Response to reviewer comments "Parameter sensitivity and identifiability for a biogeochemical model of hypoxia in the northern Gulf of Mexico", Beck MW, Lehrter JC, Lowe LL, Jarvis BM.

We thank the editor and reviewer for providing helpful comments on our manuscript. Responses to these comments are shown in italics.

Editor's comments:

The authors executed the work on parameter sensitivity and identifiability for a biogeochemical model of hypoxia in the northern Gulf of Mexico. The authors did not provide proper conceptual diagram of the model.

The previous conceptual diagram was for CGEM that included both the biogeochemical components of FishTank and hydrodynamic components. A new figure was created that inludes only the FishTank components.

The model is not presented in detail, it is necessary to give more details on original model.

We have added Table 1 that includes a full list and description of parameters that were evaluated for the sensitivity analysis. This table includes short and long descriptions, units, and starting values.

Full details of the model are provided as an appendix in Lehrter et al. 2017. We have not included the content here because it includes nearly twenty pages which we do not wish to duplicate. We refer the reader to the appendix twice in section 2.1, which includes a new sentence in the first paragraph: 'A full description of the model structure, equations, and parameters is described in appendices A-F in Lehrter et al. (2017).'

Discussion part is unnecessary long, lot of introduction in the discussion. This part needs major revisions and complete restructure.

The discussion was shortened and introductory material was removed. Specifically, the first paragraph of the discussion was removed and the first paragraph of sec. 4.1 was shortened. This reduced the length by over one page.

The sensitivity analysis was adequately implemented but further clarification is needed for how the water quality model was applied in Weeks Bay.

We present a thorough recalibration of the model with an alternative dataset described below. Please see our responses to the reviewer.

In view of the above comments and reports of reviewers, your manuscript has been evaluated and you are informed to resubmit the manuscript after proper and thorough revisions in accordance with the comments of editor/reviewers. Review reports are appended below. Thank you for your submission in Ecological Modelling.

Reviewer 1:

The authors present the sensitivity analysis of a 0-Dimensional biochemical model implemented in Weeks Bay Alabama. The authors evaluated the sensitivity of the model predictions of DO, ammonium, nitrate, chl-a and irradiance to perturbations in several model parameters. In general, the theoretical basis of the sensitivity analysis is very interesting and mathematically sound and the use of the collinearity index (Eq. 3) to explore linear correlations between parameters seems to be very useful to help identify the most important parameters controlling the model performance (in terms of goodness of fit to observed datasets). The paper can be accepted pending major revisions.

My main concern with the manuscript is that while the application of the sensitivity analysis to the FishTank model seems ok, the application of the FishTank model to Weeks Bay, AL is weak and not convincing. First of all, a correct representation of the advective and dispersive transport is critical to be able to capture the fate and transport of water quality in an estuary. Without a correct representation of transport it is almost impossible to be able to obtain a predictive water quality model. The FishTank model in this case neglects all modes of transport in Weeks Bay, AL and therefore assume the estuary is like a bathtub. This is obviously incorrect and represents a clear limitation of the paper.

We agree with the reviewer that this application is not entirely appropriate given the hydrodynamic influences on the observed data. However, our objective was to demonstrate that unit-testing of the zero-dimensional is an efficient means of improving larger models that are computationally intensive. Our intent was not to present a complete predictive model with hydrodynamic forcing to Weeks Bay. To address the concern of the reviewer and to maintain our emphasis on unit-testing, we have chosen a a more appropriate dataset, described below. A sentence added to the final paragraph of the introduction emphasizes our intent of the analysis:

'In general, the analysis makes a case for unit-testing with lower-dimensional models, particularly when computional limitations preclude the use of conventional optimization algorithms with larger three-dimensional models.'

