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.