 1% or less of "red" can still be considered in attainment. Unfortunately, even with the 1% rule, some segments and designated uses still do not meet WQS for both models. In the development of the TMDL, CH3D-ICM was tested using progressively more stringent nutrient reduction scenarios to explore just how much of a potential impact nutrient reductions could have. In some segments, the model never went to full attainment even when pre-industrial nutrient loads were simulated. Since all of the problem segments were located in the Maryland portion of the Bay, Maryland was able to account for these segments that would not fall into traditional attainment by allotting a "variance." The variances (Table 3) are defined by Maryland state regulation rather than in the TMDL regulation and only impact those segments identified as unable to meet WQS with the mandated nutrient reduction. The regulation states that the variances must be reviewed every 3 years as the modeling and understanding of the ecosystem are continually improving.

Only in the CB3MH Deep Water does ChesROMS-ECB require a variance in order to fall within attainment. CH3D-ICM, on the other hand, requires variances in five of the segments/designated uses. As discussed previously, the Chester River is a special case that is currently being studied by the CBP. However, the iteration of CH3D-ICM used in this analysis results in the Chester River Deep Water not attaining even with the variance. The difference in whether or not the models need the variances to meet WQS is important to note. The results of ChesROMS-ECB potentially indicate that some of the variances are the result of modeling artifacts such as inadequate boundary conditions or grid limitations and not the environment. This is critical to note, since the WQS are biologically based and exceedances of 16 or even 7%, as allowed by the variances, could prove biologically detrimental considering there are many important Bay species unable to tolerate low-DO conditions (USEPA 2003).

While the models differed in terms of their estimates of WQS attainment, neither model exhibited a strong sensitivity to the choice of study period when examining non-attainment relative to the entire volume of a designated use. The EPA underwent a complex process to identify the best 3-year period on which to base the hydrologic conditions of the TMDL. While there are certainly individual segments that exhibit sensitivity to the specific 3-year period chosen for the analysis, the research performed in this study (Fig. 7) indicates that when looking at an entire designated use, the models are relatively insensitive to the baseline hydrologic conditions in terms of attainment of WQS: in most cases, both models simulate an exceedance of less than 1% across entire designated uses. The models also similarly exhibit changes between 3-year periods with a higher percent exceedance in the wetter 3-year periods than in the drier ones. This is to be expected as there is an observed correlation between years with high freshwater flow and years with large hypoxic volumes (Murphy et al. 2011); thus, using a wetter 3-year period produces a more conservative result for both models.

What Is the Utility of the Similarity Index?

The Similarity Index compares models beyond their raw output and across the multiple metrics of habitat, time, and methodology to identify how models compare in terms of the impact of management decisions and changing environmental conditions. This broad comparison affords an understanding of how model similarity should inform confidence in the impact of management decisions.

The overall goal of this research was to quantify the similarity between multiple models in the attainment of WQS resulting from nutrient reduction. The Similarity Index

developed here is a tool to enhance understanding of the confidence in the potential impact of a management decision and offers insight into the segments where the models behave most similarly, i.e., where we have high relative confidence in their projection of the impact of reduced nutrient inputs, and where they behave least similarly, i.e., where we have relatively low confidence in the impact of reduced nutrient inputs. The utility of the SI is how it compares model simulations beyond a typical skill assessment analysis and informs confidence of the pass/fail outcome of a management decision.

As implemented here for the Chesapeake Bay, the SI includes the assessment of similarity across three metrics: habitat, time, and methodology. For some segments, such as Mobjack Bay (MOBPH), the two models used in this analysis produce similar results across all three metrics. This indicates that the models are similarly resolving the relative impact on DO resulting from nutrient reduction (across methodology), the status of attainment vs. non-attainment in different designated uses (across habitat), and the impact of the baseline hydrology (across time). If the model results are similar across this variety of metrics, confidence in the model projections should be relatively high. On the contrary, for a segment like the Chester River where the models perform very differently, the dissimilarity can trigger further research that can lead to an enhancement to our understanding of the system and a better modeling product. Ideally, this type of assessment will be redone after the identified issues have been resolved to see if similarity, and hence confidence in the impacts of nutrient reduction, has increased. This is an excellent example of how multiple model comparisons can result in the improvement of overall understanding and the identification of potential issues and shortcomings of individual models.

