Responses to comments on: "Daily hypoxia forecasting and uncertainty assessment via Bayesian mechanistic model for the Northern Gulf of Mexico" (Referee #1)

Our responses are in blue.

General comments

One important piece of information that is not mentioned explicitly enough in the introduction or the abstract is that the main part of the manuscript is about using a statistical model to generate suitable parameters for an existing mechanistic model (referred to DMO20). Reading these sections for the second time, it becomes a bit more clear but it would be beneficial to the reader to describe this more clearly early in the manuscript.

Thank you for your feedback. We will clarify in the introduction that in this manuscript we are using an existing mechanistic model and associated posterior parameter distributions, which were determined through Bayesian inference.

The authors are careful in creating a forecasting scenario in which no summer data is included. However, the entire set of historical data is used to produce the linear regressions, which may include future data w.r.t. year for which the forecast is produced.

We appreciate the comment and understand the concern of the referee. Note that we do not use the predictions of these regressions directly in the forecast; we only apply them for defining the 10 relevant years. To double-check the validity of these regressions, we performed leave-one year-out cross validation (LOOCV), where we excluded years by one, calibrated the models to the reduced dataset and predicted for the excluded year. The performance of August–September dropped dramatically for cross-validated results compared to models built on the full dataset. However, as highlighted in the Results (Lines 173-175), we are only using the June–July regressions for relevant years selection. We will add the description of this additional analysis to the manuscript and expand Table 1.

Model	R^2 from Table 1	LOOCV R^2	% Change
$\sqrt{Q_{A6}}$	0.79	0.75	-5.1
$\sqrt{Q_{A7}}$	0.47	0.38	-19.1
$\sqrt{Q_{A8}}$	0.28	0.14	-50.0
$\sqrt{Q_{A9}}$	0.13	0.04	-69.2
$\sqrt{L_{A6}}$	0.76	0.71	-6.6
$\sqrt{L_{A7}}$	0.51	0.41	-19.6
$\sqrt{L_{A8}}$	0.3	0.17	-43.3
$\sqrt{L_{A9}}$	0.11	-0.02	-118.2
$\sqrt{Q_{M6}}$	0.77	0.73	-5.2
$\sqrt{Q_{M7}}$	0.48	0.38	-20.8
$\sqrt{Q_{M8}}$	0.28	0.15	-46.4

$\sqrt{Q_{M9}}$	0.13	0.04	-69.2
$\sqrt{L_{M6}}$	0.78	0.74	-5.1
$\sqrt{L_{M7}}$	0.51	0.41	-19.6
$\sqrt{L_{M8}}$	0.25	0.12	-52.0
$\sqrt{L_{M9}}$	0.09	0.01	-88.9

