Response to Referee #1, P. MacCready:

We greatly appreciate P. MacCready's input on this manuscript and hope that we have fully addressed the comments/questions provided.

General Comments: *The authors systematically compare the skill of 8 3D models of Chesapeake Bay circulation and biogeochemistry. They focus on hypoxia, but consider other related properties such as mixed layer depth. They find that all models do a reasonable job at simulating hypoxia compared to two years of ~monthly observations at 13 stations. Like temperature (with which the models also have high skill) oxygen has a large seasonal cycle, contributing to its predictability. All models had poor skill at predicting the depth of the start of the hypoxic layer (very important for the ecosystem and management). The authors show that this problem is related to lack of skill predicting the density mixed layer depth.*

This is an important piece of work. This level of model inter-comparison has rarely or never occurred for estuarine systems. The paper is well-written, and the figures well-chosen. I have only a few smaller comments, and recommend that it be accepted with minor revisions.

Response: Thank you for your support of this manuscript.

Smaller Comments:

1. Page 20371, lines 7-9. The "minimum stratification criterion" is mentioned here, and it seems like a good way to ignore casts with minimal gradients, but I could not find where this criterion is defined. Please clarify.

1. Response: Thank you for identifying this unclear statement. The limitation imposed of only considering stratification to be present if there is at least a 10% change per meter in a given variable's profile is the minimum stratification criterion.

1. Manuscript Edit: The wording of the manuscript has been changed to: "The minimum stratification criterion utilized in this analysis requiring a profile to pass the 10% threshold also ensures that observations where very little…"

2. Page 20372, lines 5-6. The phrase about "the skill of a model defined as the mean of the observations" was unclear. In general the authors do a good job explaining the statistical tests, but in this case another sentence might help.

2. Response: Thank you for identifying this unclear statement.

2. Manuscript Edit: The following sentence was added directly after the referred to statement: "In effect, this means that if a model falls within the unit circle, it exhibits a skill that is greater than the skill obtained if one were to simply use the mean of the observations."

3. Page 20373, bottom. It might help to explain when and where it is of greatest value (e.g. to managers) to get the DO right, and why.

3. Response: As this is the results section, we do not feel this is the appropriate place to address the impact of DO for management. We do, however, speak to this point in the discussion in section 4.3 where we indicate that "the summer months are when the mismatch has the greatest potential to impact the available habitat for oxygen-dependent species" and later in section 4.3, that "while not all of the main stem stations develop hypoxic water each year, most mesohaline stations experience a dramatic drawdown of oxygen to levels during the summer that effectively remove a large portion of the Bay from habitable space."

3. Manuscript Edit: To emphasize this point at the beginning of the manuscript, the following was added to the second paragraph in the introduction: "The greatest impact of low DO concentrations spatially will depend on the specific living resource; however, temporally, late spring to early fall is of most concern."

4. Page 20374, line 13. Why is it that all the models have the same biases in the stratification field?

4. Response: A great question. There are likely quite a few reasons why this may be the case. The first that comes to mind is that the advection schemes implemented by these models are generally overly diffusive. This tends to smooth out sharp gradients in the vertical. As for the MLD occurring too high in the water column, this is potentially partially due to the model bathymetries being shallower than the true bathymetry at these stations (this issue is not the case for CH3D-ICM, which employs a z-grid that can more easily match the true bathymetry). In addition, this may also be a result of an underestimation of the wind field or the lack of diffuse freshwater input around the Bay.

4. Manuscript Edit: To address these two points the following text was added to the second and third paragraphs of section 4.2, respectively: "The underestimation of the vertical gradient across all models is largely due to the numerical diffusion that characterizes all of these hydrodynamic models, but may also be partially due to an underestimation of the winds or the lack of diffuse freshwater input around the Bay." "Furthermore, Model A employs a *z* grid that matches the bathymetry in trench areas better than the sigma grids used by the other models."

5. Page 20375, line 20. It is interesting that the mean of the models has these timing errors, but I'm not sure what it shows. Two years is too few to say anything statistical about timing. Probably OK to mention, though.

5. Response: We agree that this time period does not allow us to say anything specific about timing.

5. Manuscript Edit: To ensure an interannual comparison is not implied by the text, the sentence has been changed to: "While this study does not allow for a true interannual comparison, it is interesting that at station CB4.1C whereas the model ensemble closely matches the timing…"

6. Page 20378, lines 24-25. This sentence is unclear. In what way are the biological drivers, "not… spatially explicit"?

6. Response: This was meant to mean that the biological drivers do not necessarily need to be found at that exact location to have an impact on the DO at that exact location.

6. Manuscript Edit: To increase clarity and brevity, the sentence has been edited as: "This is likely due to the ephemeral nature of the biological drivers of DO."

7. Fig. 1. Give the length scale in km.

7. Response: Good point. This same issue was in Fig 2.

7. Manuscript Edit: Correction has been made in both figures.

8. Fig. 8b. The very poor correlation of Chl seems to make sense for this inherently patchy process. What is more perplexing is the large underestimate of the standard deviation. Any thoughts?

8. Response: Great point. It is true that the patchiness causes the poor correlation to make sense. However, the poor correlation is not completely due to the fact that the models are patchy in the wrong locations, but rather that they are less patchy and thus generally have a more constant spatial distribution of Chl than is observed. The models do not exhibit the strong variability demonstrated by the observed Chl distributions since the Chl is not patchily distributed in the models, but rather, it is more (although, certainly not completely) uniformly distributed. This is why the standard deviation is substantially underestimated.

8. Manuscript Edit: To address this, the following was added to the first paragraph of Section 4.2: "While the models generally simulate the total amount of chlorophyll adequately, it is more uniformly spatially distributed in the models rather than in patchy blooms as in nature, leading to the underestimation of chlorophyll variability across all models."

9. Fig. 9. Please make the x-axis ticks more regular so that the same time in different years is easier to compare visually.

9. Response: Good point.

9. Manuscript Edit: Figure has been edited so that every fourth month is indicated.

Response to Referee #2, Anonymous:

We greatly appreciate your input on this manuscript and hope that we have fully addressed the comments/questions provided.

General Comments: *Overall, this is a well-conceived modeling study that compares predictions of eight coupled hydrodynamic-biogeochemical models that were independently developed for the Chesapeake Bay against the data collected on biweekly to monthly monitoring cruises conducted during 2004-2005. In terms of the number of models involved, this is certainly one of the more comprehensive model comparisons conducted for coastal ecosystems. The members of the team are skillful modelers that have extensively published on the subject and the methods and conclusions are generally sound and scientifically defensible. The paper is well written and suitable for publication, subject to minor revision as suggested below.*

Response: Thank you for your support and comments on this manuscript.

Comments:
1. The conclusion that all models predict the seasonal dynamics of dissolved oxygen reasonably well, regardless of their structural complexity of spatial resolution, is not surprising. Extensive model comparisons conducted with climate models have taught us a very important modeling lesson – eight climate models that produce nearly identical hindcasts for the past 2,500 years, strongly disagree in their predictions for the next 85 years for the same climate scenario. I guess there is a simple answer to that – calibration. Modelers have become very good in calibrating their models, and given sufficient time and data, even a model of dubious mechanistic value will end up displaying a remarkable skill. The only way to critically evaluate the model results would be if they are subjected to a rigorous validation using data to which the models were not exposed during calibration. Because of the different data requirements, this would be very difficult to accomplish with a large number of fairly complex models, and I am not suggesting that the authors embark on that journey. However, some discussion would be needed to clarify whether the 2004-2005 data set that was used for model comparison was also used for model calibration.

1. Response: This is a very important point. All models were calibrated independently by the individual researchers/research groups, but the calibration was not exclusively focused on these 13 stations (there are ~100 stations in the Bay with monthly/semi-monthly data), these specific two years (data are available from 1985-2015) or these particular variables (many other variables, e.g. ammonium, total suspended solids, particulate organic nitrogen, etc…. are also available). Thus although the data included in this study may have been used in the overall evaluation/calibration analyses of the individual models, it is unlikely that it played a major role in this process as it represents only a small subset of data available for this purpose. In general, model evaluation/calibration was completed before we requested output from the individual modeling teams, and thus the teams were not trying to fit the specific data used in this study.

1. Manuscript Edit: The following text has been added to Section 2.2: "While these observations were publicly available for model assessment during calibration of all of the models, they represent a very small subset of the 30 years of EPA observations across over 100 Bay stations. The models compared here were calibrated based on access to the larger data set and for conditions in the Bay in general, not specifically for the 13 stations and two years considered here."

2. My second point is that I would like to have seen a more detailed analysis of the model-data comparison. For example, Fig. 9 shows that models collectively predict a duration of hypoxia compared to the measurements, and that the predicted onset of hypoxia during 2005 lags substantially with respect to the measurements. As much as I appreciate Taylor and target diagrams, I think that simple scatter plots of predicted versus observed DO values for individual models would have been very useful in that regard.

2. Response: A larger discussion on the timing issues evident in Fig. 9 was not included, since the short 2-year time frame of the comparison precluded a full analysis of interannual variability. To the referee's point regarding the value of a set of model to observations scatter plots, unfortunately scatter plots will not provide information on timing or duration of hypoxia. In addition, these seem unnecessary since the information garnered from a scatter plot is manifested in the Taylor (correlation) and target (bias) diagrams. Figure R1 below illustrates the bottom and surface DO concentrations for all 13 stations and all observation times for the model mean. From the diagram, one can see that the model mean is biased towards underestimating the observations at the surface and slightly biased towards overestimating the observations at the bottom, while the correlations for both are fairly high. This same information is demonstrated on Fig 9. If we included this type of figure for every variable (7) and every model (8), this would mean an additional 56 figures, even if we included multiple depths on each figure as in the example shown. To minimize the number of necessary figures, we believe the combination of using Taylor and target diagrams together is sufficient to demonstrate the skill of the models; however we do note that the scatter diagram provided below will of course be permanently available to readers online on the manuscript discussion page as part of this response to the reviews.

2. Manuscript Edit: No edits were made.

3. My third point concerns the selection of model data for monthly comparison. I am not sure what the word "monthly" refers to. Were the model results outputted to match the dates of the biweekly to monthly monitoring cruises, or were they averages for the entire month?

3. Response: Thank you for pointing out this unclear and important aspect. Some of this information was presented in the third paragraph of section 2.4, but we have now clarified this section further.

3. Manuscript Edit: The following text was added/edited in the third paragraph of section 2.4: "Model skill was assessed using the hourly model output (daily for CH3D-ICM chlorophyll and nitrate) that was nearest in time to that of the observation and from the grid cell that encompassed the observation location. For months with two observations, each observation was individually matched to the model output and the skill statistics from those comparisons were averaged for that month."

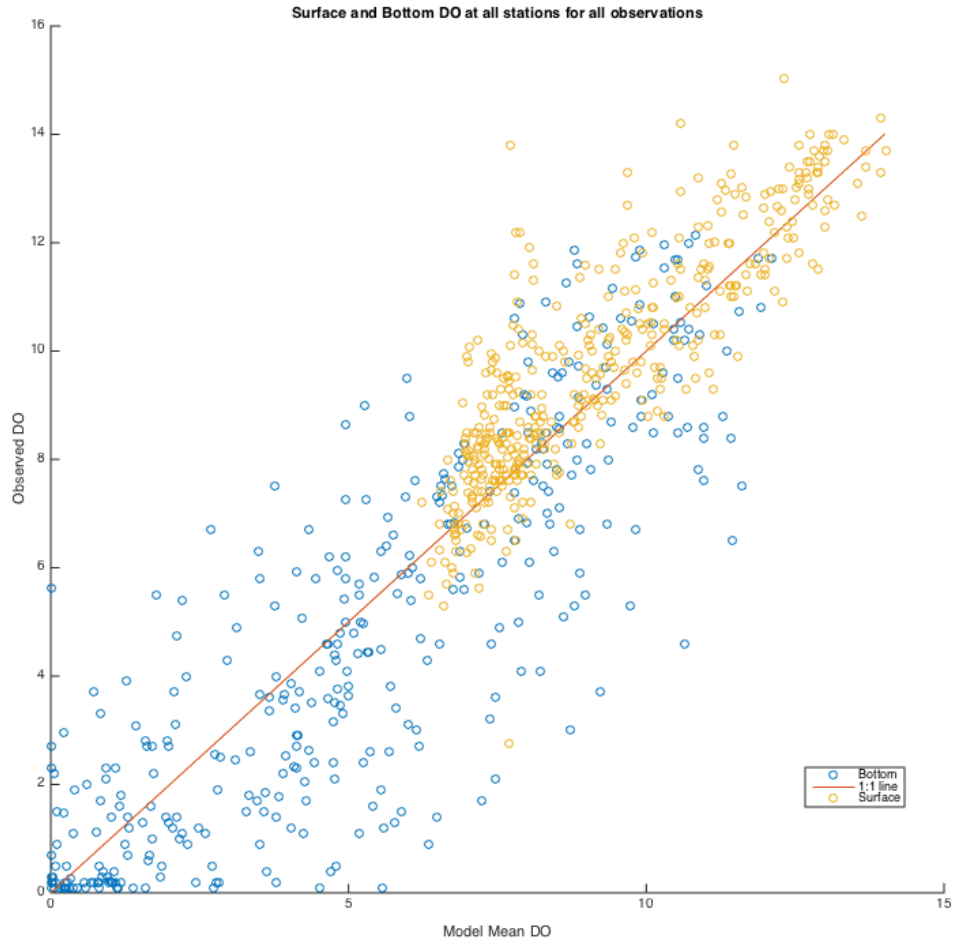

Figure R1. Scatter diagram illustrating the relationship between observed DO and the Model Mean DO for all stations at the surface (yellow) and bottom (blue.) The 1:1 line is shown in red.