An alternative dataset was used that included flask experiments of oxygen production from Pensacola Bay, Florida. This dataset included light and dark treatments of water samples in closed bottles for an approximate 12-hour period, including before and after measurements of DO concentrations with initial nutrient values for each treatment and replicate. The dataset is described in detail in Murrell et al. in revision (Murrell, M. C., Caffrey, J. C., Marcovich, D. T., Beck, M. W., Jarvis, B. M., III, J. D. H., in revision. Seasonal oxygen dynamics in a warm temperate estuary: Effects of hydrologic variability on measurements of primary production, respiration, and net metabolism. Estuaries Coasts - (-), .). Methods were added to the manuscript, Sec. 2.4:

'Estimates of oxygen production from closed bottle experiments were used to calibrate the 0-D model. Surface water was collected from a fixed location in Pensacola Bay, Florida (30.35° N, 87.20° W) on September 25th, 2013 (full details in Murrell et al. in revision). Water samples were placed into closed 300 mL glass bottles and transported back to the lab for treatments. Concentrations of O_2 and nutrients (NH_4^+, NO_2^-, NO_3^-, PO_4^{3-}, and SiO_3^{2-}) were extracted from the samples within one hour of collection. Oxygen concentrations were established using Winkler reagents (Parsons et al. 1985) and titration using a MetrOhm Titrando with thiosulfate titrant calibrated against an iodate standard and electrochemical endpoint detection. Ammonium was analyzed using a fluorometric method (Holmes et al. 1999); other nutrients were analyzed using either a Thermo Fisher Aquakem 200 discrete analyzer or an Astoria-Pacific continuous flow analyzer using standard colorimetric methods (APHA 2005).

Oxygen production after an approximate 12 hour period (morning to evening) was estimated within the bottles using light and dark treaments. All treatments were exposed to ambient outdoor conditions in a water bath at the laboratory. The dark treatments used opaque black bottles to ensure no exposure of ambient light to the samples. The exposure experiments began at 7am and were concluded at 7pm, after which the concentrations of O_2 in each bottle were measured. Each treatment had four replicates and HOBO® pendants were used to measure continuous temperature and light. Starting O_2 concentrations for the experiments were 200.78 (± 4.47) mmol O_2 m^{-3} and ending concentrations were 229.57 (± 2.66) mmol O_2 m^{-3} for the light treatments and 195.51 (± 1.9) mmol O_2 m^{-3} for the dark treatments. Initial nutrient concentrations were 0.71 mmol m^{-3} for NH_4^+ (single sample), 0.13 mmol m^{-3} for PO_4^{3-}, and 38.4 mmol m^{-3} for SiO_3^{2-}. Concentrations for NO_2^- and NO_3^- were below detection limit.

Each parameter subset for O_2 from the selection heuristics was calibrated to the ending oxygen concentration for each of the four replicates of each treatment. The intial nutrient and oxygen concentrations were used as starting conditions for FishTank. The corresponding continuous temperature and light measurements were also used as forcing conditions. Parameter values were searched using the optim function... '

For the sensitivity analysis the authors manipulated the observations of DO in the estuary filtering the tidal impacts to be able to implement the FishTank model. However, there is no evidence of the resulting time series of DO or of the other variables analyzed. I am concerned that while removing the DO variations due to tidal cycles, the authors could have also removed the daily variation of DO which can be associated to diurnal cycles of sunlight and temperature. If this is the case, then the DO calibration is based almost on synthetic and unrealistic DO data. The authors must include a section showing the data that was used for the calibration analysis. This section should include a location figure showing Weeks Bay, AL, the stations used to extract the original datasets, plots of the unprocessed and processed time series of DO, ammonium, nitrate, chl-a and irradiance as well as plots showing the processed data against the calibrated model outputs for each variable. The authors must also include a table with the calibrated parameter values. This information is standard in all modeling studies.

Our use of the flask experimental data addresses these concerns. The initial and ending DO concentrations for the experimental data, including intial nutrient concentrations, were reported in our revision to the methods. We have also added Table 7 for more details on the calibration results for each parameter subset. In addition, Figure 7 was added (old Figure 7 is now Figure 8) to demonstrate the opimization routine using the phytoplankton parameter set in Table 3 for dissolved oxygen.