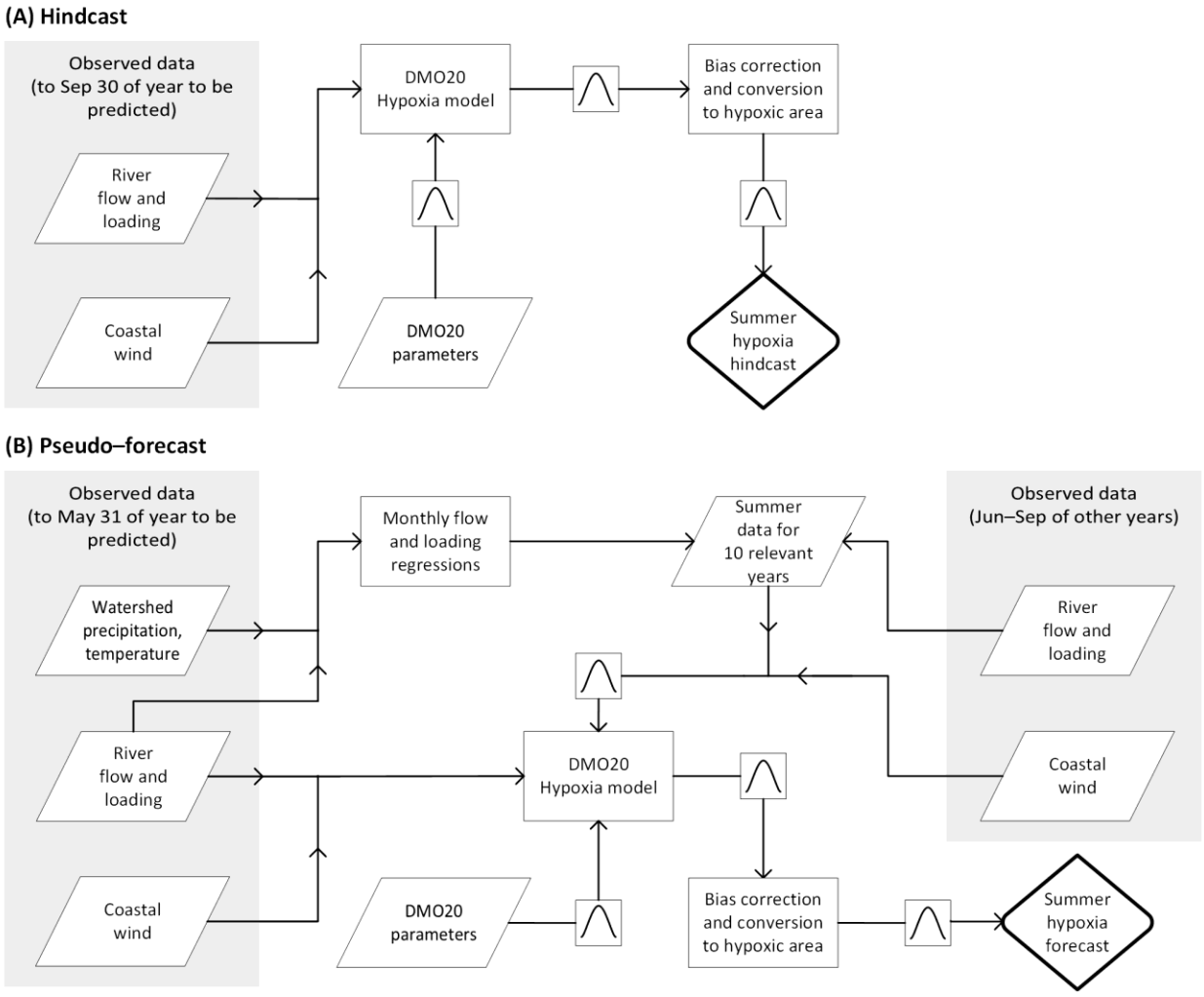
If I understand the approach correctly, even in the Case 4 setup, the forecast may include information from summer data of the current year, if the current year is "relevant" (as defined in the manuscript). As a result, does the forecasting system produce significantly better results for the "relevant" years compared to other years? In addition to the 4 cases currently included in the manuscript, I would suggest to add a Case 5 that excludes all data from the future (forecasts for the first few years with little data could be skipped) or, alternatively, excludes the data from the current forecast year, even if it is relevant.

Actually, the pseudo-forecasts do not include the information from the summer of the current year (forecast year) in any of the forecasting scenarios, as stated in the Methods at Lines 139-140. In other words, when the procedure selects the 10 relevant years, the summer data for the forecast year is excluded from selection. We will add a brief note to the Results (Section 3.2) to remind readers of this important point. We do not think that the suggested Case 5 is warranted for this reason and because it does not directly address the study's research objectives (Lines 63-68).

We will modify Figure 1 (see draft below) to clarify that the forecasts only use observed data up until 31 May of the forecast year (since the nominal forecast release date is 1 June). Data for after 31 May are sampled only from "other years".

Looking at the results in Fig. 2 and 3, it appears as if the forecast-hindcast as well as the forecast-observation comparisons show a pattern in August and September: The forecast appears to overestimate BWDO and HA for low values and underestimate it for high values and this pattern appears to increase in time. The authors already introduce a linear regression for the purpose of bias reduction but apply it only to the June forecast. A similar linear correction could be applied to correct this pattern which appears to increase with lead time. Yet, I am a bit hesitant to recommend such a correction because it, just like the bias reduction, adds a non-mechanistic element to the model.

