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."

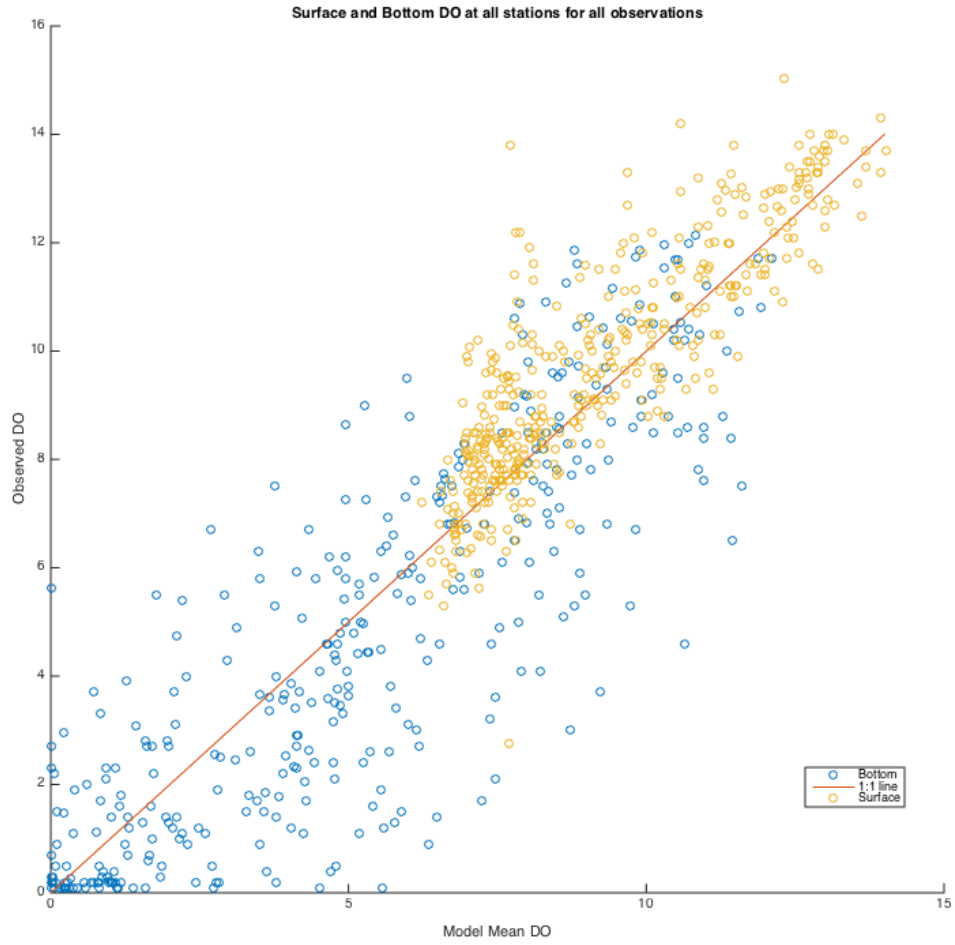

Figure R1. Scatter diagram illustrating the relationship between observed DO and the Model Mean DO for all stations at the surface (yellow) and bottom (blue.) The 1:1 line is shown in red.