Responses to comments by Referee #2

Comment on bg-2022-4
Anonymous Referee #2

In their manuscript, Ou et al develop a novel statistical model to forecast/hindcast the size of the hypoxic area in the northern Gulf of Mexico. They use the model to test the feasibility of using HYCOM output and atmospheric data (reanalysis and forecast) to forecast the size of the hypoxic zone. The manuscript is well written and the statistical model seems to be able to retrieve the hypoxic area simulated with the ROMS model (part I paper). I am not familiar with the GLM/GAM statistical techniques and hopefully another reviewer can verify this part of the methodology. My overall assessment is that some improvements are required before the manuscript should be considered for publication. There are a few points that I think are important and would like raise below. Other, more specific comments are listed afterwards.

1) In its current form, the manuscript is mostly methodological and therefore I don't know if BG is the best fit for it. This could be solved with some improvements. For instance, the Discussion section presents an example of how to use the forecast model. This is a really interesting approach but it feels like a quick addition to justify the model development, that will be "further improved" in the future. A proper set of "forecasts" that are tested against observations would make a much more compelling case for the model's ability to forecast hypoxia. 1985-2021 mid-summer observations are available for this test; I believe that HYCOM and atmospheric forcing data are available in recent years to carry out this analysis. The forecast input data come with (high?) uncertainty and it would be interesting to know the effect on the hypoxia forecast (compared with the reanalysis input).

Authors' Response: We did compare the predicted hypoxic area by the ensemble model with the Shelfwide observations during the composition of this manuscript. We will expand our Discussion section by 1) comparing observed, predicted, ROM hindcast, and NOAA forecast hypoxic area; 2) assessing the sensitivity of input data to the predicted hypoxic area. We will also perform a long prediction according to HyCOM's availability.

2) My second point is a follow up from above. The manuscript relies exclusively on models. This is fine as a methodological paper but not if the authors aim at improving the current (seasonal) hypoxia forecasts and providing a tool for managers. For instance, it is assumed that the ROMS hindcast is a true representation of LaTex hypoxia. This is obviously not the case (as with any models) and it seems important to include observations in the manuscript to see how/if the forecasts drift away from the observations as we go from ROMS to GLM/GAM to HYCOM. Also note some of the reviewers comments on the Part I paper referenced here. Furthermore, the model provides a highly temporally resolved forecast, but it is not clear to me if, as a forecast, it does better than the seasonal forecast models (cited in the Introduction) that are, for some of them, spatially and temporally resolved. Some comparison with those (available annually through NOAA, e.g. https://www.noaa.gov/news-release/noaa-forecasts-averagesized-dead-zone-for-gulf-of-mexico) would strengthen the manuscript.

Authors' Response: We agree with this comment that observation should be involved in evaluating the model performance. The ROMS hindcast does capture the annual variability of the observed hypoxic area as in the Part I paper. It is also very interesting to compare our prediction with other forecast products like those by NOAA. We will add more discussion in the revision.

To the best of our knowledge, no forecast model is capable of providing a hypoxia forecast map. We would like to resolve it in a future study.

3) The part that needs significant improvement is the Discussion, which is not really available in the current version of the manuscript. Rather, the Discussion section presents an attempt at a "real" forecast using HYCOM. This could be moved to the Results section and a real Discussion section should be provided. What does this new technique brings to the knowledge of LaTex hypoxia? How does it compare with earlier models? How is this useful to managers? What are the caveats and limitations? what are the future developments? How is this technique portable to other systems? All of those are legitimate points that should be discussed.

Authors' Response: We will improve the Discussion section by adding related discussions as recommended.

Specific comments

L36/53: Those are seasonal forecast and cannot include the wind since it is not predictable at this time scale.

Response: The model by Turner et al. (2006) was not built using wind information. Instead, it was built based on May nitrogen load and observations of the hypoxic area under fair weather. Therefore, their model can provide a robust annual prediction when no strong wind is present.

L56: Stratification is included indirectly in the statistical models

Response: The previous models as mentioned in the previous paragraph did consider stratification-relevant predictors like wind speed, water transport, and riverine nutrient loads (usually correlated to river discharges). However, water stratification is affected by all these variables and can be represented directly by the water density profiles. We plan to change this statement to "The effects of water column stratification are considered only implicitly by the associated wind speeds, water transport, and riverine nutrient loads (usually highly correlated to river discharges), although stratification is documented as a crucial factor in regulating HA variability."

L58: They are not pseudo forecast, they forecast the mid summer hypoxic area (well in advance). Therefore, they are seasonal forecasts, which is different from the short-term forecasts provided by HYCOM.

Response: We will use a seasonal forecast model instead of a pseudo forecast model here.

L58-59: "fail whenever winds are strong in summers": Note that some of these models provide information on the effect of the wind on the forecast

Response: The wind input they used is from historical records (e.g., Katin et al., 2021 and Laurent and Fennel, 2019). We will revise this sentence.