An SI like the one developed here can enhance the understanding of impacts from a change in the environment in any system. While simple model skill assessment comparisons are important and increasingly employed, when assessing the impact of management decisions, it is critical to compare how models respond to management decisions and environmental changes such as nutrient reduction. Furthermore, comparing models across multiple metrics significantly enhances the quality and quantity of understanding derived from a comparison of this sort. This study utilized three metrics that compared the actual answers the models were offering (across habitat), the impact of interannual variability (across time), and the intermediate steps required to get from raw model output to management application (across methodology). Not every application of a SI has to be for a management decision as formalized as the Chesapeake Bay TMDL. A study can compare multiple models across model results, the time-dependent variability of those results, and the intermediate steps to get those results. To enhance the value of the SI even more, additional models can be added.

Summary and Conclusions

Both models analyzed in this study simulate a similar level of attainment of Chesapeake Bay water quality standards as a result of regulatory nutrient reduction. While the models differ in their simulated absolute change in dissolved oxygen concentrations resulting from the nutrient reduction scenario, the relative change in DO between the models is quite similar. Since the methodology for evaluating the impact of nutrient reduction is based on a relative change within each model between the standard run and the nutrient reduction scenario, the models can differ in their simulation of the absolute change in DO while still simulating a similar level of water quality standards attainment.

Although the predicted attainment of water quality standards between the models is similar, there are locations in the Bay where there is particularly high similarity (and potentially a resultant high confidence) and locations where there is relatively low similarity (and potentially a resultant low confidence) in these projections. The parts of the Bay where similarity is lowest are the Chester River and Eastern Bay. The deep main stem (CB4MH and CB5MH) is also identified as a low similarity region, albeit slightly higher than for the Chester River and Eastern Bay. While specific modeling issues can potentially explain the particularly low similarity in the Chester River and Eastern Bay, the low similarity in the mid-Bay main stem is primarily a result of the models differing in the location of greatest impact from the nutrient reduction scenario with CH3D-ICM placing the greatest impact slightly further north than ChesROMS-ECB.

Although this study identified locations and sources of low similarity in estimates of the attainment of water quality resulting from nutrient reductions, overall the results presented here highlight that the similarities between the two sets of model results far outweighed the differences. This lends an increased degree of confidence in the anticipated impact of the regulated nutrient reduction of the Chesapeake Bay TMDL. Furthermore, the framework for assessing confidence in model predictions of water quality standard attainment, via the Similarity Index, can be expanded beyond the two models evaluated in this research. Including additional Chesapeake Bay models in the future will further inform confidence by enhancing understanding of the potential impact of nutrient reduction. The concept of evaluating the impact of management decisions with multiple models and across multiple metrics, such as habitat, time, and methodology, is new and can be used for other systems where management decisions are informed by computer models (e.g., HELCOM 2009, BSC 2009).

While this study utilizes a multiple model approach to evaluate model projections of the future, the future examined here is one with similar environmental and climatological conditions as the present day. This leads to the question of whether or not these results would stand if climate change impacts were added to the analysis. Although this study demonstrated

that the current regulatory nutrient reductions are likely to eventually produce the required DO improvements under the present climate, it is not clear that the established nutrient loads will be adequate under near-term future climate conditions that include rising temperature and sea level along with changes in precipitation patterns (Irby et al. 2018).

References

Bever, A.J., M.A.M. Friedrichs, C.T. Friedrichs, M.E. Scully, and L.W. Lanerolle. 2013. Combining observations and numerical model results to improve estimates of hypoxic volume within the Chesapeake Bay, USA. *Journal of Geophysical Research: Oceans* 118 (10): 4924–4944. https://doi.org/10.1002/jgrc.20331.

BSC. 2009. *Strategic action plan for the environmental protection and rehabilitation of the Black Sea*. Black Sea Commission, Sofia, Bulgaria.

Buchheister, A., C.F. Bonzek, J. Gartland, and R.J. Latour. 2013. Patterns and drivers of the demersal fish community of Chesapeake Bay. *Marine Ecology Progress Series* 481: 161–180. https://doi.org/10.3354/meps10253.

Cahill, B., J. Wilkin, K. Fennel, D. Vandemark, and M.A.M. Friedrichs. 2016. Inter-annual and seasonal variability in air-sea CO_2 fluxes along the U.S. Eastern Continental Shelf and their sensitivity to increasing air temperatures and variable winds. *Journal of Geophysical Research: Biogeosciences* 121 (2): 295–311. https://doi.org/10.1002/2015JG002939.