Thank you for the comment, in fact we agree with the reviewer regarding some bias presence in September for the east section. However, we decided not to introduce a September bias correction for two main reasons. First, there are fewer measurements of hypoxia available for September, so the case for a bias correction is less strong (Fig. S2.2). Second, we see that HA is predicted surprisingly well in September, even without a bias correction (Fig. S.3). We also note that hypoxia is typically reduced in September compared to June–August, so September is of less interest to management and fisheries compared to other months.



How difficult would it be to extend the approach presented in this study and estimate August and September values from all data available until then? I am not suggesting that this needs to be done in this manuscript, yet creating successive 2 month forecasts appear a suitable course of action for producing more accurate estimates. This could be mentioned in the discussion.

We appreciate the comment and agree that being able to update and extend the forecast over the course of the summer could be beneficial for fisheries management. In this manuscript, we focus on a 1 June forecast release date, consisting with current Gulf forecasting practices. However, the methods should be transferable to other forecast release dates. We will expand the Discussion, pointing this out as a future opportunity.

Overall, while the manuscript is well written it sometimes overestimates the study-specific knowledge of the reader. Including more information explicitly would benefit readers. In some instances, I had to read ahead to answer questions which could have been addressed right away. I have listed some of those instances below.

Thank you for the positive feedback. In addition to addressing your specific comments, we will review and edit the entire manuscript for clarity.

Specific comments

L 8: “Several models” Here it would be helpful to specify what type of model is meant, e.g. “dynamical”, “statistical” etc.

We appreciate the comment and will add clarification to the abstract. Furthermore, we expand on the modeling approaches to predict and forecast hypoxia later in introduction (Lines 38-40, 46-55).

L 52: The same group has previously considered different sources of uncertainty in the 3d model, finding that variations in wind forcing had the largest impact on hypoxia estimates (J. P. Mattern, K. Fennel, and M. Dowd (2013), Sensitivity and Uncertainty Analysis of Model Hypoxia Estimates for the Texas-Louisiana Shelf, Journal of Geophysical Research, doi: 10.1002/jgrc.20130).

We agree that the mentioned manuscript provides a more elaborate analysis of uncertainty than Laurent and Fennel (2019), but not in a forecasting framework. We will add this reference to Line 55, for readers who are interested in exploring these uncertainties more.

L 77: After the first read, I am assuming that the DMO20 model has four compartments, an eastern and a western one, each divided into two layers. This could be made a bit more explicit in the text.

Thank you for the comment, we will add clarification in Section 2.1.

L 91: Is there any indication about the cause of this bias? It is nice to have an underlying parsimonious mechanistic model, yet the bias correction introduces a non-mechanistic element. By the way, it would be helpful to mention again that June is the start of the prediction interval and that the bias disappears over the course of several weeks, so that it can be neglected in the following month.

Thank you for the comment, the June bias might be due to multiple factors, including limitations of the DMO20 model formulation. We will add the suggested clarifications.

L 110: Is this done for one or multiple years? All years with data? This information is probably given later but it would be useful to mention it here already or even earlier.

Thank you for the comment, summer riverine inputs were sampled from multiple years, and we will add clarification.

L 118: Do the ten most relevant years represent the full time period accurately?

We appreciate the comment and the results of this study suggest that 10 relevant years perform well in representing the summer riverine inputs, which is reflected in increase of R^2 compared to randomly selected years (Table 2, compare overall for Cases 2 and 4).

Note that our goal is not to represent the entire study period, but to represent the range of summer conditions indicated by the flow and loading regressions (considering spring flows, rainfall, and loading). We recognize that there is some ambiguity regarding the appropriate number of relevant years to include. Thus, we have performed additional analysis (running the forecasts using only 5 relevant years) to test the sensitivity of the model to the number of relevant years included. We will add a brief discussion of this additional analysis to the manuscript.

L 144: It may be good to give some examples of the model parameters that contribute to the uncertainty here, so that the reader does not need to consult the DMO20 paper to get this information.