Challenges associated with modeling low-oxygen waters in Chesapeake Bay: a multiple model comparison

I. D. Irby[1], M. A. M. Friedrichs[1], C. T. Friedrichs[1], A. J. Bever[2], R. R. Hood[3], L. W. J. Lanerolle[4,5], M. Li[6], L. Linker[7], M. E. Scully[8], K. Sellner[9], J. Shen[1], J. Testa[6], H. Wang[3], P. Wang[10], and M. Xia[11]

[1]Virginia Institute of Marine Science, College of William & Mary, P.O. Box 1346, Gloucester Point, VA 23062, USA

[2]Anchor QEA, LLC, 130 Battery Street, Suite 400, San Francisco, CA 94111, USA

[3]Horn Point Laboratory, University of Maryland Center for Environmental Science, P.O. Box 775, Cambridge, MD 21613, USA

[4]NOAA/NOS/OCS Coast Survey Development Laboratory, 1315 East–West Highway, Silver Spring, MD 20910, USA

[5]ERT Inc., 14401 Sweitzer Lane Suite 300, Laurel, MD 20707, USA

[6]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, P.O. Box 38, Solomons, MD 20688, USA

[7]US Environmental Protection Agency Chesapeake Bay Program Office, 410 Severn Avenue, Annapolis, MD 21403, USA

[8]Woods Hole Oceanographic Institution, Applied Ocean Physics and Engineering Department, Woods Hole, MA 02543, USA

[9]Chesapeake Research Consortium, 645 Contees Wharf Road, Edgewater, MD 21037, USA

[10]VIMS/Chesapeake Bay Program Office, 410 Severn Avenue, Annapolis, MD 21403, USA

[11]Department of Natural Sciences, University of Maryland Eastern Shore, MD, USA

Abstract

As three-dimensional (3-D) aquatic ecosystem models are used more frequently for operational water quality forecasts and ecological management decisions, it is important to understand the relative strengths and limitations of existing 3-D models of varying spatial resolution and biogeochemical complexity. To this end, two-year simulations of the Chesapeake Bay from eight hydrodynamic-oxygen models have been statistically compared to each other and to historical monitoring data. Results show that although models have difficulty resolving the variables typically thought to be the main drivers of dissolved oxygen variability (stratification, nutrients, and chlorophyll), all eight models have significant skill in reproducing the mean and seasonal variability of dissolved oxygen. In addition, models with constant net respiration rates independent of nutrient supply and temperature reproduced observed dissolved oxygen concentrations about as well as much more complex, nutrient-dependent biogeochemical models. This finding has significant ramifications for short-term hypoxia forecasts in the Chesapeake Bay, which may be possible with very simple oxygen parameterizations, in contrast to the more complex full biogeochemical models required for scenario-based forecasting. However, models have difficulty simulating correct density and oxygen mixed layer depths, which are important ecologically in terms of habitat compression. Observations indicate a much stronger correlation between the depths of the top of the pycnocline and oxycline than between their maximum vertical gradients, highlighting the importance of the mixing depth in defining the region of aerobic habitat in the Chesapeake Bay when low-oxygen bottom waters are present. Improvement in hypoxia simulations will thus depend more on the ability of models to reproduce the correct mean and variability of the depth of the physically driven surface mixed layer than the precise magnitude of the vertical density gradient.

1 Introduction

Since the middle of the last century, anthropogenic impacts have dramatically decreased water quality throughout the Chesapeake Bay (Boesch et al., 2001), one of the largest estuaries in North America. Land-use change along with the industrialization and urbanization of the Chesapeake Bay watershed have caused dramatic increases in nutrient inputs to the Bay (Kemp et al., 2005), spurring additional primary production and phytoplankton abundance (Harding and Perry, 1997). Because increased primary production leads to more organic matter throughout the water column that is eventually decomposed by bacteria, these increased nutrient inputs to the Bay have led to a corresponding decrease in dissolved oxygen (DO) concentrations (Hagy et al., 2004). Hypoxia, generally defined as the condition in which DO concentrations are below $2 mg L^{-1}$, usually initiates seasonally in the northern portion of the Bay and expands southward as summer develops (Kemp et al., 2009; Testa and Kemp, 2014). Although hypoxia in the Chesapeake Bay has likely existed since European colonization (Cooper and Brush, 1991, 1993), recent studies have highlighted an accelerated rise in the number and spatial extent of hypoxic, as well as anoxic (DO concentrations $<0.2 mg L^{-1}$), events in the Bay since the 1950's, primarily attributed to increased anthropogenic nutrient input (Hagy et al., 2004; Kemp et al., 2005; Gilbert et al., 2010). These impacts are likely to be exacerbated by future climate change (Najjar et al., 2010; Meire et al., 2013; Harding et al., 2015).

Interest in the ecological impacts of reduced DO concentrations has been elevated due to the observed proliferation of hypoxic events in the world's coastal oceans, creating vast dead zone areas that compress suitable habitat for many marine species (Diaz, 2001; Diaz and Rosenberg, 2008; Pierson et al., 2009). Low-DO waters can greatly impact the abundance and health of important ecological species, potentially resulting in suffocation and major kills of fish, crabs, and shellfish (Breitburg, 2002; Ekau et al., 2010; Levin et al., 2009). While the presence of DO concentrations $< 2 mg L^{-1}$ have been shown to decrease the abundance of fish larvae (Keister et al., 2000), some species can incur negative health impacts and modify their behavior at significantly higher DO

concentrations (Vaquer-Sunyer and Duarte, 2008). DO concentrations of ~ 4 mg L^{-1} have been found to compress demersal fish habitat as fish seek out more oxygenated waters (Buchheister et al., 2013). Zooplankton, a crucial food source for valuable species, have also been found to exhibit changes in distribution and predation when subject to large volumes of low-DO water, potentially leading to further impacts along the food chain (Breitburg et al., 1997; Pierson et al., 2009). Invertebrates have similarly been found to alter their behavior under low-DO conditions (Riedel et al., 2014). In the Chesapeake Bay, multiple regulated fish species, such as striped bass and American shad, require oxygen restoration targets as high as 5mgL^{-1} (USEPA, 2010). The greatest impact of low DO concentrations spatially will depend on the specific living resource; however, temporally, late spring to early fall is of most concern. As a result of the significant ecological importance of oxygen on living resources in the Bay, DO concentrations are used as a primary indicator in assessing water quality for Chesapeake Bay regulations (Keisman and Shenk, 2013).

Improving the health of the Chesapeake Bay has become a priority for the Environmental Protection Agency (EPA) along with the six states and Washington, DC that make up the Bay watershed (Fig. 1), and together they have committed to utilizing a suite of regulatory models to inform their management decisions (USEPA, 2010). The Chesapeake Bay Program (CBP), a regional partnership that has led and directed the restoration of the Chesapeake Bay since 1983, has undertaken an extensive modeling effort of the Bay (Cerco and Cole, 1993; Cerco et al., 2002; Cerco and Noel, 2004, 2013). This modeling system is being used by the CBP to estimate the aggregate effect of changes in management practices, including land use, atmospheric deposition, animal populations, and fertilizer and manure application. Recently, the modeling system has been used to conduct scenario simulations to assess management actions needed to achieve desired Bay water quality standards (USEPA, 2010). Ultimately this model was used to establish a regulatory set of total maximum daily loads of nutrients and sediment delivered from the watershed, with the goal of significantly improving water quality throughout the Bay (USEPA, 2010).

Many 3-D hydrodynamic-oxygen models of varying complexity stemming from the academic research community have also been used to simulate DO concentrations throughout the Chesapeake Bay (Scully, 2010, 2013; Hong and Shen, 2013; Feng et al., 2015; Testa et al., 2014; Li et al., 2015). Bever et al. (2013) specifically demonstrated that multiple models of varying complexity are able to generate skillful estimates of hypoxic volume in the Bay. Some of these models are being used in the Bay to simulate short-term and/or seasonal forecasts of DO conditions. Furthermore, some models are also being used to generate scenario forecasts, or projections, that assess the impact of changes in management practices on estuarine DO concentrations, in some cases taking into account the impacts of future changes in climate.

As ecosystem and water quality models are increasingly used for operational forecasts as well as scenario-based management decisions by the regulatory and academic research communities, it is important to understand the relative strengths and limitations of existing models of varying complexity. The ability to discern which variables must be most accurately simulated in order to adequately reproduce the temporal and spatial variability of Bay oxygen concentrations is a necessary prerequisite for fully understanding how volumes of low-DO water are initiated and sustained within water quality models. The utilization of multiple models can also inform projections by providing independent confidence bounds for management decisions. To those ends, the overarching goals of this research are to compare the relative skill of various three-dimensional (3-D) Chesapeake Bay models characterized by different levels of biogeochemical complexity and spatial resolution, to better understand factors limiting their ability to reproduce observed DO distributions, and to suggest approaches for the continued improvement of these models.

2 Methods

2.1 Participating Chesapeake Bay models

Eight 3-D models were evaluated in this study (Table 1), each of which includes hydrodynamic and DO components. Among the eight models, there are four different hydrodynamic base models. Models B, C, D, F, and G utilize the Regional Ocean

Modeling System (ROMS; Shchepetkin and McWilliams, 2005; Haidvogel et al., 2008) that employs a structured grid with sigma layers in the vertical dimension. Specifically, Models B, C, and F use a ROMS implementation developed for the Chesapeake Bay based on Xu et al. (2012; ChesROMS). Model D employs a ROMS implementation for the Chesapeake Bay based on Li et al. (2005), while Model G uses the ROMS-based Chesapeake Bay Operational Forecast System (CBOFS; Lanerolle et al., 2011). Models A, E, and H each use a different hydrodynamic base model: the Curvilinear Hydrodynamics in Three Dimensions model (CH3D; Cerco et al., 2010), the Finite-Volume Community Ocean Model (FVCOM; Jiang and Xia, 2015), and the Hydrodynamic-Eutrophication Model – Fluid Dynamics Code (EFDC; Park et al., 1995; Hong and Shen, 2012; Du and Shen, 2015), respectively. The only model that employs a non-sigma vertical grid is Model A and the only model utilizing an unstructured horizontal grid is Model E. While Model E contains 10 sigma vertical layers, all of the other sigma grids use 20 layers. All of the grids vary in terms of their horizontal resolution, with Models A and G utilizing the highest resolution horizontal grids.

These four hydrodynamic models are coupled to five different models used to simulate DO (Table 1). Models A, B, C, D, and E utilize full biogeochemical models that include as state variables various combinations of oxygen, phytoplankton, zooplankton, and multiple inorganic and organic nutrients. Specifically, Models A and E employ a version of the Integrated Compartment Model (ICM; Cerco et al., 2010; Jiang et al., 2015), Model B uses the Estuarine Carbon Biogeochemistry model (ECB; Feng et al., 2015), Model C uses the Biogeochemistry model (BGC; Brown et al., 2013), and Model D uses the Row-Column AESOP model (RCA; Testa et al., 2014). In terms of food web complexity the models vary considerably: Models B and C employ a single phytoplankton group whereas Model D uses two phytoplankton groups, Model E uses three, and Model A, the most complex of the participating models, uses five.

In contrast to the full biogeochemical models discussed above (Models A through E), Models F, G, and H represent oxygen dynamics as simply as possible and therefore do not utilize a full biogeochemical component. Rather, the models impose a biological oxygen consumption rate that is model-specific, but constant in both space and time. This

component is referred to as a constant-respiration model (CRM). In this model, DO is introduced to the estuary via the river and ocean boundaries and is set to saturation at the estuarine surface. This constant-respiration oxygen parameterization (Scully, 2010) is simplistic, yet has been shown to adequately represent Chesapeake Bay oxygen dynamics (Scully, 2010, 2013; Bever et al., 2013).

The major difference in forcing between the eight model implementations is that Models A and B use riverine input derived from watershed models, whereas Models C–H used the measured flow from United States Geological Survey gauging stations, extrapolated using various techniques. Model A utilized the CBP's regulatory watershed model (Shenk and Linker, 2013), while Model B utilized the Dynamic Land Ecosystem Model (Yang et al., 2014, 2015; Tian et al., 2015). At the open boundary with the Atlantic Ocean, Models B, C, D, F, G, and H utilize a sub-tidal elevation extrapolated from tidal stations on either side of the open boundary. Model E uses the TPXO tidal model, while Model A uses a mix of observational and model forcing (Cerco et al., 2010). While Model B utilizes wind forcing based on observations from the Thomas Point Light, Models C through H use wind estimates from the North American Regional Reanalysis (NARR).

The eight models used in this analysis have been developed for a variety of purposes. Model A is a governmental regulatory model developed by the CBP that has been extensively calibrated specifically to examine water quality issues in the Chesapeake Bay (Cerco and Cole, 1993; Cerco and Noel, 2004, 2013; Cerco et al., 2010) and has been used in the development of the 2010 Chesapeake Bay Total Maximum Daily Load (USEPA, 2010). The National Oceanic and Atmospheric Administration employs the hydrodynamic component of Model F for operational forecasts of a variety of physical estuarine parameters for the Chesapeake Bay (http://www.tidesandcurrents.noaa.gov/ofs/cbofs/cbofs.html). The other six models are academic models used in diverse research efforts focused on the Chesapeake Bay but not necessarily specifically on DO dynamics.

Finally, a ninth model is calculated as the mean of the results from the eight models

described above, and is referred to here as Model Mean, or Model M.

2.2 Available Chesapeake Bay observations

Model simulations were compared to cruise data from the CBP for 2004 and 2005 from 13 stations along the main stem of the Bay (Table 2, Fig. 2). The years 2004 and 2005 were selected to represent relatively wet and average years, respectively, and the 13 stations were chosen as they have been found to offer optimal estimates of Bay-wide hypoxic volume (Bever et al., 2013). Stations were sampled on up to 34 cruises over the two years (Table 2), generally twice a month from April to August and once a month for the remainder of the year. Observational data can be downloaded from the CBP Water Quality Database (http://www.chesapeakebay.net/data/downloads/cbp_water_quality_database_1984_present). Variables downloaded from the CBP website and used in this study were temperature, salinity, DO, nitrate + nitrite (hereafter abbreviated as "nitrate"), and chlorophyll *a* (hereafter abbreviated as "chlorophyll"). For most cruises, observations of temperature, salinity, and DO were made at roughly 1 m intervals throughout the water column, whereas observations of chlorophyll and nitrate were generally made only at the surface, bottom, and sometimes one or two mid-water column locations. For further information on available water quality observations, please see USEPA (2012). While these observations were publicly available for model assessment during calibration of all of the models, they represent a very small subset of the 30 years of EPA observations across over 100 Bay stations. The models compared here were calibrated based on access to the larger data set and for conditions in the Bay in general, not specifically for the 13 stations and two years considered here.

2.3 Calculation of stratification and mixed layer depth

Stratification of the density and oxygen fields was examined to identify the maximum gradient of the pycnocline and oxycline as well as the depth of the top of the pycnocline and oxycline. In open ocean studies, the depth of the top of stratification is commonly referred to as the mixed layer depth (MLD), although this term is less frequently used in the estuarine literature. As the research presented here distinguishes between the depths of the top of the pycnocline and that of the oxycline, these will be referred to respectively

as the density (ρ) mixed layer depth (MLD_ρ) and the oxygen mixed layer depth (MLD_O). Density was calculated via a classical density formula that is also utilized by the CBP for use in the Chesapeake Bay (Fofonoff and Millard, 1983; USEPA, 2004) and is a function of temperature and salinity.

