Referee general comment:
Thank you very much for the opportunity to review this manuscript describing the internal (physical and biological) and external factors leading to inter-annual variability in the extent anoxia in Lake Mendota. This study uses a combination of three very different types of models to evaluate these various factors. I found this paper very interesting, very well written, and may be very useful to the scientific community. I applaud the authors in using this multi-model approach. However, I think two of the three models have serious flaws that need to be addressed prior to publication.
My main concern is that one of the main takeaways from this paper (internal productivity has very limited effect on interannual differences in anoxia) may not be true. It may be true that physical mixing drives the overall extent of anoxia (baseline), but I think it is too early to say interannual variability in productivity has little affect. I think two of the models need to be reevaluated prior to making those conclusions:

Referee comment:
GLM-AED2. GLM-AED2 simulated the annual progression of anoxia very well, and simulated the importance of stratification driving not only the average changes in DO depletion but also much of the interannual variability in DO associated with stratification. But the model did not capture the interannual variability in surface productivity that may drive the other interannual variability in DO. It clearly could not reproduce the interannual variability in AF. This model had an R2 of only 0.08 and a negative NSE. Part of the problem may be that the model is trying to simulate two very different lakes (one without spiny water fleas and one with them) - all with one set of coefficients (that may not even represent the lake in the first place). Without simulating the big biological change, I am not sure you can get there with this model.
Referee suggestion: Use GLM-AED2 to only simulate one of the periods, either prior to or after the change in biology. If this does not improve the overall ability to predict AF, then the phytoplankton parameters may have to be adjusted. Without being able to predict most of the variability in AF, I really don't see its use in this paper.
Author response:
We are very thankful for this comment by the reviewer, which gives us the chance to discuss the GLM-AED2 performance and hopefully improve the overall manuscript. We make two over-arching points here. The first is that a visible shift in AF occurred in 2010 (Fig. 10b), and this may be explained by changes in the foodweb that affect primary production and organic matter cycling. We have no conclusive evidence of the cause, but the shift is coincident with the invasion of the predacious zooplankton, *Bythotrephes*. We discuss this and have added it to the abstract. The second is that our model reproduces well the ecosystem dynamics prior to 2010, and as the reviewer suggests, the lake is likely in different states, separated by the shift that occurs in 2010. Further, we also acknowledge that the GPP is actually an important driver of the variability in summer anoxia (rel. importance 15 %). The main text was revised accordingly in the abstract and in the discussion:

> L27-32: The summer heat budget, the timing of thermal stratification, and the gross primary production in the epilimnion were the most important predictors of the spatial and temporal extent of summer anoxia periods in Lake Mendota. Inter-annual variability in anoxia was largely driven by physical factors: earlier onset of thermal stratification in combination with a higher vertical stability strongly affected the duration and spatial extent of summer anoxia. A step change upward in summer anoxia in 2010 was unexplained by the GLM-AED2 model. Although the cause remains unknown, possible factors include invasion by the predacious zooplankton*, Bythothrephes longiman*us.
>
> L441-443: We also acknowledge that a step change in the Anoxic Factor occurred in 2010 and was unexplained by our model. Although the cause remains unknown, the timing was coincident with large increases in the invasive zooplankton, *Bythotrephes* (Walsh et al., 2017).

To the point about the model not capturing variability in surface productivity, we added a new figure to the Appendix: Figure A8 which shows the time-series comparison between observed and modeled DOC concentrations. Here, you can see that the model replicated the overall dynamics of DOC concentrations in three different depths over time, which highlights its ability to replicate net aquatic production in the surface layer and its contribution to dissolved organic matter. Fig. 9a of the main text showed that the model overestimates surface dissolved oxygen concentration. This overestimation must have a concomitant increase in organic matter as a consequence of photosynthesis, and in this case is particulate organic matter (POM). Considering our proxy for phytoplankton biomass is well predicted (Fig. 5), this suggests our over-estimate of primary production results in increase in POM that is exported from the epilimnion to the hypolimnion. Unfortunately, we do not have observed POM to calibrate this part of the model, but we feel it is likely that our model has overestimated the contribution of primary production to hypolimnetic organic matter and subsequent oxygen depletion. This underlies our conclusion that primary production may be less important to inter-annual variability than physical factors. We added these sentences to the main text to state this:

> L481-487: Although the model replicated well the long-term DOC dynamics (Appendix Figure A8), it also overestimated surface layer dissolved oxygen concentrations compared to the observed data. This overestimation must have a concomitant increase in organic matter as a consequence of photosynthesis, and in this case in POC. Considering our proxy for the dynamics of phytoplankton biomass is reasonably well predicted (Fig. 5), this suggests our over-estimate of primary production results in increase in POC that is exported from the epilimnion to the hypolimnion. Unfortunately, we do not have observed POC to calibrate this part of the model, but we feel it is likely that our model has overestimated the contribution of primary production to hypolimnetic organic matter and subsequent oxygen depletion.