L76: FYI (related to the main comment above), looking at the comparison between ROMS and observed mid-summer hypoxic area in Part I manuscript, the r-square is 0.58.

L79: could you define the geographical limits that you use for the LaTex shelf? That would be helpful to have a sense of your comparisons as it is not clear if you use the same area as the mid-summer sampling cruises to calculate the hypoxic zone.

Response: The LaTex Shelf we mentioned here covers the region shown in the below figure (Fig. 1). We may need to add a map of the LaTex Shelf in the supplementary materials. The shelf region we chose is larger than the coverage of the Shelfwide cruise survey because 1) The Shelfwide cruise surveys do not usually extend to the west of 93.5W, however, in some summers (like 2017), the survey did reach the west of 94W along the Texas coastal where hypoxia was reported. 2) According to the SEAMAP summer Groundfish Survey (Fig. 2 below), hypoxic bottom water can be found to the west of 94W along with coastal Texas.

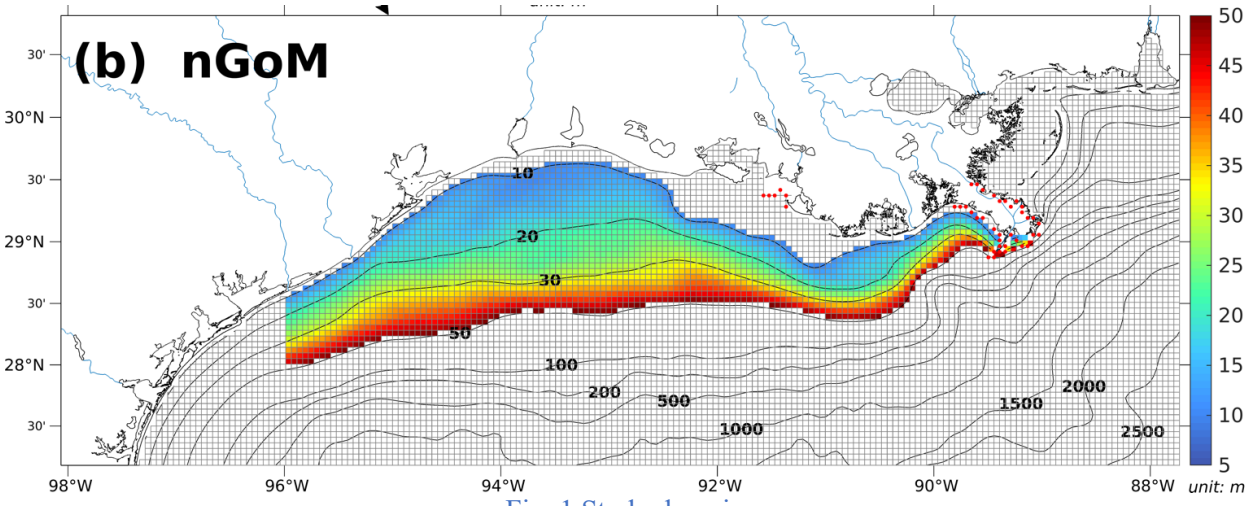

Fig. 1 Study domain.

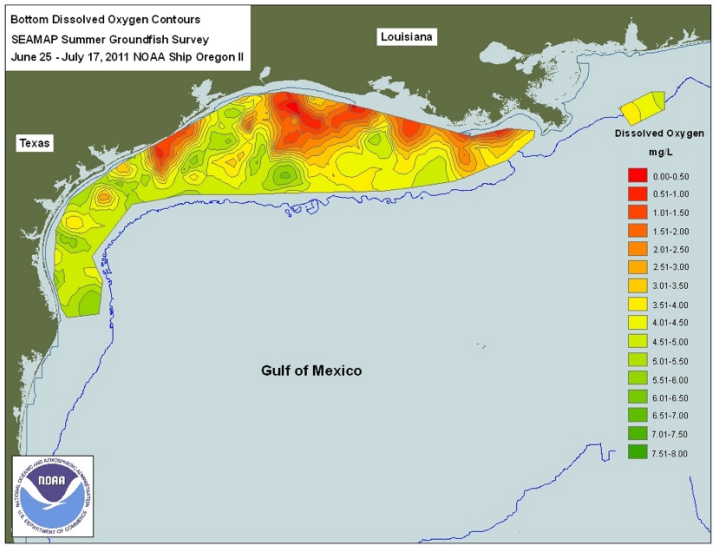

Fig. 2 Distribution of bottom dissolved oxygen concentration provided by the SEAMAP Summer Groundfish Survey in 2011 summer.

L91: what do you mean by up to?