The CBP defines the top and bottom of stratification in order to distinguish individual designated use areas for water quality management purposes (USEPA, 2004). They suggest that the top of the pycnocline be defined as the shallowest occurrence of a density gradient of 0.1 kg m^{-4} or greater as resolved by CBP profile observations, which are typically spaced at 0.5 to 2 m depth intervals. If density gradients throughout the water column are less than 0.1 kg m^{-4}, they define the water to be unstratified. The 0.1 kg m^{-4} threshold definition is designed to identify any initiation of stratification that may serve to cut off vertical mixing from a nearly perfectly well mixed layer.

While the CBP definition described above delineates between designated use boundaries according to density, our research focuses on the relationship between the pycnocline and oxycline, requiring an alternate definition that can be applied to both the density and oxygen distributions. In addition, the CBP definition often generates estimates for the depth of the top of the pycnocline that are too shallow compared to the maximum depth of surface mixing (Fig. 3). As a result, a percentage threshold criterion was developed that identifies the bottom of the reasonably well-mixed layer, rather than perfectly mixed layer, and is used in this analysis. The percentage threshold method defines a density or DO profile as being stratified if a change of 10 % of the difference between the profile's maximum and minimum values occurs within a single meter (Fig. 3). For example, if the maximum DO concentration throughout the water column on an individual sampling date is 10 mg L^{-1} and the minimum concentration is 1 mg L^{-1}, stratification is defined to be present if a difference of 0.9 mg L^{-1} is present within one meter. As recommended by the CBP, the uppermost meter of the water column is not considered (USEPA, 2004). The mixed layer depth is therefore defined as the shallowest level (below 1 m depth) where stratification is identified. The minimum stratification criterion utilized in this analysis requiring a profile to pass the 10% threshold also ensures that observations where very

little stratification exists do not bias the stratification results while also allowing for a single criterion to be used across multiple stratification variables.

2.4 Model skill metrics

Simulations of the Chesapeake Bay from the eight models described above were statistically compared to historical monitoring data using a variety of skill metrics including: root-mean squared difference (RMSD), bias, standard deviation, and correlation coefficient. These metrics are illustrated on Taylor and target diagrams (Taylor, 2001; Hofmann et al., 2008; Jolliff et al., 2009), which offer a compact way of assessing model skill by displaying a number of different skill metrics. Target diagrams illustrate the bias and total RMSD of model output, which Taylor diagrams do not. Taylor diagrams include quantitative information on the standard deviations and correlations between the model output and the observations, which target diagrams do not. Both diagrams, however, represent unbiased RMSD, sometimes called "centered-pattern RMSD". On target diagrams, a model symbol above the horizontal axis overestimates the mean of the observations and a model symbol to the right of the vertical axis overestimates the variability of the observations. (See Hofmann et al. (2008) and Jolliff et al. (2009) for a more detailed description of these diagrams.) On Taylor diagrams, a model symbol lying on the horizontal axis exactly correlates to the observations and a model symbol further from the origin than the observation symbol overestimates the standard deviation of the observations. (See Taylor et al. (2001) for a more detailed description of these diagrams).

Taylor and target diagrams presented here are normalized to the standard deviation of the observations, allowing multiple variables be represented on the same plot. This also conveniently allows the unit circle on a target diagram to represent the skill of a model defined as the mean of the observations. In effect, this means that if a model falls within the unit circle, it exhibits a skill that is greater than the skill obtained if one were to simply use the mean of the observations. The Taylor and target plots are either temporal (displaying model skill at a single station over the study period) or spatial (displaying model skill during a single month over the entire set of study stations). In addition,

summary diagrams are presented which combine both temporal (examining the seasonal changes at each individual station) and spatial (examining differences across the Bay during an individual month) variability.

Model skill was assessed using the hourly model output (daily for CH3D-ICM chlorophyll and nitrate) that was nearest in time to that of the observation and from the grid cell that encompassed the observation location. For months with two observations, each observation was individually matched to the model output and the skill statistics from those comparisons were averaged for that month. The native horizontal resolution and bathymetry of the individual model grids was preserved in the comparison so as not to bias the analysis through varying interpolation methodologies. For stratification variables, the models and observations were interpolated to a 1m vertical grid that extended only as deep as the individual models' bathymetry or deepest observation in order to preserve the differences in bathymetric grids while allowing for a direct comparison of the observations to the models. Model-data comparisons at the bottom of the water column were not necessarily based on the same depths, since in many cases the modeled bathymetry was shallower (or at times, deeper) than the deepest data point at a given station. In order to avoid issues with extrapolation and/or grid stretching, data at the bottom of the water column were always compared with model estimates from the deepest grid cell provided by each particular model. Model-data comparisons for stratification and mixed layer depths only included stations and times for which stratification was defined to exist in both the observed and simulated fields.

3 Results

An analysis of model skill of the combined temporal and spatial variability of DO at the surface and bottom of the water column, as well as at the observed MLD_O, indicates that all models, regardless of biogeochemical complexity or spatial resolution, exhibit a high degree of skill in reproducing observed DO (Fig. 4). Specifically, all models produce DO concentrations at the surface and bottom that have a normalized total RMSD less than one. The same is true for nearly all models for DO at the observed MLD_O. However, most models underestimate observed DO both at the surface and at the MLD_O (Fig. 4a).

The correlation between the observed and modeled DO is relatively constant with depth (Fig. 4b), though on average slightly higher at the bottom (0.85) than at the surface (0.80). Further, on average, the models simulate DO at the surface and bottom better than they do at the MLD_O. No statistical difference exists between the skill of models that utilize a full biogeochemical component and those that utilize the simple constant-respiration oxygen parameterization. Based on an analysis of variance (ANOVA) comparing the full biogeochemical models to the CRM models, the two model types do not perform differently in terms of their ability to reproduce the combined temporal and spatial variability of bottom DO as measured by total RMSD ($p = 0.48$). Overall, Model M (the mean of the 8 models) consistently performs better than any individual model across all depths examined (Fig. 4).

The monthly temporal variability of bottom DO at each station over the two years studied is resolved similarly well by all of the models (Fig. 5a), but the models have difficulty simulating spatial DO variability during each month (Fig. 5b). Due to the stations chosen for this analysis (Fig. 2), the spatial variability being examined here is essentially the north to south variability. Most models exhibit a latitudinal gradient with respect to their skill in reproducing the temporal variability of bottom DO, with models overestimating DO at the more northern stations (Fig. 5a). Some models differ in their ability to reproduce summer (May to September) DO concentrations and winter (October to April) DO concentrations (Fig. 5b). Models B, F, and G all distinctively overestimate mean DO in the summer compared to the winter. In contrast, Models A and C perform similarly well in both seasons (Fig. 5b). In addition, all three constant respiration models as well as Models D and E substantially underestimate DO at several stations in the winter.

All eight models generally resolve the pycnocline and oxycline with similar skill (Fig. 6). All models consistently underestimate the mean and standard deviation of the maximum strength of stratification within the pycnocline and oxycline, defined herein as the maximum vertical gradients of density and oxygen (Fig. 6a). All models, except for Model A (see Sect. 4.2), also underestimate the mixed layer depth, regardless of whether it is computed in terms of density or oxygen. (Note that these model symbols in Fig. 6a are located above the *y* axis despite this negative bias in MLD because the vertical

coordinate system is oriented upwards.) Thus the models are producing stratification that is both weaker than observed and higher (shallower) in the water column. The correlation coefficient for these metrics is low, ranging between 0.1–0.6, and indicates that all models are missing the majority of variability associated with the magnitude and location of the pycnocline and oxycline (Fig. 6b). However, there is slightly more consistency and better correlation coefficients among the models for the strength of stratification than the depth of the mixed layers.

All eight models are also characterized by similar skill in representing the temporal and spatial variability of density stratification and MLD_{ρ} (Fig. 7). There is a latitudinal difference in skill of the models in reproducing the magnitude of the pycnocline and MLD_{ρ}, with model skill generally lower at the northern stations (Fig. 7a). Contrary to the pattern shown for bottom DO (Fig. 5b), none of the models exhibit a significant seasonal pattern between summer and winter in reproducing spatial variability of $d\rho/dz$ or MLD_{ρ} (Fig. 7b). However, Model A differentiates itself from the rest of the models in its pattern of skill at reproducing the spatial and temporal variability of the MLD_{ρ} (see Sect. 4.2). Temporal and spatial patterns for oxycline stratification (dO/dz) and MLD_O closely match those of $d\rho/dz$ and MLD_{ρ} (not shown).

All eight models reproduce the variability of bottom DO better than the variables that are generally thought of as being the primary drivers of hypoxic conditions, including stratification (Fig. 6), salinity, chlorophyll and nitrate (Fig. 8, Table 3). However, all models reproduce patterns in temperature across the Bay and through time better than any of the other variables in this model comparison (Fig. 8). All eight models, as well as the Model Mean, are characterized by very low bias in modeled temperature, and correlation coefficients of approximately 0.99; this high skill results from the very strong and predictable seasonal temperature variability. Even though the five models with full biogeochemical components (Models A, B, C, D, and E) are characterized by large differences in their mechanistic approaches to modeling nitrate and chlorophyll, they produce similar total RMSDs for all of the variables examined at both the surface and at the bottom (Table 3).

The mean of the eight models (Model M) has a higher model skill (lower RMSD) than any individual model across nearly every variable examined (Table 3). In addition, for nearly all observations at all stations, the 95 % confidence interval of all model hindcasts encapsulates the observed bottom DO concentration (Fig. 9), even though any individual model may overestimate or underestimate observed DO. Models generally fall into greater agreement during the summer, when DO is low, and into lesser agreement in the winter when DO is replete. While this study does not allow for a true interannual comparison, it is interesting that at station CB4.1C whereas the model ensemble closely matches the timing of the drawdown of DO in the spring of 2004 (Fig. 9), it produces a summer rather than spring initiation of hypoxic conditions in 2005. In addition, the model ensemble produces a premature relaxing of hypoxic conditions for both years at this observation station.

In order to better understand the impact of stratification on DO concentrations throughout the water column, the relationship between the observed pycnocline strength and MLD_ρ were compared to the observed oxycline strength and MLD_O. Observations from 1998 to 2006 demonstrate that while there is not a strong correlation between the strengths of the pycnocline and oxycline, there is a very strong correlation between MLD_ρ and MLD_O (Fig. 10). Depending on the criteria used for defining the existence of stratification (see Sect. 2.3), the correlation of the pycnocline and oxycline strengths range between r^2 = 0.18 to 0.26 and the correlations of MLD_ρ and MLD_O range between $r^2 = 0.51$ to 0.82 (Table 4). Furthermore, correlation of the relationship between the MLD_ρ and MLD_O is stronger for more severe stratification (Table 4). The relationship between the two mixed layer depths is biased towards the MLD_O being slightly located deeper in the water column than the MLD_ρ. As the cut-off criteria for the existence of stratification becomes more stringent, the relationship becomes closer to 1:1.

4 Discussion

4.1 How does the skill of various hydrodynamically-based DO models compare?

– In examining the eight 3-D models in this study, there is not a statistical

difference between the ability of simple and complex models to simulate the mean and monthly variability of bottom DO; in addition, models with higher spatial resolution do not necessarily produce better estimates of DO.

Models currently simulating hypoxia throughout Chesapeake Bay compute oxygen concentrations in essentially two distinct ways: they either utilize a simple constant respiration model or a full biogeochemical model. In this study, the relative skill of both types of models is compared. Specifically, in examining results of the comparison between five biogeochemical models (A, B, C, D, and E) and three simplistic constant respiration models (F, G, and H), the two groups of models performed statistically similar in their skill of reproducing bottom DO concentrations (Fig. 3, Table 3). These results support those of Bever et al. (2013) who compared three constant respiration models with the CBP regulatory model (Model A) and similarly found that all four of the models were equally skillful in terms of reproducing the seasonal variability in bottom DO throughout the Bay in 2004 and 2005. Consistent with the results of Scully (2013), this result implies that the seasonal variability of DO in the Chesapeake Bay is primarily dependent on underlying hydrodynamic mechanisms which are nearly identical for all eight models, rather than on aspects related to the biogeochemical cycling which vary dramatically between models and in fact are constant in three of the eight models. It should be noted, however, that the two years studied here were relatively wet years and an analysis of dry years may offer different results.

Many previous studies have examined the costs and benefits of adding complexity to biogeochemical models. For example, increasing biogeochemical complexity has been found to improve skill in some biogeochemical data assimilative parameter optimization studies (Friedrichs et al., 2006, 2007; Lehmann et al., 2009; Bagniewski et al., 2011; Ward et al., 2013; Xiao and Friedrichs, 2014). The additional parameters associated with increased complexity generally provide more parameters that are available for additional tuning and subsequent improved model-data agreement. This is in contrast to the results of this analysis demonstrating that increased biogeochemical complexity does not necessarily improve model-data agreement. In this case the increase in model complexity has likely outpaced the ability of the researchers to fully tune the model to the available

observations. However, even past studies that have invoked formal parameter optimization methodologies such as genetic algorithms and variational adjoint methods (Friedrichs et al., 2007; Ward et al., 2010; Xiao and Friedrichs, 2014) have found that under certain conditions, adding too much complexity does not necessarily improve model skill and in fact can decrease model skill and portability, primarily due to artifacts resulting from overtuning. This mirrors findings from the larger ecosystem modeling community where the best-fit models are often those with intermediate complexity (Fulton et al., 2003).

In this study, horizontal grid resolution differed significantly between model implementations, with the most highly resolved grid (Model G) including more than nine times more grid cells than the lower resolution grids (Table 1). A certain degree of resolution is clearly required to successfully simulate dynamic processes, and a model with 8–10 km resolution will not be able to correctly simulate the hydrodynamic processes within the Bay (Feng et al., 2015). However, an increase in horizontal grid resolution from $\sim$ 1.8 to $\sim$ 0.6 km, which results in a run-time change of a factor of nine, or possibly of 27 if the time step is accordingly decreased by a factor of three, does not necessarily result in a significant improvement in simulation skill of either stratification or bottom oxygen. Although not shown here, additional sensitivity experiments with Model G revealed that doubling the vertical resolution of this model had no significant effect on the model's ability to resolve the depth of stratification or the maximum magnitude of stratification. Thus, when selecting the optimal model resolution for a simulation, it is critical to weigh the advantages of increased resolution with the increased time required for simulation. With a given level of computational resources, fewer sensitivity experiments can be conducted with a model using a more highly resolved grid.