The new results for the experimental data present some interesting challenges that demonstrated how unit-testing can inform model understanding. The implications of the analysis were described in an addition to the discussion, sec. 4.1:

'..However, calibration to most parameter sets showed little improvement in model fit despite relatively large changes in parameter values. This result underscores the challenge of model optimization when conidering many parameters. Although we demonstrated model sensitivity to individual parameters, the interactive effects are also important. For example, optimization to zooplankton parameters showed almost no change in model RMSE despite a relatively large sensitivity of the model to individual parameters. A more comprehensive approach that considers the interaction of zooplankton parameters with others, such as phytoplankton, is warranted given the structural dependencies betwen parameter categories. Indeed, the reduction in RMSE for selection heuristics that equally considered the parameter categories (third heuristic) showed the second largest reduction in model error. This subset included both phytoplankton and zooplankton parameters, suggesting more careful consideration of these sensitive parameters could further improve model calibration.'

Please also describe what are the forcing conditions of the FishTank model. There are two tributary rivers in Weeks Bay and one outlet condition. How is each one simulated in your FishTank model?.

Please see the description above justifying our use of the new dataset.

Other comments:

Introduction. I suggest the authors to include also references of couple hydrodynamic and water quality studies available in the region. For example, the work of Camacho et al. (2014a) in St. Louis Bay, MS and part of Mississippi Sound is a relevant citation that is missing in this investigation. The work of Dortch et al. (2007) in Mississippi Sound should also be cited.

Citations for Dortch et al. 2007 and Camacho et al. 2014a were added to the first paragraph.

Sensitivity analyses have also been conducted in the past in the region. For example, Camacho et al (2014b) conducted an uncertainty and sensitivity analysis of a hydrodynamic model of Weeks Bay Alabama. I suggest the authors to include these references. The sensitivity estimates provided by Camacho et al. (2014b) seem to be based on the same equation Eq. (1) included in this manuscript. Can the authors briefly indicate (perhaps one or two sentences) if both methods are the same?

Yes, this approach is similarly described by Camacho et al. 2014b. The following text was added to paragraph one in sec. 2.2: 'This approach is described as a First Order Variance Analysis that linearly propagates changes from model parameters to model predictions (Camacho et al. 2014b).'

lines 24 - 39 page 3. This part of the text is full of vague statements. Please be more explicit. What are the "characteristics expected from the output" and the "information represented by the structural components". What do you mean with "Given that these characteristics cannot be simultaneously achieved, models are developed in partial dependence of reality and theoretical constructs, completely separate from both, or dependent on one or the other".

This section was revised as follows:

'The development of a model represents a tradeoff between achieving predictive accuracy and a realistic representation of environmental processes [...] Given that these characteristics cannot be simultaneously achieved, models are developed to balance predictive accuracy with environmental realism or often favor one at the expense of the other.'

lines 24 - 39 page 3. I think the reference Levins (1966) is extremely old and must be removed.

This citation was referenced in Ganju et al. 2016, from which this information was obtained. We have not removed the citation because we feel that age does not make a publication inadequate. However, we have added the citation to Ganju et al. 2016 as a more current reference.

It is interesting to see that the authors constantly use the term "precision" to denote goodness of fit. In general, precision is used to denote the number of digits of a number. Note that a number can be precise but inaccurate. I suggest the authors to revise the text and replace "precision" with a better term to denote goodness of fit. e.g. level of agreement between observations and simulations.

We have found that precision is commonly used in the literature to describe goodness of fit, although we understand the distinction made by the reviewer. Considering an example where observed data represent a biased sample of an environmental process (e.g., infrequent or uneven sampling of an event that varies over time), model calibration to these data can only ever achieve precision. Although practitioners would not knowingly use observed data that are biased for model calibration, we are certain that many ecological datasets are incomplete representations of actual processes. We argue then that model calibration in many cases is better described as seeking precision, although in practice it is stated as predictive accuracy given an incomplete knowledge of the observed data relative to the true *environmental process. Regardless, we have replaced precision with "goodness of fit" or "fit" as a more generic term.*

Line 9 page 10. Selection or Selected?

Changed to: 'Heurestics for parameter selection also recognized...'

Line 4 page 12. what are the default parameter conditions?. You should include a table with the default parameter values and parameter values under 50% modified conditions.

A new table (Table 1) was added for the full parameter set including names, long description, units, values, and the 50% increase in values for the sensitivity analyses.

Lines 29-30 page 16. Please include the units in the reported RMSEs. mmol O2?

RMSE units are reported as mmol o2 m-3, which was added to the text.