Response: We wanted to state that the correlation between PEA and SS is -0.88 which is high. We will correct this sentence as:

Indeed, the correlation of regionally averaged PEA and SSS is significantly as high as -0.88 ($p<0.001$; Figure 1a) which emphasizes the importance of freshwater-induced stratification.

L148: It might be helpful to include these equations here.

Response: We will include these equations in the revision.

L158: Can you discuss the biological meaning of this time lag? It seems to indicate that mid-summer hypoxia is fuelled by early summer loads and therefore that there is no relationship between May load and summer hypoxia.

Response: The time lag here represents the time between the occurrence of massive organic matter consumption on the sediment and massive nutrient supply by rivers. The maximum nutrient loads do not always occur in May but sometimes in June and July (see below Fig. 3, we will provide daily time series

of riverine nutrients load in the supplementary materials). The correlation shown in Figure A2(a) is around 0.68 when the Mississippi nitrogen load leads by 60 days. The relationship between May load and mid-summer hypoxia is also statistically significant. But the correlation reaches the maximum when the Mississippi nitrogen load leads by 19 days.

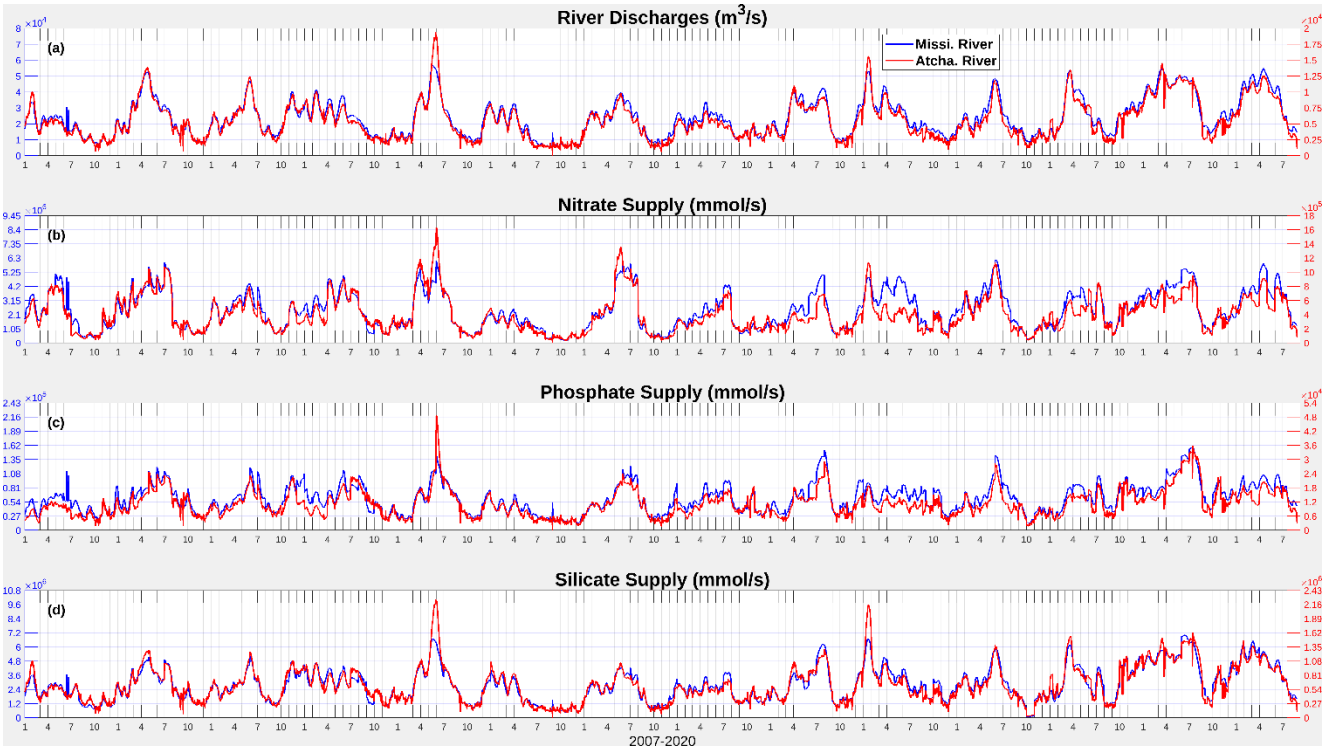

Fig. 3 Daily time series of river discharge, nitrate supply, phosphate supply, and silicate supply by the Mississippi River and the Atchafalaya River from 2007 to 2020. Data is derived from the USGS.

L176 (Table 1): "Hypoxic area" would be better than "Area of extremely low dissolved oxygen concentration"
Response: We will correct it.

L194: Figures are not presented in order, please reorder
Response: We will move this sentence around Figure 4.