Accurately simulating the observed spatial variability of DO (Fig. 4b) was a greater challenge than simulating the temporal variability of DO (Fig. 4a) for all eight models participating in this intercomparison. This is especially true in the winter months when the vast majority of the Bay is oxygen replete and the models have difficulty representing the observed variability from station to station. The majority of the models tend to slightly overestimate mean bottom DO in the summer whereas multiple models (e.g.,

Models D, E, F, and G) exhibit a strong negative bias during January and/or February of 2005, primarily at stations in the middle to southern portion of the Bay's deep channel. Interestingly, increased biological complexity and higher grid resolution do not completely resolve this issue, as this is true for models utilizing full biogeochemical models (Models D, E) as well as those using highly resolved model grids (Model G). This is likely due to the ephemeral nature of the biological divers of DO.

The strong performance of the constant respiration models implies that these models may be excellent candidates for providing short-term bottom oxygen forecasts. The high DO skill of the CRM models primarily results from the fact that seasonal variations in physical processes (primarily wind mixing and temperature) play a dominant role in controlling the seasonal cycle of oxygen (Scully, 2013). Because the underlying hydrodynamic models all use similar physical forcing, the constant respiration models are able to simulate the seasonal cycle of DO with similar skill as the more complex biogeochemical models. As a result, these simple models that are easier to tune and require less in the way of computational resources than full biogeochemical models, may be efficiently used to produce short-term (on the order of days) DO forecasts. On the contrary, the more complex full biogeochemical models will be necessary for scenario-based and long-term (on the order of months to years) forecasting which requires that models respond to prescribed changes in the biogeochemical environment, such as increased rates of nutrient loading due to changes in land use, land cover, and/or climate.

4.2 How does model skill of DO compare to that of the primary drivers of DO variability?

> – Overall, model DO skill is greater than that of the variables generally considered to drive DO variability, such as stratification, salinity, mixed layer depth, chlorophyll, and nitrate; only modeled temperature has higher skill than modeled DO.

Since dissolved oxygen concentrations in the Chesapeake Bay are controlled by physical processes (e.g., advection, wind mixing, heating/cooling, and stratification), as well as biological processes (e.g., photosynthesis and respiration), it is critical to understand the

skill of the models in terms of how well they reproduce the many factors influencing oxygen concentrations. As expected, the five models containing a specific biogeochemical model component had more difficulty simulating the observed chlorophyll and nitrate concentrations than the physical variables (temperature and salinity), both at the surface (Table 3) and the bottom (Fig. 8). Replicating the correct location, magnitude, and timing of phytoplankton blooms and nutrient cycling is a complex issue, and as a result, these features are generally not well simulated in the models. While the models generally simulate the total amount of chlorophyll adequately, it is more uniformly spatially distributed in the models rather than in patchy blooms as in nature, leading to the underestimation of chlorophyll variability across all models. Although all models produced a relatively high correlation between observed and modeled temperature and salinity (Fig. 8), the correlation coefficients for chlorophyll and nitrate were much lower. The correlations for observed vs. modeled DO was more similar to that of the physical variables (temperature, salinity) than the biological variables (chlorophyll and nitrate), highlighting that the seasonal variability in bottom DO is regulated more by physical than biological factors. This also explains the success of the constant respiration models, which by definition contain no biological variability yet reproduce DO variability nearly as well as the most complex biogeochemical models.

In this study, model skill was also considerably higher for bottom oxygen than it was for the vertical gradient of stratification and mixed layer depths (Figs. 6 and 8). The underestimation of the vertical gradient across all models is largely due to the numerical diffusion that characterizes all of these hydrodynamic models, but may also be partially due to an underestimation of the winds or a lack of diffuse freshwater input around the Bay. Even though the models all underestimated the strength of stratification (Figs. 4 and 6), modeled stratification in summer was strong enough to prevent mixing with the relatively well-oxygenated surface waters. This result suggests, somewhat surprisingly, that simulating the correct vertical gradient of stratification is not absolutely necessary for skillful bottom DO simulations. Models need only simulate *enough* stratification to effectively cut off vertical mixing in order to develop an isolated bottom layer that can then experience a draw down in oxygen via respiration. In addition, the models must also

correctly simulate the horizontal advection of oxygen (Scully, 2013; Li et al., 2015). The fact that bottom DO is simulated so well by the eight models analyzed here suggests that not only is the advection of oxygen well represented in the models, but also the strength of stratification, i.e., the maximum vertical gradients of density and oxygen, produced by these models is sufficient. Thus, although novel and somewhat unexpected, these results are not contradictory to previous studies demonstrating the importance stratification plays in initiating summer hypoxia in the Chesapeake Bay (Murphy et al., 2011).

Model skill in terms of reproducing observed mixed layer depths was likewise much lower than model skill of reproducing observed oxygen concentrations. All models, except Model A, produced mixed layer depths (MLD_O and MLD_ρ) that were generally too shallow in the water column (Fig. 6a). Note that Model A is a regulatory model that has been used for many years by the Chesapeake Bay Program, and has thus undergone more extensive calibration aimed at matching the mean salinity and oxygen characteristics of the Bay (Cerco and Cole, 1993). Furthermore, Model A employs a *z* grid that matches the bathymetry in trench areas better than the sigma grids used by the other models. Although Model A produced mixed layer depths that were generally in the correct location within the water column (Fig. 6a), they were too variable (Fig. 6b). This variability may partly be a result of the 1.5m *z* grid employed by Model A causing large jumps between vertical grid cells and hence resulting in overestimates of MLD variability. All other models use sigma grids typically with more highly resolved vertical resolution at the depth of maximum stratification.

The two variables for which the models have greatest skill are DO and temperature (Fig. 8). This is because oxygen variability is driven primarily by seasonal variability in physical processes such as solubility and wind mixing and to a lesser degree by variability in oxygen consumption (Scully, 2013). As a result, the models using a constant mean respiration rate produce as realistic hypoxia simulations as the biogeochemically complex models. Observations clearly show this strong seasonal variability in bottom DO (Fig. 11a) and, to a slightly lesser extent, clear seasonal variability in DO at the bottom of the bottom of the oxygen mixed layer (MLD_O; Fig.

11b). But a seasonal cycle is not manifested in the MLD_O itself (Fig. 11c). The lack of such a strong seasonal cycle in the observed mixed layer depths makes this a more difficult variable for the models to simulate. As a result, the models can relatively skillfully simulate the combined spatial and temporal variability of DO while simultaneously missing the MLD_O.

4.3 Why is it important for DO models to simulate the MLD_O correctly?

– Most of the aerobic habitat in the Bay during the summer is located above the MLD_O, thus it is critical for living resource managers to use models that accurately simulate this variable.

On average, the models miss the observed depth of the MLD_O by 3.4m, which equates to roughly a 60 % error in the modeled mixed layer depths. While the models have difficulty simulating the MLD_O throughout the entire year (Figs. 6 and 7b), the summer months are when the mismatch has the greatest potential to impact the available habitat for oxygen-dependent species. Each year during this time period low-oxygen waters occupy nearly the entire water column below the mixed layer. At Station CB4.1C, a representative mesohaline deep trough station, the contours of low-oxygen ($5 mgL^{-1}$) and hypoxic ($2 mgL^{-1}$) waters are located just below the MLD_O from late spring until late fall (Fig. 12). The severe depletion of oxygen below the mixed layer compresses the habitable space at this station to roughly 10 m (from a maximum of 32 m) during the annual low-oxygen event.

The impact of habitat compression can be substantial, as many Bay species require DO concentrations well above the traditional hypoxic threshold (USEPA, 2010). While not all of the main stem stations develop hypoxic water each year, most mesohaline stations experience a dramatic drawdown of oxygen to levels during the summer that effectively remove a large portion of the Bay from habitable space (Murphy et al., 2011; Schlenger et al., 2013). Studies have shown that some species modify their behavior based on the oxycline depth, which acts to constrict the habitable space in the water column (Prince and Goodyear, 2006; Pierson et al., 2009; Elliot et al., 2013). Since species can be

negatively impacted by low-DO concentrations as high as 5mgL^{-1} (Breitburg, 2002; Vaquer-Sunyer and Duarte, 2008; USEPA, 2010), the location of the oxycline is not only important for habitat compression in the summer months, but can also be important in the winter months when an occasional lack of vertical mixing can substantially decrease bottom DO concentrations. Furthermore, in order to accurately estimate hypoxic volume, models must correctly simulate the depth of the mixed layer, since the MLD_O closely follows the depth of the 2 mg L^{-1} contour.

4.4 How can DO simulations in the Bay be improved for management of water quality and living resources?

– To better simulate DO conditions and summer habitat compression due to low-DO water, simulations of the depth of the top of the pycnocline (MLD_ρ) must be improved.

Although the suite of models examined reproduce DO concentrations relatively well overall (Fig. 4), the models typically overestimate summer habitat compression by producing low DO concentrations too high in the water column (Fig. 6). Observations from the Chesapeake Bay Program show a strong correlation between the depths of the oxygen and density-defined mixed layers (Fig. 10b). The models analyzed here also clearly exhibit a close relationship between their skill in simulating the depths of the oxygen and density-defined mixed layers (Fig. 6). These strong relationships between the depths of the oxygen and density-defined mixed layers result from the fact that the pycnocline represents the physical barrier that leads to the development of the oxycline. Therefore, the inability of the models to accurately simulate habitat compression is an artifact of their lack of skill in simulating the depth of the density-defined mixed layer. In contrast, the strength of density stratification is not well correlated to the strength of oxygen stratification. This is because a relative wide range of intensities of density stratification is still sufficient to cut off vertical mixing, leading to the observed draw-down in bottom DO. Thus, even though all models underestimate the strength of the pycnocline, they still produce enough stratification to greatly reduce mixing. The results from this paper thus indicate that to further improve DO simulations and better estimate

summertime habitat compression, it is even more critical for models to accurately simulate the depth of the top of the pycnocline than to accurately simulate the absolute strength of the pycnocline.

4.5 What is the utility of the multi-model ensemble and Model Mean?

– The multi-model ensemble approach allows for the development of a Model Mean, which taken as its own model, is the most skilled model when examining the combined suite of variables analyzed in this study.

The model skill assessment presented here demonstrates that the average of all eight models, or five models in the case of chlorophyll and nitrate, does better than any individual model if looking across the suite of variables analyzed. This finding is similar to that of other studies that examined the value of the model mean from a multi-model ensemble (e.g., Gneiting and Raftery, 2005; Hagedorn et al., 2005). While the concept of using a multi-model ensemble has been most extensively employed by atmospheric, climatic, and global circulation modelers, such as the Intergovernmental Panel on Climate Change (e.g., Collins et al., 2013), the tool's utility for aquatic ecosystem modeling is gaining traction (Meier et al., 2012; Trolle et al., 2014; Janssen et al., 2015). As models are increasingly used in regulatory decisions regarding aquatic ecosystems, a cohort of similarly skilled models can be used to help inform a set of confidence bounds around an environmental forecast. Due to the restrictions placed on models used in regulatory actions, utilization of a multi-model ensemble may not be realistic for all environmental and resource managers; however, multiple models can be integrated into the decision-making process even when the ultimate decision must be based on a single model. For example, a confidence interval plot could help identify where regulatory model output might be acting out of sync with other skilled water quality models of the same system, thereby informing managers of the potential shortfalls associated with the regulatory model. Furthermore, if the models tend to be predicting similar DO concentrations, a cohort of models could enhance the confidence in regulatory decisions based on a single regulatory model (Friedrichs et al., 2012; Weller et al., 2013). Comparing multiple models can also help inform how to better improve models in the

future, as this study has aimed to do.

5 Conclusions

All models analyzed here exhibited a high degree of skill in simulating dissolved oxygen concentrations within the main stem of the Chesapeake Bay in two years corresponding to relatively wet and average years. Their high skill results from the fact that physical processes (e.g., solubility, wind-mixing, and advection) exert a first order influence on the seasonal cycle of oxygen. As a result, the models' ability to reproduce dissolved oxygen concentrations is independent of the complexity of the biogeochemical parameterizations: the simplest constant respiration models were found to reproduce observed oxygen concentrations as well as the most biologically complex models. Essentially, all models are equally capable of respiring most of the available oxygen in the lower water column during summer.

This study also suggests that for use as management tools for water quality and living resources, it is more critical for these models to adequately resolve the depth of the mixed layer than the absolute strength of stratification (as long as modeled stratification is strong enough to limit vertical mixing). This is critical because observations show that during warmer months, oxygen-depleted water fills the water column to where stratification limits further mixing, which effectively cuts off waters below the mixed layer for use by the majority of the Chesapeake Bay's most recognized and valued living resources. These results furthermore suggest that modelers should focus their efforts on improving the hydrodynamics of their models in an effort to improve simulations of mixed layer depth dynamics and variability.

These findings have significant ramifications for short-term bottom DO forecasts, which may be successful with very simple oxygen parameterizations embedded in hydrodynamic models. In contrast, scenario-based water quality forecasts are likely to benefit from more complex models, which must adequately reproduce the longer-term response of the oxygen field to changes in nutrient and organic matter loads. This study also helps to demonstrate how multiple community models from governmental agencies and academic institutions may be used together to provide a model mean and a set of

confidence bounds for regulatory model results that could be used to inform management decisions.

Acknowledgements. This work was supported by the NOAA IOOS program as part of the Coastal Ocean Modeling Testbed. We thank Yun Li and Younjoo Lee for help with the ROMS-RCA simulations used in this analysis and Ray Najjar for his insights and comments. This is VIMS contribution 3520 and UMCES contribution 5130.

References

Bagniewski, W., Fennel, K., Perry, M. J., and D'Asaro, E. A.: Optimizing models of the North Atlantic spring bloom using physical, chemical and bio-optical observations from a Lagrangian float, Biogeosciences, 8, 1291–1307, doi:10.5194/bg-8-1291-2011, 2011.

Bever, A. J., Friedrichs, M. A. M., Friedrichs, C. T., Scully, M. E., and Lanerolle, L. W.: Combining observations and numerical model results to improve estimates of hypoxic volume within the Chesapeake Bay, USA, J. Geophys. Res-Oceans, 118, 4924–4944, doi:10.1002/jgrc.20331, 2013.

Boesch, D. F., Brinsfield, R. B., and Magnien, R. E.: Chesapeake Bay Eutrophication: scientific understanding, ecosystem restoration, and challenges for agriculture, J. Environ. Qual., 30, 303–320, 2001.

Breitburg, D.: Effects of hypoxia, and the balance between hypoxia and enrichment, on coastal fishes and fisheries, Estuaries, 25, 767–781, 2002.

Breitburg, D. L., Loher, T., Pacey, C. A., and Gerstein, A.: Varying effects of low dissolved oxygen on trophic interactions in an estuarine food web, Ecol. Monogr., 67, 489–507, 1997.