Cerco, C.F., and M.R. Noel. 2013. Twenty-one-year simulation of Chesapeake Bay water quality using the CE-QUAL-ICM eutrophication model. *Journal of the American Water Resources Association* 49: 1119–1133. https://doi.org/10.1111/lawr.12107.

Cerco, C., S.C. Kim, and M. Noel. 2010. *The 2010 Chesapeake Bay Eutrophication Model—a report to the US Environmental Protection Agency Chesapeake Bay Program and to The US Army Engineer Baltimore District*. Vicksburg: US Army Engineer Research and Development Center.

Cooper, S.R., and G.S. Brush. 1993. A 2,500-year history of anoxia and eutrophication in the Chesapeake Bay. *Estuaries* 16 (3): 617–626.

Curtin, P.D., G.S. Brush, and G.W. Fisher. 2001. *Discovering the Chesapeake: The history of an ecosystem*. Baltimore: Johns Hopkins University Press.

Da, F., M.A.M. Friedrichs, and P. St-Laurent. 2018. Impacts of atmospheric nitrogen deposition and coastal nitrogen fluxes on oxygen concentrations in Chesapeake Bay. *Journal of Geophysical Research, Oceans* 123. https://doi.org/10.1029/2018jc014009.

DiToro, D., 2001. Sediment Flux Modeling. John Wiley and Sons, New York City, New York, ISBN 0-471-13535-6.

Druon, J.N., A. Mannino, S. Sigorini, C. McClain, M. Friedrichs, J. Wilkin, and K. Fennel. 2010. Modeling the dynamics and export of dissolved organic matter in the Northeastern U.S. continental shelf. *Estuarine Coastal Shelf Science* 88 (4): 488–507. https://doi.org/10.1016/j.ecss.2010.05.010.

Feng, Y., M.A.M. Friedrichs, J. Wilkin, H. Tian, Q. Yang, E.E. Hofmann, J.D. Wiggert, and R.R. Hood. 2015. Chesapeake Bay nitrogen fluxes derived from a land-estuarine ocean biogeochemical modeling system: Model description, evaluation, and nitrogen budgets. *Journal of Geophysical Research: Biogeochemistry* 120 (8): 1666–1695. https://doi.org/10.1002/2015JG002931.

Fennel, K., J. Wilkin, J. Levin, J. Moisan, J. O'Reilly, and D. Haidvogel. 2006. Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Global Biogeochemical Cycles* 20 (3). https://doi.org/10.1029/2005GB002456.

Friedrichs, M.A.M., R.R. Hood, and J.D. Wiggert, 2006. Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data. Deep-Sea Research II, 53, 576-600.

Friedrichs, M.A.M., J. Dusenberry, L. Anderson, R. Armstrong, F. Chai, J. Christian, S.C. Doney, J. Dunne, M. Fujii, R. Hood, D. McGillicuddy, K. Moore, M. Schartau, Y. H. Spitz, J. Wiggert, 2007. Assessment of skill and portability in regional marine biogeochemical models: role of multiple phytoplankton groups. Journal of Geophysical Research , 112, CO8001, https://doi.org/10.1029/2006JC003852.

Friedrichs, M.A.M., M.-E. Carr, R. Barber, M. Scardi, D. Antoine, R.A. Armstrong, I. Asanuma, M.J. Behrenfeld, E.T. Buitenhuis, F. Chai, J.R. Christian, A.M. Ciotti, S.C. Doney, M. Dowell, J. Dunne, B. Gentili, W. Gregg, N. Hoepffner, J. Ishizaka, T. Kameda, I. Lima, J. Marra, F. Mélin, J.K. Moore, A. Morel, R.T. O'Malley, J. O'Reilly, V.S. Saba, M. Schmeltz, T.J. Smyth, J. Tjiputra, K. Waters, T.K. Westberry, A. Winguth, 2009. Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean. Journal of Marine Systems, 76, 113-133, https://doi.org/10.1016/j.jmarsys.2008.05.010

Friedrichs, M., K.G. Sellner, and M.A. Johnston. 2012. *Using multiple models for management in the Chesapeake Bay: A shallow water pilot project*. Chesapeake Bay Program Scientific and Technical Advisory Committee Report. No. 12-003, Edgewater, MD, 11 pp.

Hagy, J.D., W.R. Boyton, C.W. Keefe, and K.V. Wood. 2004. Hypoxia in Chesapeake Bay, 1950-2001: Long-term change in relation to nutrient loading and river low. *Estuaries* 27 (4): 634–658.