L198 (Figure 1a): The lack of relationship between SOCalt and botT is a bit concerning, can you comment?
Response: The SOCalt not only depends on DCP_{Temp} but also on the riverine nutrient supply. Riverine nutrient supply does not reach the maximum as the temperature reaches the maximum in August. The maximum riverine nutrient supply was usually found from May to June. This is the reason why we had a low correlation between SOCalt and DCP_{Temp}.

L198 (Figure 1g): What is the time range of these data, all year, spring-summer, springfall?
Response: The date is provided daily from 1 January 2007 to 26 August 2020.

L223-228: I didn't get how this added term solves the high level of correlation between predictors
Response: The multicollinearity problem is hard to solve, but easy to be quantified by variance inflation factors (VIFs). We first developed the model and then quantified the multicollinearity among the selected predictors. As we stated in L259-260, the VIFs among the selected predictors are 2.60, 2.43, and 1.23 for

PEA, SOCalt, and DCPTemp, respectively. A VIFs value lower than 5 is usually considered as weak multicollinearity.

L258: "impaired"
Response: Corrected.

L264-274: Not sure if that is a good test of model skill. Excluding randomly half of the years (or 30-40%) would have provided a good dataset for testing. Can you discuss why you did not split the hypoxia data into years, since hypoxia is a seasonal process?
Response: We did not split the data by year but instead by the data feature of the hypoxic area. As shown in Figure 1(b), the distribution of daily hypoxic area is highly right-skewed with much fewer values in the medium- and high-value ranges. Splitting the data based on years does not guarantee the hypoxic area in the training set covers the entire range, which would weaken the model performance. Instead, we split the data maintaining the distribution of hypoxic area in both the training set and test set. More specifically, 80% of samples with hypoxic area within a given range (e.g., 0, (0, 5000], (5000, 10000], etc.) are chosen randomly for the training set, the rest 20% are put into the test set. Thus, the ranges of the hypoxic area in both sets can be guaranteed the same range as the entire dataset. This is the reason why we did not consider the daily data as a time series. Since hypoxia occurs annually, the low, medium and high range of hypoxic areas are corresponding to non-summer seasons, early summer, and mid- and late summer, although the comparison shown in Figure 4 looks like time series, it is not.

L281 (Figure 4): You should add observations.
L376: It is an interesting technique but lacks observations, why didn't you do a real forecast, i.e. a week ahead of the mid-summer cruise, for each year where the input data are available?
L373: Why not doing that for the entire time series?
L386: Your model forecast doesn't seem to do better than the seasonal forecast in 2019 and misses the pre-sampling mixing event, can you comment? The 2020 mid-summer hypoxic area is also largely overestimated (~20,000 vs 5,000) and seem to be doing worst than seasonal forecasts despite the model ability to take into account the effect of wind (there was a tropical storm before the mid summer sampling that year)
Response: We will compare our prediction with shelfwide observations, and the NOAA forecast for the entire time series.

The HyCOM global products can hardly capture all the hydrodynamical features due to the relatively low temporal resolution (monthly) of riverine forcing in the model. As shown in Figure 6a, the PEA is more underestimated in the HyCOM dataset than in the ROMS hindcast results. Even though the HyCOM products are scaled according to the relationship with the ROMS hindcast, the HyCOM-derived PEA is still overestimated or underestimated when compared to the ROMS hindcast. This is the main drawback of using the HYCOM products in forecasting. The low performance of the pre-sampling mixing event in 2019 can be attributed to this reason.

The 2020 mid-summer hypoxic area is not largely overestimated. The observed value is 5480 km^2 on around day 570 (Figure 7) when the prediction is around 7000 km^2.

L289: the correlation doesn't seem to be significant
Response: The SOCalt and DCP_{Temp} are not significantly correlated. However, DCP_{Temp} is a proxy of the decomposition rate of organic matter and also a proxy of sediment oxygen consumption. We cannot see the mechanism from the statistical regression model, but can attribute the significant coefficients to some explainable mechanisms according to the reference of predictors.

L293 (Table 2): What is Pr? does it make any sense to provide a Pr of <1e-16?
Response: Pr is the p-value. The p-values are all < 2E-16 not < 1E-16.

L299: "procedure"
Response: Corrected.

L316-317: Early summer or spring? It looks like hypoxia develops in Spring in the time series
Response: The comparison shown in Figure 4 is not a time series comparison. According to Figure 7 in the Part I paper, the hypoxia develops rapidly in late spring and early summer.

L316-323: Do you see all that in Figure 5?
Response: The predicted hypoxic area is a summation of s(PEA), s(SOCalt), and s(DCP_{Temp}). Thus, a greater smooth function of the corresponding predictor indicates greater influences on the hypoxic area.

L343: This is not a discussion, see main comment above.
Response: We will improve our Discussion section according to the comments.

L377: "slight": ~20+% difference
Response: We will provide the percentage changes of prediction compared to the hindcast results to further illustrate it.