Brown, C. W., Hood, R. R., Long, W., Jacobs, J., Ramers, D. L., Wazniak, C., Wiggert, J. D., Wood, R., and Xu, J.: Ecological forecasting in Chesapeake Bay: using a mechanistic-empirical modeling approach, J. Marine Syst., 125, 113–125, doi:10.1016/j.jmarsys.2012.12.007, 2013.

Buchheister, A., Bonzek, C. F., Gartland, J., and Latour, R. J.: Patterns and drivers of the demersal fish community of Chesapeake Bay, Mar. Ecol.-Prog. Ser., 481, 161–180, doi:10.3354/meps10253, 2013.

Cerco, C., Johnson, B., and Wang, H.: Tributary Refinements to the Chesapeake Bay Model, ERDC TR-02-4, US Army Engineer Research and Development Center, Vicksburg, MS, 2002.

Cerco, C., Kim, S.-C., and Noel, M.: 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, US Army Engineer Research and Development Center, Vicksburg, MS, 2010.

Cerco, C. F. and Cole, T.: Three-dimensional eutrophication model of Chesapeake Bay, J. Environ. Eng.-ASCE, 119, 1006–10025, 1993.

Cerco, C. F. and Noel, M. R.: The 2002 Chesapeake Bay Eutrophication Model, EPA 903-R-04-004, US Army Corps of Engineers, Waterways Experiment Stations, Vicksburg, MS, 2004.

Cerco, C. F. and Noel, M. R.: Twenty-one-year simulation of Chesapeake Bay water quality using the CE-QUAL-ICM eutrophication model, J. Am. Water Resour. As., 49, 1119–1133, doi:10.1111/jawr.12107, 2013.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., and Wehner, M.: Long-term climate change: projections, commitments and irreversibility, in: Climate Change

2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1029–1136, 2013.

Cooper, S. R. and Brush, G. S.: Long-term history of Chesapeake Bay anoxia, Science, 254, 992–996, 1991.

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

Diaz, R. J.: Overview of hypoxia around the world, J. Environ. Qual., 30, 275–281, 2001.

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

Du, J. and Shen, J.: Decoupling the influence of biological and physical processes on the dissolved oxygen in the Chesapeake Bay, J. Geophys. Res.-Oceans, 120, 78–93, doi:10.1002/2014JC010422, 2015.

Ekau, W., Auel, H., Pörtner, H.-O., and Gilbert, D.: Impacts of hypoxia on the structure and processes in pelagic communities (zooplankton, macro-invertebrates and fish), Biogeosciences, 7, 1669–1699, doi:10.5194/bg-7-1669-2010, 2010.

Elliott, D. T., Pierson, J. J., Roman, M. R.: Predicting the effects of coastal hypoxia on vital rates of the planktonic copepod *Acartia tonsa* dana, PLoS ONE, 8, e63987, doi:10.1371/journal.pone.0063987, 2013.

Feng, Y., Friedrichs, M. A. M., Wilkin, J., Tian, H., Yang, Q., Hofmann, E. E., Wiggert, J. D., and Hood, R. R.: Chesapeake Bay nitrogen fluxes derived from a land-estuarine ocean biogeochemical modeling system: model description, evaluation, and nitrogen budgets, J. Geophys. Res.-Biogeo., 120, 1666–1695, doi:10.1002/2015JG002931, 2015.

Fofonoff, N. P. and Millard, R. C.: Algorithms for Computations of Fundamental Properties of Seawater, UNESCO Technical Papers in Marine Science, 44, Paris, France, 53 pp., 1983.

Friedrichs, M., Sellner, K. G., and Johnston, M. A.: 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, 2012.

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

Friedrichs, M. A. M., Dusenberry, J., Anderson, L., Armstrong, R., Chai, F., Christian, J., Doney, S. C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D., Moore, K., Schartau, M., Sptiz, Y. H., and Wiggert, J.: Assessment of skill and portability in regional marine biogeochemical models: role of multiple phytoplankton groups, J. Geophys. Res., 112, C08001, doi:10.1029/2006JC003852, 2007.

Fulton, E. A., Smith, A. D. M., and Johnson, C. R.: Effect of complexity on marine ecosystem

models, Mar. Ecol.-Prog. Ser., 253, 1–16, 2003.

Gilbert, D., Rabalais, N. N., Díaz, R. J., and Zhang, J.: Evidence for greater oxygen decline rates in the coastal ocean than in the open ocean, Biogeosciences, 7, 2283–2296, doi:10.5194/bg-7-2283-2010, 2010.

Gneiting, T. and Raftery, A. E.: Weather forecasting with ensemble methods, Science, 310, 248–249, doi:10.1126/science.1115255, 2005.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, Tellus A, 57, 219–233, doi:10.1111/j.1600-0870.2005.00103.x, 2005.

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

Haidvogel, D. B., Arango, H., Budgell, W. P., Cornuelle, B. D., Curchitser, E., Di Lorenzo, E., Fennel, K., Geyer, W. R., Hermann, A. J., Lanerolle, L., Levin, J., McWilliams, J. C., Miller, A. J., Moore, A. M., Powell, T. M., Shchepetkin, A. F., Sherwood, C. R., Signell, R. P., Warner, J. C., and Wilkin, J.: Ocean forecasting in terrain-following coordinates: formulation and skill assessment of the Regional Ocean Modeling System, J. Comput. Phys., 227, 3595–3624, doi:10.1016/j.jcp.2007.06.016, 2008.

Harding Jr., L. W. and Perry, E. S.: Long-term increase of phytoplankton biomass in Chesapeake Bay, 1950–1994, Mar. Ecol.-Prog. Ser., 157, 39–52, 1997.

Harding Jr., L. W., Gallegos, C. L., Perry, E. S., Miller, W. D., Adolf, J. E., Mallonee, M. E., and Paerl, H. W.: Long-term trends of nutrients and phytoplankton in Chesapeake Bay, Estuaries Coasts, doi:10.1007/s12237-015-0023-7, 2015.

Hofmann, E. E., Druon, J., Fennel, K., Friedrichs, M., Haidvogel, D., Lee, C., Mannino, A., McClain, C., Najjar, R., O'Reilly, J., Pollard, D., Previdi, M., Seitzinger, S., Siewert, J., Signorini, S., and Wilkin, J.: Eastern US continental shelf carbon budget: integrating models, data assimilation, and analysis, Oceanography, 21, 86–104, doi:10.5670/oceanog.2008.70, 2008.

Hong, B. and Shen, J.: Responses of estuarine salinity and transport processes to potential future sea-level rise in the Chesapeake Bay, Estuar. Coast. Shelf S., 104–105, 33–45, doi:10.1016/j.ecss.2012.03.014, 2012.

Hong, B. and Shen, J.: Linking dynamics of transport timescale and variations of hypoxia in the Chesapeake Bay, J. Geophys. Res.-Oceans, 118, 6017–6029, doi:10.1002/2013JC008859, 2013.

Janssen, A. B. G., Arhonditsis, G. B., Beusen, A., Bolding, K., Bruce, L., Bruggeman, J., Couture, R.-M., Downing, A. S., Elliott, J. A., Frassl, M. A., Gal, G., Gerla, D. J., Hipsey, M. R., Hu, F., Ives, S. C., Janse, J. J., Jeppsen, E., Johnk, K. D., Kneis, D., Kong, X., Kuiper, J. J., Lehmann, M. K., Lemmen, C., Ozkundakci, D., Petzoldt, T., Rinke, K., Robson, B. J., Sachse, R., Schep, S. A., Schmid, M., Scholten, H., Teurlincx, S., Trolle, D., Troost, T. A., Van Dam, A. A., Van Gerven, L. P. A., Weijerman, M., Wells, S. A., and Mooij, W. M.: Exploring, exploiting and evolving diversity of aquatic ecosystem models: a community perspective, Aquat. Ecol., 49, 513–

548, doi:10.1007/s10452-015-9544-1, 2015.

Jiang, L. and Xia, M.: Dynamics of the Chesapeake Bay outflow plume: Realistic plume simulations and its seasonal and interannual variability, J. Geophys. Res.-Oceans, 121, doi:10.1002/2015JC011191, 2016.

Jiang, L., Xia, M., Ludsin, S. A., Rutherford, E. S., Mason, D. M., Jarrin, J. M., and Pangle, K. L.: Biophysical modeling assessment of the drivers for plankton dynamics in dressenid-colonized western Lake Erie, Ecol. Model., 308, 18–33, 2015.

Jolliff, J. K., Kindle, J. C., Schulman, I., Penta, B., Friedrichs, M. A. M., Helber, R., and Arnone, R. A.: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, J. Marine Syst., 76, 64–82, doi:10.1016/j.jmarsys.2008.05.014, 2009.

Keisman, J. and Shenk, G.: Total maximum daily load criteria assessment using monitoring and modeling data, J. Am. Water Resour. As., 49, 1134–1149, doi:10.1111/jawr.12111, 2013.

Keister, J. E., Houde, E. D., and Breitburg, D. L.: Effects of bottom-layer hypoxia on abundances and depth distributions of organisms in Patuxent River, Chesapeake Bay, Mar. Ecol.-Prog. Ser., 205, 43–59, 2000.

Kemp, W. M., Boyton, W. R., Adolf, J. E., Boesch, D. F., Boicourt, W. C., Brush, G., Cornwell, J. C., Fisher, T. R., Gilbert, P. M., Hagy, J. D., Harding, L. W., Houde, E. D., Kimmel, D. G., Miller, W. D., Newell, R. I. E., Roman, M. R., Smith, E. M., and Stevenson, J. C.: Eutrophication of Chesapeake Bay: historical trends and ecological interactions, Mar. Ecol.-Prog. Ser., 303, 1–29, 2005.

Kemp, W. M., Testa, J. M., Conley, D. J., Gilbert, D., and Hagy, J. D.: Temporal responses of coastal hypoxia to nutrient loading and physical controls, Biogeosciences, 6, 2985–3008, doi:10.5194/bg-6-2985-2009, 2009.

Lanerolle, L. W., Patchen, R. C., and Aikman, F.: The Second Generation Chesapeake Bay Operational Forecast System (CBOFS2): Model Development and Skill Assessment, TR- NOS-CS-29, US Department of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, Office of Coast Survey, Coast Survey Development Laboratory, Silver Spring, MD, 2011.

Lehmann, M. K., Fennel, K., and He, R.: Statistical validation of a 3-D bio-physical model of the western North Atlantic, Biogeosciences, 6, 1961–1974, doi:10.5194/bg-6-1961-2009, 2009.

Levin, L. A., Ekau, W., Gooday, A. J., Jorissen, F., Middelburg, J. J., Naqvi, S. W. A., Neira, C., Rabalais, N. N., and Zhang, J.: Effects of natural and human-induced hypoxia on coastal benthos, Biogeosciences, 6, 2063–2098, doi:10.5194/bg-6-2063-2009, 2009.

Li, M., Zhong, L., and Boicourt, W. C.: Simulations of Chesapeake Bay estuary: sensitivity to turbulence mixing parameterizations and comparison with observations, J. Geophys. Res., 110, C12004, doi:10.1029/2004JC002585, 2005.

Li, Y., Li, M., and Kemp, W. M.: A budget analysis of bottom-water dissolved oxygen in Chesapeake Bay, Estuar. Coast., 38, 2132–2148, doi:10.1007/s12237-014-9928-9, 2015.

Meier, H. E. M., Andersson, H. C., Arheimer, B., Blenckner, T., Chubarenko, B., Donnelly, C., Eilola, K., Gustafsson, B. G., Hansson, A., Havenhand, J., Hoglund, A., Kuznetsov, I., MacKenzie, B. R., Muller-Karulis, B., Neumann, T., Niiranen, S., Piwowarczyk, J., Raudsepp, U., Reckermann, M., Ruoho-Airola, T., Savchuk, O. P., Schenk, F., Schimanke, A., Vali, G., Weslawski, J.-M., and Zorita, E.: Comparing reconstructed past variations and future projections of the Baltic Sea ecosystem – first results from multi-model ensemble simulations, Environ. Res. Lett., 7, 034005, doi:10.1088/1748-9326/7/3/034005, 2012.

Meire, L., Soetaert, K. E. R., and Meysman, F. J. R.: Impact of global change on coastal oxygen dynamics and risk of hypoxia, Biogeosciences, 10, 2633–2653, doi:10.5194/bg-10-2633- 2013, 2013.

Murphy, R. R., Kemp, W. M., Ball, W. P.: Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading, Estuar. Coast., 34, 1293–1309, doi:10.1007/s12237-011- 9413-7, 2011.

Najjar, R. G., Pyke, C. R., Adams, M. B., Breitburg, D., Hershner, C., Kemp, M., Howarth, R., Mulholland, M. R., Paolisso, M., Secor, D., Sellner, K., Wardrop, D., and Wood, R.: Potential climate-change impacts on the Chesapeake Bay, Estuar. Coast. Shelf S., 86, 1–20, doi:10.1016/j.ecss.2009.09.026, 2010.

Park, K., Kuo, A. Y., Shen, J., and Hamrick, J. M.: A three-dimensional Hydrodynamic Eutrophication Model (HEM-3D): description of water quality and sediment process submodels, in: Applied Marine Science and Ocean Engineering, Special Report, Virginia Institute of Marine Science, Gloucester Point, VA, 327, 1995.

Pierson, J. J., Roman, M. R., Kimmel, D. G., Boicourt, W. C., and Zhang, X. S.: Quantifying changes in the vertical distribution of mesozooplankton in response to hypoxic bottom waters, J. Exp. Mar. Biol. Ecol., 381, 74–79, 2009.

Prince, E. D. and Goodyear, C. P.: Hypoxia-based habitat compression of tropical pelagic fishes, Fish. Oceanogr., 15, 451–464, doi:10.1111/j.1365-2419.2005.00393.x, 2006.

Riedel, B., Pados, T., Pretterebner, K., Schiemer, L., Steckbauer, A., Haselmair, A., Zuschin, M., and Stachowitsch, M.: Effect of hypoxia and anoxia on invertebrate behaviour: ecological perspectives from species to community level, Biogeosciences, 11, 1491–1518, doi:10.5194/bg-11-1491-2014, 2014.

Schlenger, A. J., North, E. W., Schlag, Z., Li, Y., Secor, D. H., Smith, K. A., and Niklitschek, E. J.: Modeling the influence of hypoxia on the potential habitat of Atlantic sturgeon *Acipenser oxyrinchus*: a comparison of two methods, Mar. Ecol.-Prog. Ser., 483, 257–272, doi:10.3354/meps10248, 2013.