HELCOM. 2009. *Eutrophication in the Baltic Sea—An integrated thematic assessment of the effects of nutrient enrichment and eutrophication in the Baltic Sea region*. Helsinki Commission: Baltic Marine Environment Protection Commission, Balt. Sea Environ. Proc. No. 115B. Helsinki, Finland.

Hofmann, E.E., J. Druon, K. Fennel, M. Friedrichs, D. Haidvogel, C. Lee, A. Mannino, C. McClain, R. Najjar, J. O'Reilly, D. Pollard, M. Previdi, S. Seitzinger, J. Siewert, S. Sigorini, and J. Wilkin. 2008. Eastern US continental shelf carbon budget: Integrating models,

data assimilation, and analysis. *Oceanography* 21 (1): 86–104. https://doi.org/10.5670/oceanog.2008.70.

Hofmann, E.E., B. Cahill, K. Fennel, M.A.M. Friedrichs, K. Hyde, C. Lee, A. Mannino, R.G. Najjar, J.E. O'Reilly, J. Wilkin, and J. Xue. 2011. Modeling the dynamics of continental shelf carbon. *Annual Review of Marine Science* 3 (1): 93–122. https://doi.org/10.1146/annurev-marine-120709-142740.

Irby, I.D., M.A.M. Friedrichs, C.T. Friedrichs, A.J. Bever, R.R. Hood, L.W.J. Lanerolle, M. Li, L. Linker, M.E. Scully, K. Sellner, J. Shen, J. Testa, H. Wang, P. Wang, and M. Xia. 2016. Challenges associated with modeling low-oxygen waters in Chesapeake Bay: A multiple model comparison. *Biogeosciences* 13 (7): 2011–2028. https://doi.org/10.5194/bg-13-2011-2016.

Irby, I.D. 2017. Using Water Quality Models in Management – A Multiple Model Assessment, Analysis of Confidence, and Evaluation of Climate Change Impacts, Doctoral thesis, William & Mary, Williamsburg, VA, USA.

Irby, I.D., M.A.M. Friedrichs, F. Da, and K. Hinson. 2018. The competing impacts of climate change and nutrient reductions on dissolved oxygen in Chesapeake Bay. *Biogeosciences* 15: 2649–2668. https://doi.org/10.5194/bg-15-2649-2018.

Johnson, B.H., K.W. Kim, R.E. Heath, B.B. Hsieh, and H.L. Butler. 1993. Validation of a three-dimensional hydrodynamic model of Chesapeake Bay. *Journal of Hydraulic Engineering* 119 (1): 2–20.

Jolliff, J.K., J.C. Kindle, I. Schulman, B. Penta, M.A.M. Friedrichs, R. Helber, and R.A. Arnone. 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems* 76 (1-2): 64–82. https://doi.org/10.1016/j.jmarsys.2008.05.014.

Linker, L.C., G.W. Shenk, P. Wang, C.F. Cerco, A.J. Butt, P.J. Tango, and R.W. Savidge. 2002. *A comparison of Chesapeake Bay estuary model calibration with 1985–1994 observed data and method of application to water quality criteria*. A Report of the Modeling Subcommittee Chesapeake Bay Program Office, Annapolis, MD.

Linker, L.C., R. Dennis, G.W. Shenk, R.A. Batiuk, J. Grimm, and P. Wang. 2013a. Computing atmospheric nutrient loads to the Chesapeake Bay watershed and tidal waters. *Journal of the American Water Resources Association* 49: 1025–1041. https://doi.org/10.1111/jawr.12112.

Linker, L.C., R.A. Batiuk, G.W. Shenk, and C.F. Cerco. 2013b. Development of the Chesapeake Bay watershed total maximum daily load allocation. *Journal of the American Water Resources Association*: 1–21. https://doi.org/10.1111/jawr.12105.

Mistiaen, J.A., I.E. Strand, and D. Lipton. 2003. Effects of environmental stress on blue crab (Callinectes sapidus) harvests in Chesapeake Bay tributaries. *Estuaries* 26 (2): 316–322.

Moriarty, J.M., C.K. Harris, C. Rabouille, K. Fennel, M.A.M. Friedrichs, and K. Xu. 2017. The roles of resuspension, diffusion and biogeochemical processes on oxygen dynamics offshore of the Rhone River, France: A numerical modeling study. *Biogeosciences* 14 (7): 1919–1946. https://doi.org/10.5194/bg-14-1919-2017.