Thank you for the comment, and we are going to add the summary of the calibrated parameter estimates to the supplementary material.

L 150: What if one, multiple, or maybe all relevant years are in the future w.r.t. to the estimated year?

Thank you for the comment. Our approach assumes that there are sufficient records to generate a robust sample of representative years. Whether those years happen to be past or future is not directly relevant to our research questions. Clearly, there are sufficient samples moving forward from the present. If we wanted to generate a forecast in 1985, we'd need older records, but this is beyond the scope of this study, and we don't think it would be particularly meaningful to our research objectives.

L 156: "Sixteen multiple linear regressions": I assume, the 16 refers to 4 (months) * 2 (rivers) * 2 (discharge, nitrogen loading) but this could be made a bit clearer, or a reference to Table 1 could be added here already. In my opinion, it would be good to clearly state again that there are 4 regressions for each month.

We agree with the comment and will add this clarification.

L 185: The "hypoxia model" is DMO20, correct? I would suggest to include this here again.

We agree with the comment and will add this clarification.

L 192: Is there a distinction between "forecasted" in this line and "pseudo-forecast" a few lines above? It would be good to stay consistent with the use of "pseudo". Maybe even drop the "pseudo-" prefix after describing that this is the way the word forecast is used in the context of the manuscript.

Here, we use 'forecasted' as the adjective formed from a verb, while the product of forecasting in our case is 'pseudo-forecast' because we are generating the forecast post factum, as noted at Line 146.

Fig. 2: Am I correct in assuming that there are 32*30 red dots in the top panels, one for every day in June in the 32 years with data? But if the monitoring cruises are typically in late July, why are there so many red dots in the bottom panels?

The statement related to the top panels is correct. Red dots in the bottom panels indicate 34 available observations for June. There are total of 149 observations (i.e., cruises), 63 in July, 35 in August, 17 in September. Note that we are not just using LUMCON cruises. We are also using Texas A&M cruises, NOAA SEAMAP cruises, etc., so there are many cruises outside of July. See Matli et al., (2018) for more details. We will also clarify this in Section 2.2.

Fig. 4: I can only distinguish between 3 shades of gray here, yet the caption suggests there should be 4. Are the uncertainties plotted cumulatively or is the effect of parameter uncertainty generally smaller than that of the riverine and meteorological inputs?

There are four shades of grey. However, bands related to regression transformation of BWDO to HA (the lightest grey) do not add much visually to mechanistic model error, data input uncertainty and parameter uncertainty. These transformation regressions have high R^2 values and relatively small error (R^2 = 0.98 and the residual standard deviation is 706 km^2 for west section, while for east section the R^2 = 0.99 and the residual standard deviation is 216 km^2; see Del Giudice et al. (2020) for details). To address the reviewer's comment, we will switch the sequence of the shades of grey, so that the thin outer band is more prominent.

L 238: "Note that the relative magnitudes of the variance components are somewhat different from the relative magnitudes of the 95% IQR components ...": If the goal here is to say that the relative magnitudes differ because the variance has squared units I think it would be easier for the reader to state this directly, rather than drawing a line from IQR to standard deviation.

We agree with the comment and we will clarify the text.

References

Del Giudice, D., Matli, V. R. R. and Obenour, D. R.: Bayesian mechanistic modeling characterizes Gulf of Mexico hypoxia: 1968–2016 and future scenarios, Ecol. Appl., 30(2), eap.2032, doi:10.1002/eap.2032, 2020.
Laurent, A. and Fennel, K.: Time-Evolving, Spatially Explicit Forecasts of the Northern Gulf of Mexico Hypoxic Zone, Environ. Sci. Technol., 53(24), 14449–14458, doi:10.1021/acs.est.9b05790, 2019.
Matli, V. R. R., Fang, S., Guinness, J., Rabalais, N. N., Craig, J. K. and Obenour, D. R.: Space-Time Geostatistical Assessment of Hypoxia in the Northern Gulf of Mexico, Environ. Sci. Technol., 52(21), 12484–12493, doi:10.1021/acs.est.8b03474, 2018.