Scully, M. E.: The importance of climate variability to wind-driven modulation of hypoxia in Chesapeake Bay, J. Phys. Oceanogr., 40, 1435–1440, doi:10.1175/2010JPO4321.1, 2010.

Scully, M. E.: Physical controls on hypoxia in Chesapeake Bay: a numerical modeling study, J. Geophys. Res.-Oceans, 118, 1239–1256, doi:10.1002/jgrc.20138, 2013.

Shenk, G. W. and Linker, L. C.: Development and application of the 2010 Chesapeake Bay

watershed total maximum daily load model, J. Am. Water Resour. As., 49, 1–15, doi:10.1111/jawr.12109, 2013.

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

Taylor, K. E.: Summarizing multiple aspects of models performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001.

Testa, J. M. and Kemp, W. M.: Spatial and temporal patterns of winter–spring oxygen depletion in Chesapeake Bay bottom water, Estuar. Coast., 37, 1432–1448, doi:10.1007/s12237-014- 9775-8, 2014.

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

Tian, H., Yang, Q., Najjar, R., Ren, W., Friedrichs, M. A. M., Hopkinson, C. S., and Pan, S.: Anthropogenic and climatic influences on carbon fluxes from eastern North America to the Atlantic Ocean: a process-based modeling study, J. Geophys. Res.-Biogeo., 120, 752–772, doi:10.1002/2014JG002760, 2015.

Trolle, D., Elliott, J. A., Mooij, W. M., Janse, J. H., Bolding, K., Hamilton, D. P., and Jeppsen, E.: Advancing projections of phytoplankton responses to climate change through ensemble modeling, Environ. Modell. Softw., 61, 371–379, doi:10.1016/j.envsoft.2014.01.032, 2014.

USEPA: Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity, and Chlorophyll *a* for the Chesapeake Bay and its Tidal Tributaries – 2004 Addendum, EPA 903-R-03-002, US Environmental Protection Agency, USEPA Region III Chesapeake Bay Program Office, Annapolis, MD, 2004.

USEPA: Chesapeake Bay Total Maximum Daily Load for Nitrogen, Phosphorus, and Sediment, US Environmental Protection Agency, US Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, MD, 2010.

USEPA: Guide to Using Chesapeake Bay Program Water Quality Monitoring Data, EPA 903-R-12-001, US Environmental Protection Agency, US Environmental Protection Agency Chesapeake Bay Program, Annapolis, MD, 2012.

Vaquer-Sunyer, R. and Duarte, C. M.: Thresholds of hypoxia for marine biodiversity, P. Natl. Acad. Sci. USA, 105, 15452–15457, doi:10.1073/pnas.0803833195, 2008.

Ward, B. A., Friedrichs, M. A. M., Anderson, T. R., and Oschlies, A.: Parameter optimization techniques and the problem of underdetermination in marine biogeochemical models, J. Marine Syst., 81, 34–43, doi:10.1016/j.jmarsys.2009.12.005, 2010.

Ward, B. A., Schartau, M., Oschlies, A., Martin, A. P., Follows, M. J., and Anderson, T. R.: When is a biogeochemical model too complex? Objective model reduction and selection for

North Atlantic time-series sites, Prog. Oceanogr., 116, 49–65, doi:10.1016/j.pocean.2013.06.002, 2013.

Weller, D., Benham, B., Friedrichs, M., Gardner, N., Hood, R., Najjar, R., Paolisso, M., Pasquale, P., Sellner, K., and Shenk, G.: Multiple Models for Management in the Chesapeake Bay, Chesapeake Bay Program Scientific and Technical Advisory Committee Workshop Report, No. 14-004, 25–26 February 2013.

Xiao, Y. and Friedrichs, M. A. M.: 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, 3015–3030, doi:10.5194/bg-11- 3015-2014, 2014.

Xu, J., Long, W., Wiggert, J. D., Lanerolle, L. W. J., Brown, C. W., Murtugudde, R., and Hood, R. R.: Climate forcing and salinity variability in Chesapeake Bay, USA, Estuar. Coast. Shelf S., 35, 237–261, doi:10.1007/s12237-011-9423-5, 2012.

Yang, A., Tian, H., Friedrichs, M. A. M., Hopkinson, C. S, Lu, C., and Najjar, R. G.: Increased nitrogen export from eastern North America to the Atlantic Ocean due to climatic and anthropogenic changes during 1901–2008, J. Geophys. Res.-Biogeo., 120, 1046-1068, doi:10.1002/2014JG002763, 2015.

Yang, Q., Tian, H., Friedrichs, M. A. M., Liu, M., Li, X., and Yang, J.: Hydrological responses to climate and land-use changes along the North American east coast: a 110-year historical reconstruction, J. Am. Water Resour. As., 51, 47–67, doi:10.1111/jawr.12232, 2015.

Table 1. Model characteristics.

Model	A	B	C	D	E	F	G	H
Hydrodynamic model-DO model	CH3D-ICM	ChesROMS-ECB	ChesROMS-BGC	ROMS-RCA	FVCOM-ICM	ChesROMS-CRM	CBOFS-CRM	EFDC-CRM
Grid structure	Structured	Structured	Structured	Structured	Unstructured	Structured	Structured	Structured
Average wet-cell resolution	1 km	1.8 km	1.8 km	1.89 km	1.26 km	1.8 km	0.565 km	1.2 km
Vertical grid	1.52 m	20 sigma	20 sigma	20 sigma	10 sigma	20 sigma	20 sigma	20 sigma
River forcing	CBP Watershed Model	DLEM Watershed Model	USGS Data	USGS Data	USGS Data	USGS Data	USGS Data	USGS Data
Sub-tidal elevation at open boundary	Multiple efforts	Lewes, DE to Duck, NC	Lewes, DE to Duck, NC	Wachapreague, VA to Duck, NC	TPXO Tidal Model	Lewes, DE to Duck, NC	Ocean City, MD to Duck, NC	Lewes, DE to Duck, NC
Wind forcing	Multiple efforts	Thomas Point Light	NARR	NARR	NARR	NARR	NARR & NDBC buoys	NARR
Other atmospheric forcing	Multiple efforts	NARR	NARR	NARR	NARR	NARR	NARR	Norfolk & Baltimore Airports
Biogeochemical complexity	High; 5 phytoplk. groups	High; 1 phytoplk. group	High; 1 phytoplk. group	High; 2 phytoplk. groups	High; 3 phytoplk. groups	Low; constant respiration	Low; constant respiration	Low; constant respiration
Model citation	Cerco et al., 2010	Feng et al., 2015	Brown et al., 2013	Testa et al., 2014	Jiang and Xia, 2015	Scully, 2013	Lanerolle et al., 2011	Du and Shen, 2015

Table 2. Characteristics of observation stations (from USEPA, 2012).

Station	Latitude	Longitude	Station Depth	# of Cruises
CB3.2	39.1634 N	76.3063 W	12.1 m	34
CB3.3C	38.9951 N	76.3597 W	24.3 m	34
CB4.1C	38.8251 N	76.3997 W	32.3 m	34
CB4.2C	38.6448 N	76.4177 W	27.2 m	34
CB4.3C	38.5565 N	76.4347 W	26.9 m	34
CB4.4	38.4132 N	76.3430 W	30.3 m	34
CB5.1	38.3185 N	76.2930 W	34.1 m	34
CB5.2	38.13678N	76.2280 W	30.6 m	34
CB5.4	37.8001 N	76.1747 W	31.1 m	26
CB6.2	37.4868 N	76.1563 W	10.5 m	30
CB6.4	37.2365 N	76.2080 W	10.2 m	29
CB7.1	37.6835 N	75.9897 W	20.9 m	27
LE2.3	38.0215 N	76.3477 W	20.1 m	34

Table 3. Mean and standard deviation (STD) of observations and total normalized RMSD for each model.

	Mean ± STD of Obs	Normalized RMSD								
		A	B	C	D	E	F	G	H	M
Surface Temp. (ºC)	17.44±8.82	0.13	0.13	0.12	0.09	0.13	0.13	0.16	0.19	0.10
Bottom Temp. (ºC)	15.75±8.02	0.24	0.35	0.35	0.23	0.22	0.35	0.17	0.19	0.23
Surface Salinity (PSU)	10.92±4.32	0.37	0.62	0.53	0.36	0.46	0.61	0.57	0.41	0.35
Bottom Salinity (PSU)	18.17±3.14	0.72	0.85	0.73	1.55	1.28	0.78	1.03	0.97	0.75
Max. dρ/dz (kg m^{-4})	~1.64±1.15	1.03	1.09	1.07	1.09	1.25	1.01	1.23	1.02	N/A
MLDρ (m)	~5.32±3.99	1.01	1.13	1.11	1.41	1.39	1.12	1.38	1.13	N/A
Surface DO (mg L^{-1})	9.74±2.15	0.67	0.58	0.89	0.80	1.00	0.63	0.64	0.69	0.57
DO at MLD_O (mg L^{-1})	~8.44±2.53	0.54	0.57	0.74	0.93	0.83	0.81	0.95	1.09	0.62
Bottom DO (mg L^{-1})	4.42±3.61	0.51	0.59	0.81	0.61	0.54	0.46	0.61	0.60	0.46
Max. dDO/dz (mg L^{-1} m^{-1})	~1.81±1.12	1.19	1.21	1.34	1.09	1.35	1.12	1.23	1.19	N/A
MLDo (m)	~6.62±4.01	1.24	1.01	1.10	1.33	1.33	1.05	1.30	1.29	N/A
Surface Chl *a* (mg m^{-3})	11.19±9.04	0.92	1.22	1.60	1.23	0.89	N/A	N/A	N/A	1.16
Bottom Chl *a* (mg m^{-3})	9.02±11.52	0.87	1.10	1.07	1.05	1.01	N/A	N/A	N/A	0.90
Surface Nitrate (mmolN m^{-3})	0.32±0.33	0.61	0.79	1.03	0.61	0.52	N/A	N/A	N/A	0.79
Bottom Nitrate (mmolN m^{-3})	0.12±0.13	1.08	1.38	1.38	0.92	1.46	N/A	N/A	N/A	0.85

Table 4. Pycnocline and oxycline correlation statistics (all correlations have p-values << 0.01).

Stratification Threshold Percentage	Max dρ/dz vs. Max dO/dz	MLDρ vs. MLD_O	Profiles with Stratification
10%	0.18	0.51	1613
15%	0.22	0.59	1303
20%	0.22	0.70	916
25%	0.26	0.82	575

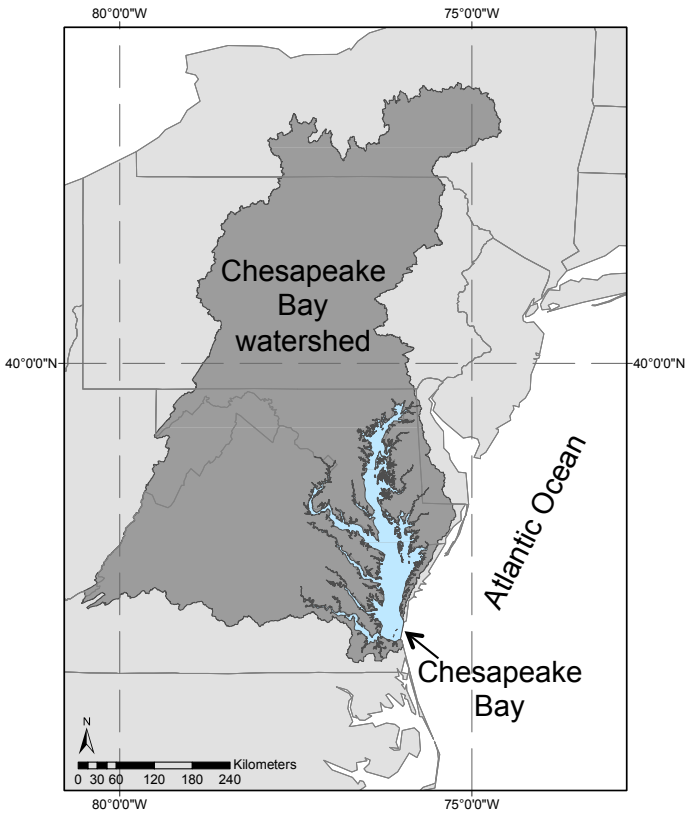

Figure 1. Map of the Chesapeake Bay and its watershed.

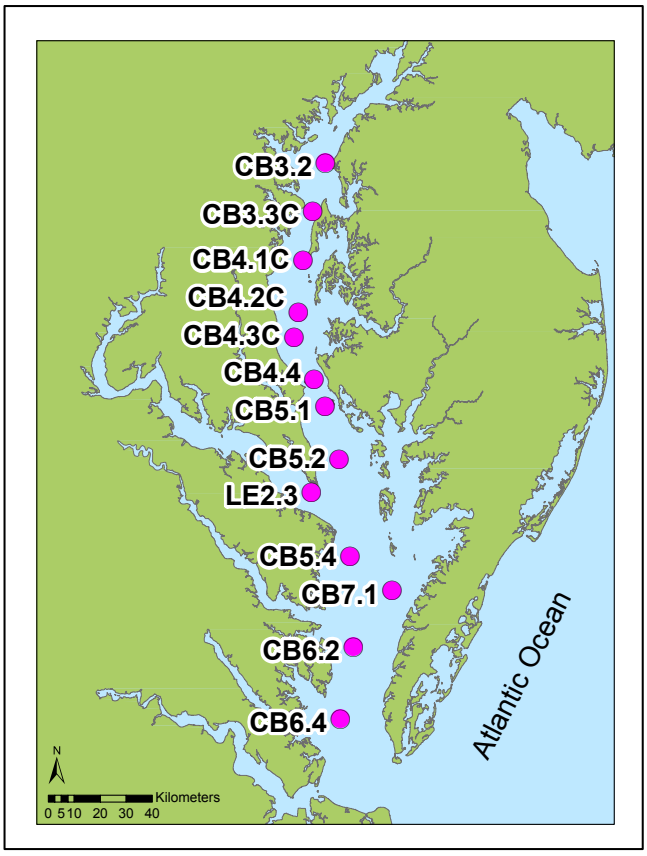

Figure 2. Location of the CBP Water Quality Monitoring stations used in this study.