Murphy, R.R., W.M. Kemp, and W.P. Ball. 2011. Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts* 34 (6): 1293–1309. https://doi.org/10.1007/s12237-011-9413-7.

Nelson, R.H. 2014. How to save the Chesapeake Bay TMDL: The critical role of nutrient offsets. *William & Mary Environmental Law and Policy Review* 38.

Scully, M.E. 2010. The importance of climate variability to wind-driven modulation of hypoxia in Chesapeake Bay. *Journal of Physical*

Oceanography 40 (6): 1435–1440. https://doi.org/10.1175/2010JPO4321.1.

Shchepetkin, A.F., and J.C. McWilliams. 2005. The Regional Ocean Modeling System (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model* 9 (4): 347–404. https://doi.org/10.1016/j.ocemod.2004.08.002.

Shenk, G.W., and L.C. Linker. 2013. Development and application of the 2010 Chesapeake Bay watershed total maximum daily load model. *Journal of the American Water Resources Association* 49: 1–15. https://doi.org/10.1111/lawr.12109.

Stow, C.A., J.K. Jolliff, D.J. McGillicuddy, Jr., S.C. Doney, M.A.M. Friedrichs, J.I. Allen, K.A. Rose, and P. Wallhead, 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, 76, 4-15, https://doi.org/10.1016/j.marsys.2008.03.011

Tango, P.J., and R.A. Batiuk. 2013. Deriving Chesapeake Bay water quality standards. *Journal of the American Water Resources Association* 1–18. doi:https://doi.org/10.1111/jawr.12108.

Testa, J.M., Y. Li, Y.J. Lee, M. Li, D.C. Brady, D.M. Di Toro, W.M. Kemp, and J.J. Fitzpatrick. 2014. Quantifying the effects of nutrient loading on dissolved O_2 cycling and hypoxia in Chesapeake Bay using a coupled hydrodynamic-biogeochemical model. *Journal of Marine Systems* 139: 139–158. https://doi.org/10.1016/j.jmarsys.2014.05.018.

Tian, H., Q. Yang, R.G. Najjar, W. Ren, M.A.M. Friedrichs, C.S. Hopkinson, and S. Pan. 2015. Anthropogenic and climatic influences on carbon fluxes from eastern North America to the Atlantic Ocean: A process-based modeling study. *Journal of Geophysical Research: Biogeosciences* 120 (4): 757–772. https://doi.org/10.1002/2014JG002760.

USEPA. 2003. *Ambient water quality criteria for dissolved oxygen, water clarity and chlorophyll a for the Chesapeake Bay and its tidal tributaries*. U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD. EPA 903-R-03-002.

USEPA. 2010a. *Chesapeake Bay total maximum daily load for nitrogen, phosphorus, and sediment*. U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD.

USEPA. 2010b. *Ambient water quality criteria for dissolved oxygen, water clarity and chlorophyll a for the Chesapeake Bay and its tidal tributaries: 2010*. Technical Support for Criteria Assessment Protocols Addendum. U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD. EPA 903-R-10-002.

USEPA. 2012. *Guide to using Chesapeake Bay program water quality monitoring data*. U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD. EPA 903-R-12-001.

USEPA Appendix H. 2010. *Chesapeake Bay total maximum daily load for nitrogen, phosphorus, and sediment: appendix H—Criteria assessment procedures using model scenario output with bay monitoring data*. U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD.

Xiao, Y., and M.A.M. Friedrichs. 2014. Using biogeochemical data assimilation to assess the relative skill of multiple ecosystem models in the Mid-Atlantic Bight: Effects of increasing the complexity of the planktonic food web. *Biogeosciences* 11 (11): 3015–3030. https://doi.org/10.5194/bg-11-3015-2014.

Xu, J., R.R. Hood, and S.Y. Chao. 2005. Modeling the light field in a partially mixed estuary. *Estuaries* 28 (4): 572–580.

Xu, J., W. Long, J.D. Wiggert, L.W.J. Lanerolle, C.W. Brown, R. Murtugudde, and R.R. Hood. 2012. Climate forcing and salinity variability in Chesapeake Bay, USA. *Estuarine, Coastal, and Shelf Science* 35 (1): 237–261. https://doi.org/10.1007/s12237-011-9423-5.