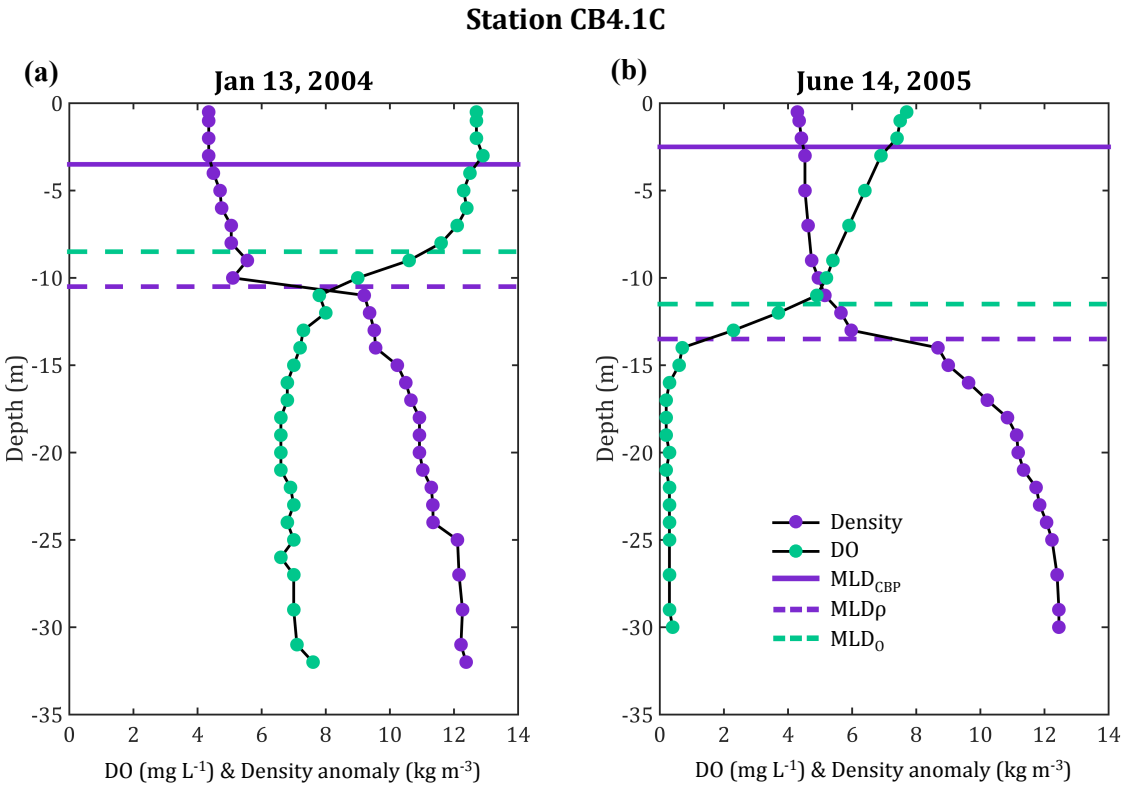

Figure 3. Density and dissolved oxygen profiles for a mid-Bay station (CB4.1C) on (a) January 13, 2004 and (b) June 14, 2005, comparing the 0.1 kg m^{-4} stratification definition used by the CBP (MLD_{CBP}) with the 10% threshold definitions used here for density (MLD_{ρ}) and oxygen (MLD_{O}).

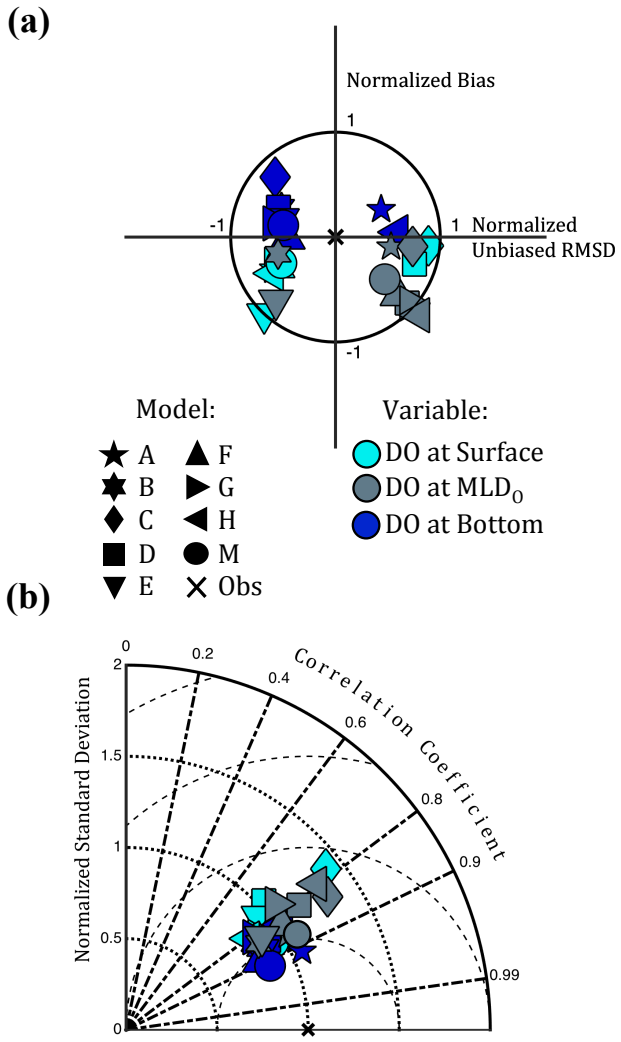

Figure 4. Normalized summary (a) target and (b) Taylor diagrams illustrating model skill of dissolved oxygen at the surface, MLD_O, and bottom for 13 Chesapeake Bay stations in 2004-2005. The "x" represents the skill of a model that perfectly reproduces the observations. The dotted, dashed-dot, and dashed lines on the Taylor diagram represent lines of constant standard deviation, correlation coefficient, and unbiased RMSD, respectively.

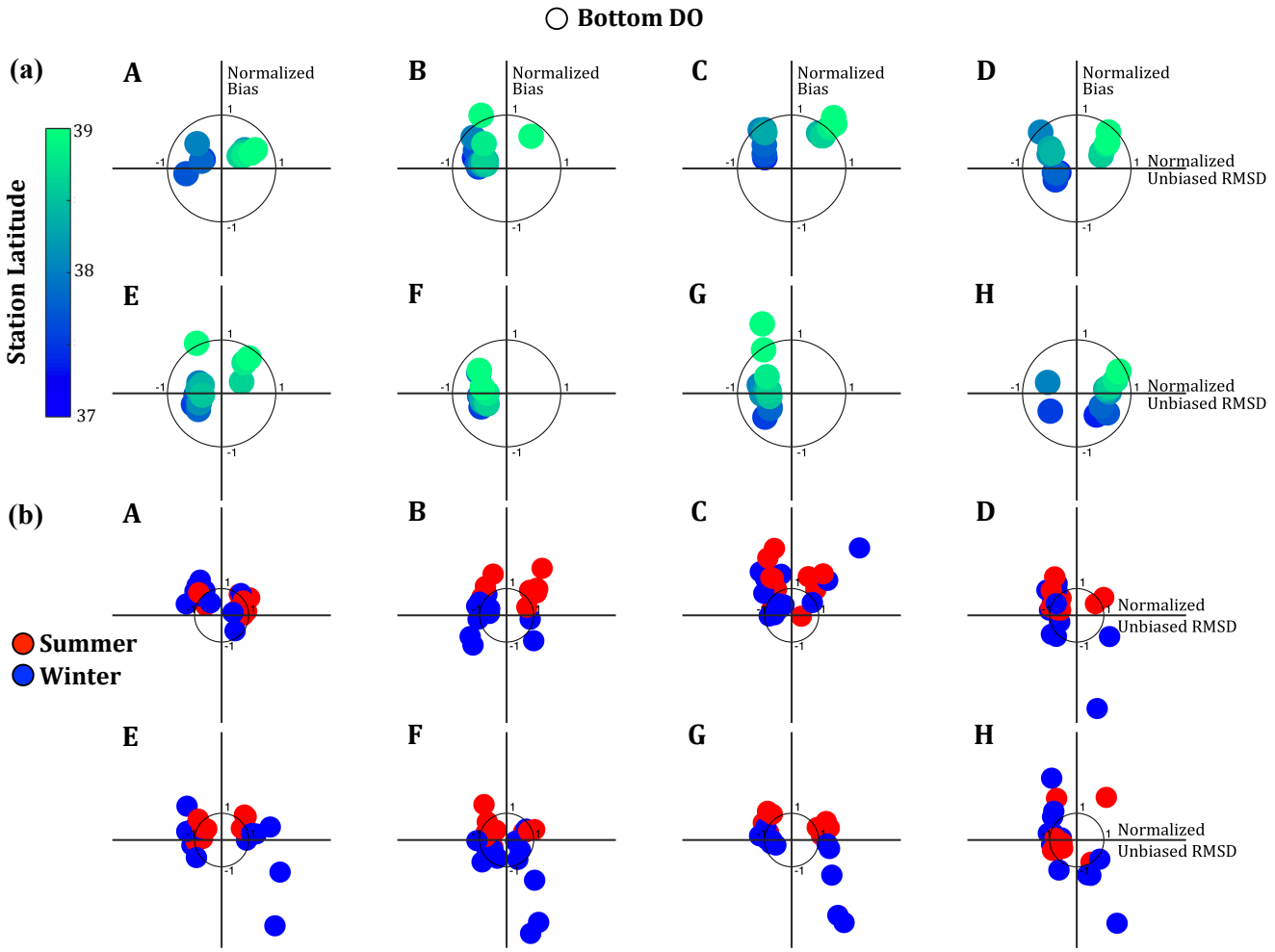

Figure 5. Normalized target diagrams for Models A-H demonstrating the (a) temporal and (b) spatial skill in resolving the variability of bottom dissolved oxygen concentrations. In (a) the individual dots represent the 13 stations along the main stem of the Chesapeake Bay. In (b) the dots represent the 24 months of 2004-2005 and are delineated by color: red = summer (May-September) and blue = winter (October-April).

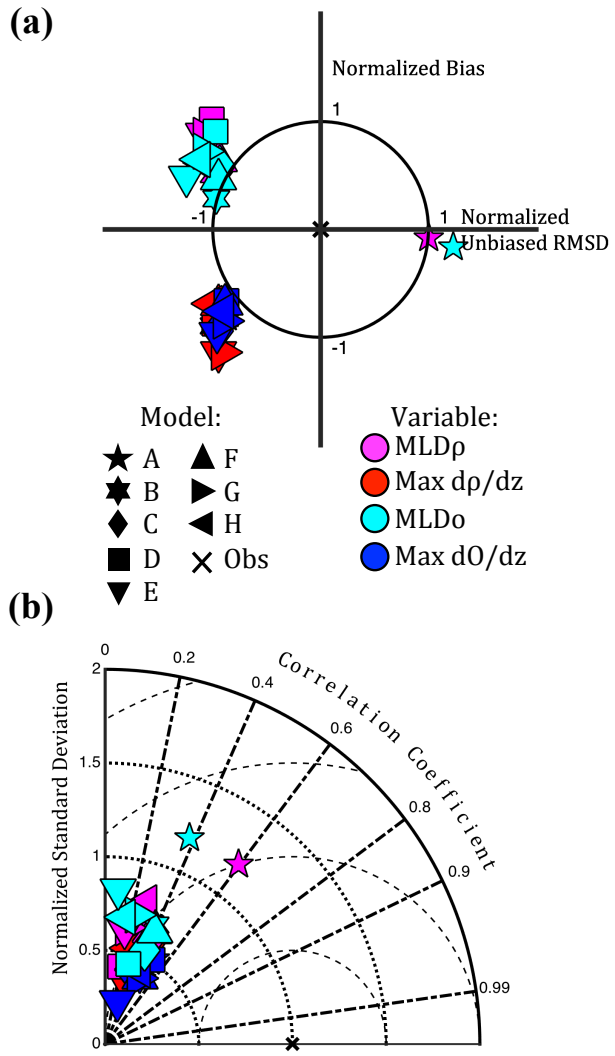

Figure 6. Normalized summary (a) target and (b) Taylor diagram illustrating model skill of MLD_{ρ} and MLD_{O}, max dρ/dz, and max dO/dz at 13 Chesapeake Bay stations for 2004-2005. The "x" represents the skill of a model that perfectly reproduces the observations. Since RMSD of stratification is only computed at stations where both the observations and model exhibit stratification, the Model Mean is not calculable for these variables.

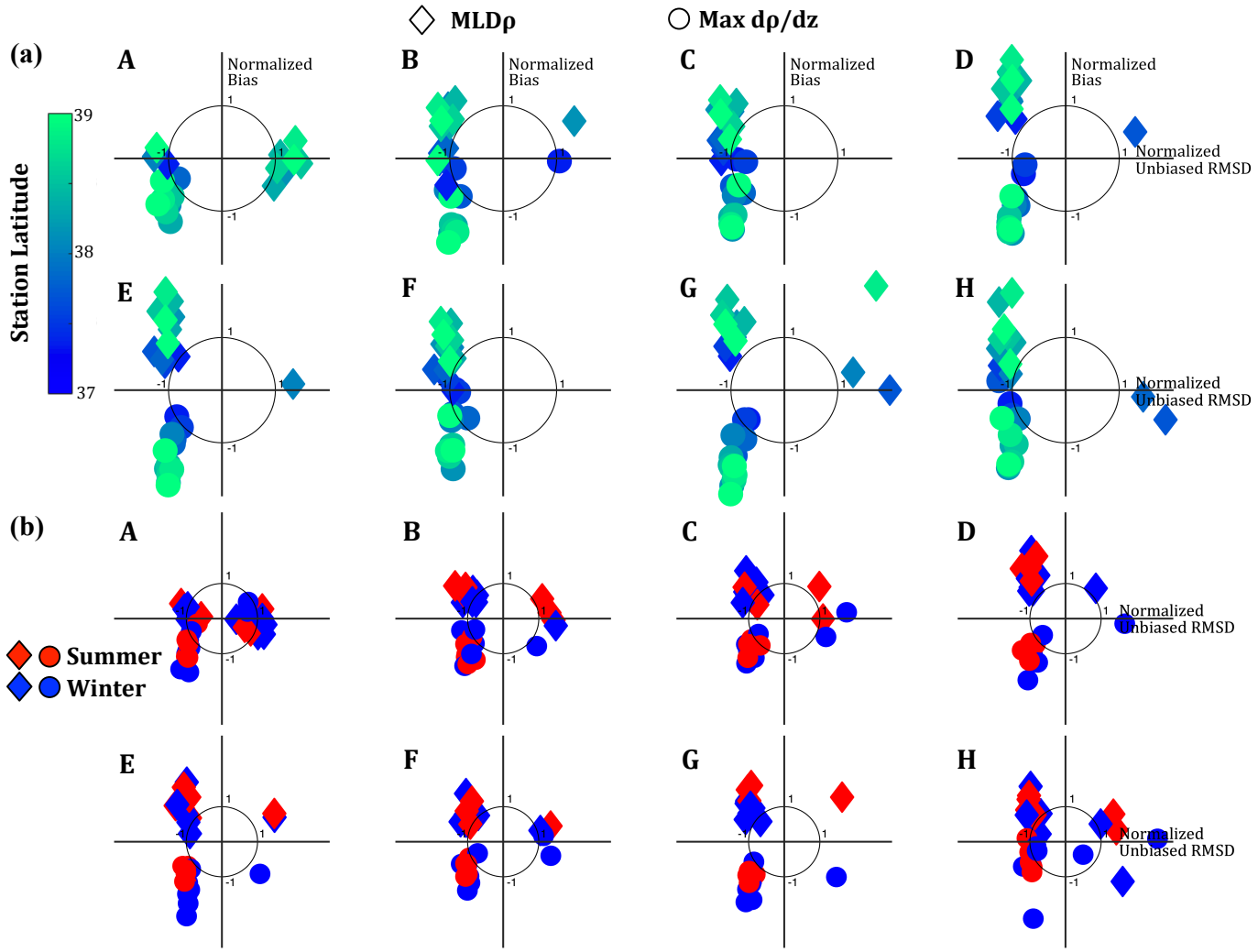

Figure 7. Normalized target diagrams for Models A-H demonstrating the (a) temporal and (b) spatial skill in resolving the variability of the strength of density stratification (circles) and the depth of pycnocline initiation (diamonds). In (a) the individual dots represent the 13 stations along the main stem of the Chesapeake Bay. In (b) the dots represent the 24 months of 2004-2005 and are delineated by color: red = summer (May-September) and blue = winter (October-April).

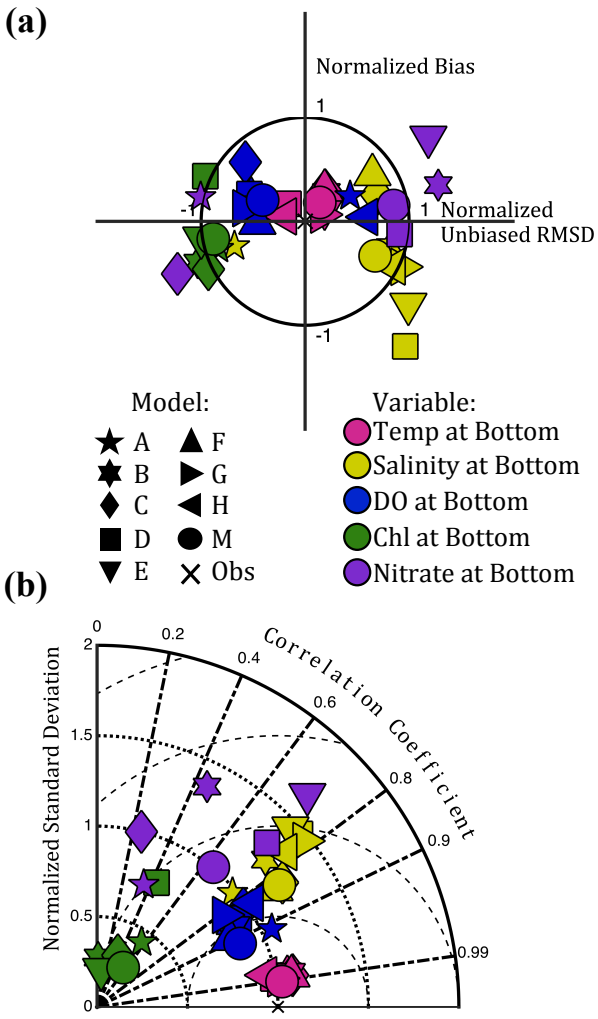

Figure 8. Normalized summary (a) target and (b) Taylor diagram illustrating model skill of bottom temperature, salinity, chlorophyll, nitrate, and dissolved oxygen at 13 Chesapeake Bay stations for 2004-2005. The "x" represents the skill of a model that perfectly reproduces the observations.

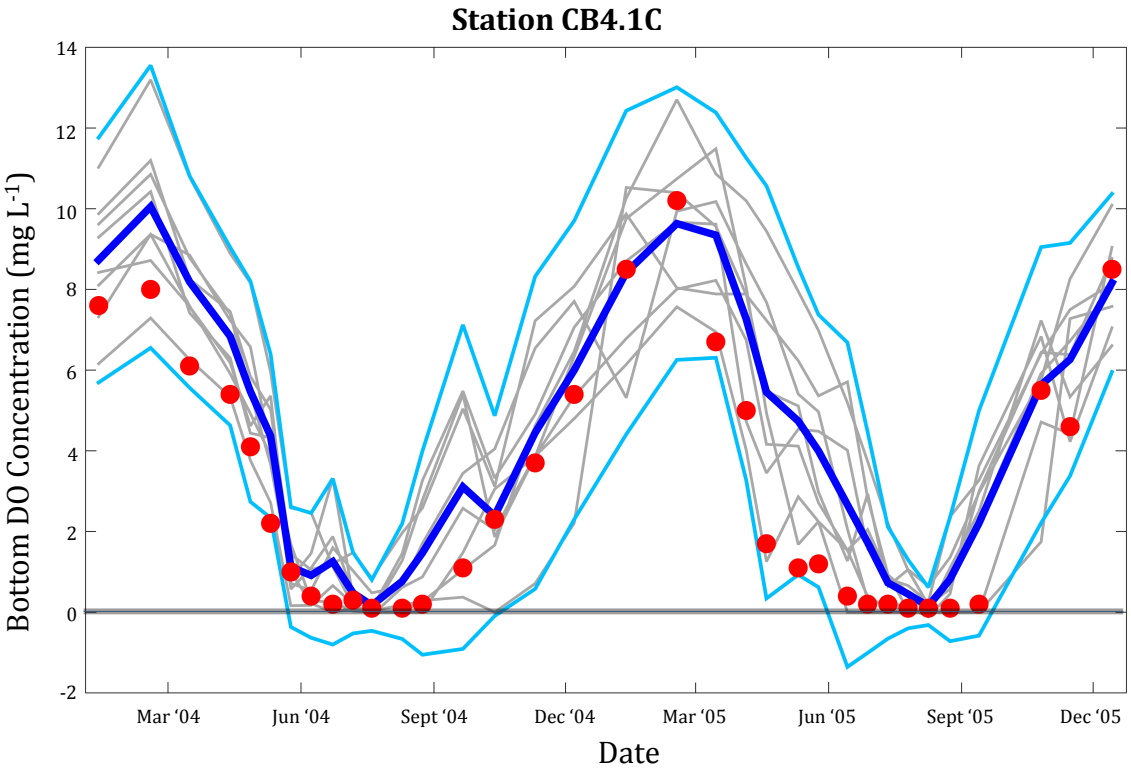

Figure 9. Time series of bottom dissolved concentrations for station CB4.1C. Red dots represent the 34 observations made during 2004-2005. Grey lines are the individual model simulations. The dark blue line represents the model mean while the cyan line represents the 95% confidence interval of the model simulations.

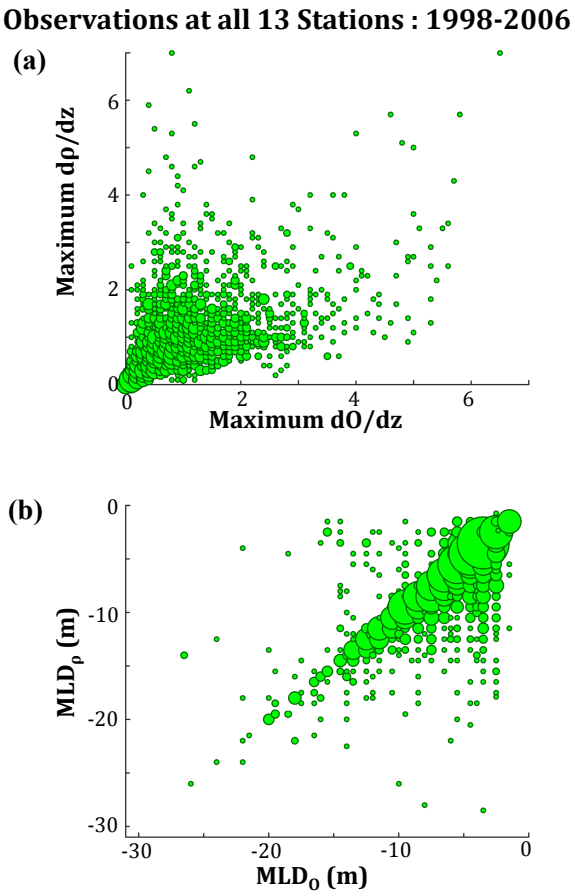

Figure 10. Scatter plots comparing observations of (a) the strengths of stratification of the pycnocline and oxycline and (b) the oxygen- and density-defined mixed layer depths. Size of the circles is proportional to the number of observations. Observations are from 1998-2006 at the 13 Chesapeake Bay stations shown in Figure 2.

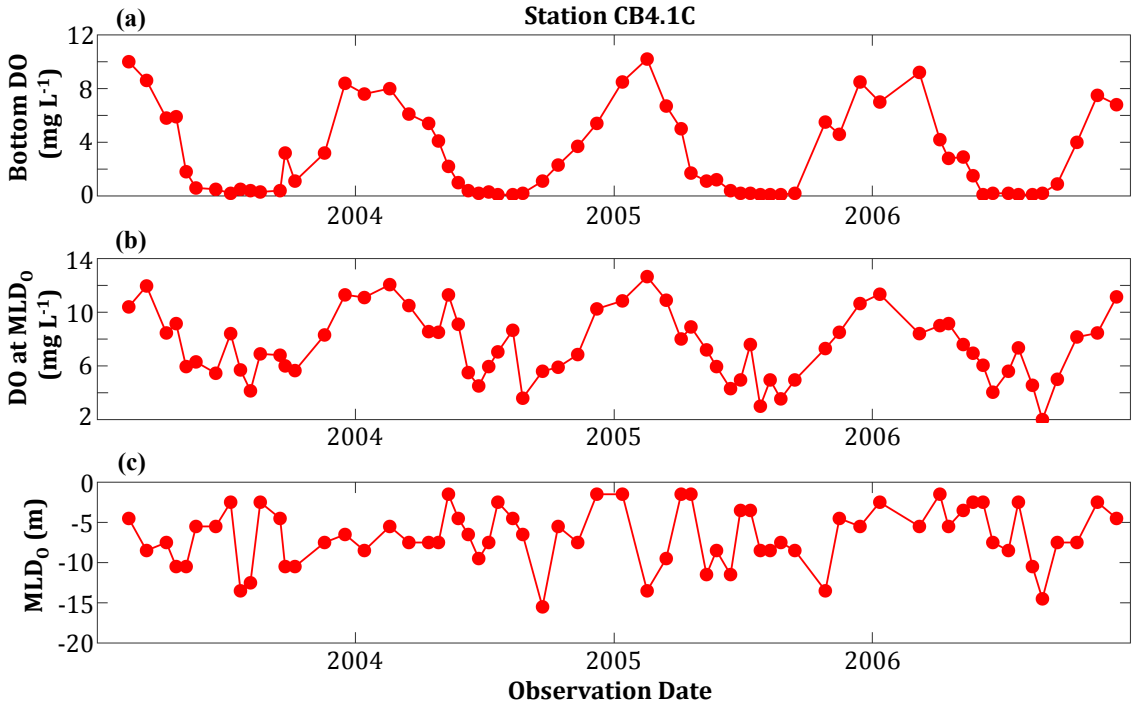

Figure 11. Time series of observations at Station CB4.1C from 2003 – 2006 for (a) bottom dissolved oxygen, (b) dissolved oxygen at the MLD_O, and (c) MLD_O.

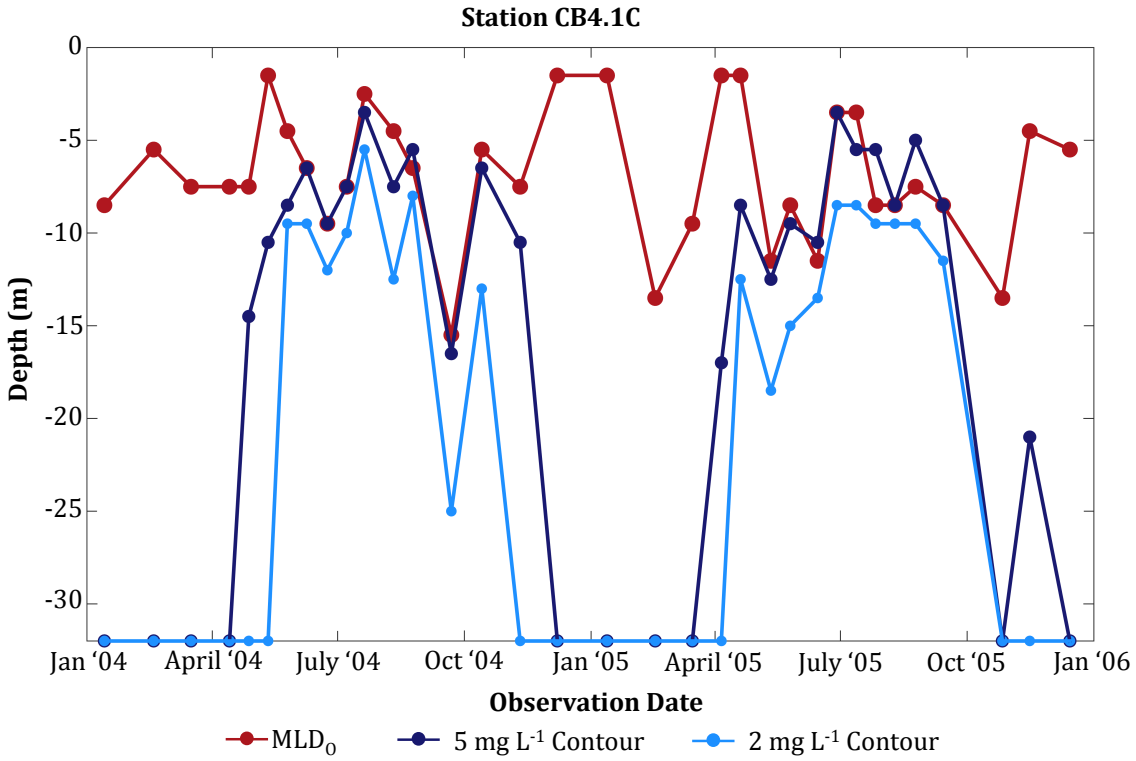

Figure 12. Time series of observations of dissolved oxygen and MLD_O contours at Station CB4.1C for 2004 and 2005.