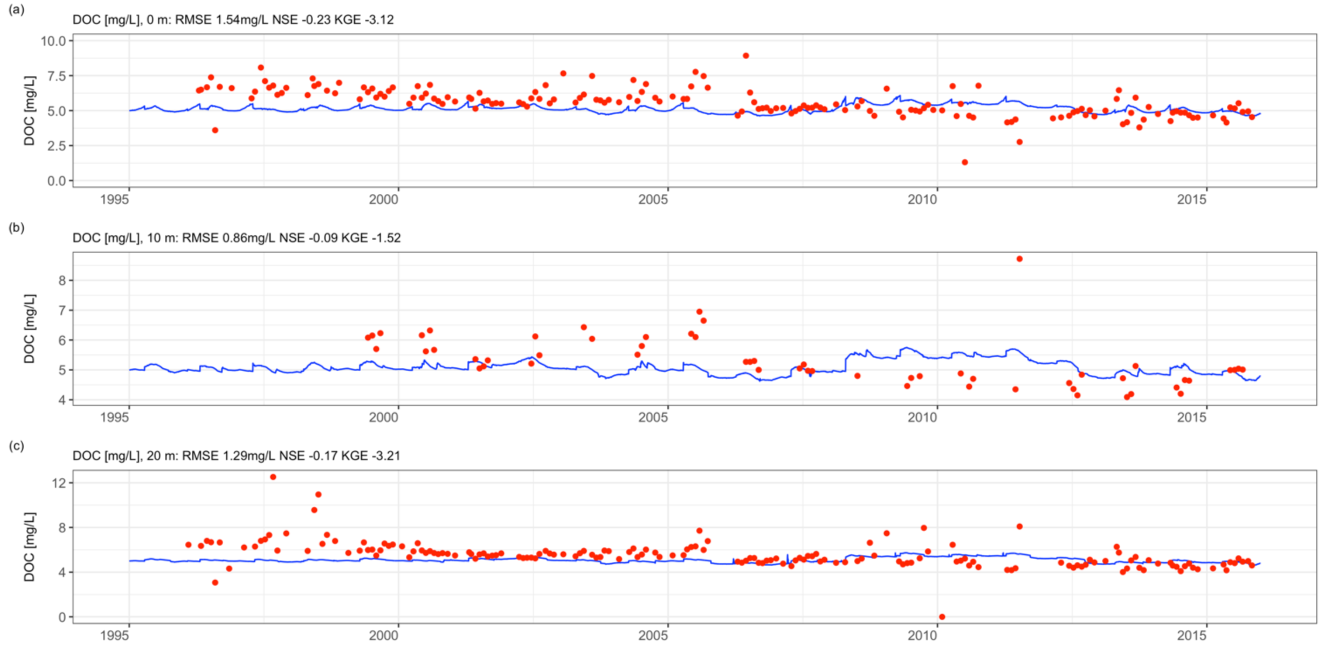


Figure A8 Time-series comparison between observed (red dots) and modeled dissolved organic carbon concentrations (blue lines). The fit criteria root-mean square error (RMSE), Nash-Sutcliffe coefficient of efficiency (NSE) and Kling-Gupta coefficient of efficiency (KGE).

Regarding the capture of interannual anoxia dynamics: Yes, it seems there was a shift in the ecosystem happening beginning in 2010 with higher annual Anoxic Factors. We changed Figure 10 to also show the comparison between simulated Anoxic Factor and the observed data for the periods pre-2010 and post-2010. We also recalculated goodness of fit separately for the two time periods. For the total time period (Fig 10b, 1992-2015 when observed data was available) the model achieved an RMSE of 7.12 d, NSE of -0.22, KGE of 0.26 and r of 0.28 showing that on average it was a week off in replicating the Anoxic Factor, but the KGE and r values proved that the general dynamics and interannual variability could be replicated. When comparing with the pre-2010 period (Fig 10c), the model achieved an RMSE of 6.79 d, an NSE of -0.25, an KGE of 0.44 and r of 0.45, which highlights the model's ability to replicate anoxia dynamics in this period (please note that the model was calibrated for the period 2005-2015 which proves, at least in our opinion, the success of the calibration if there indeed was an ecosystem shift). When comparing with the post-2010 period (Fig 10d), the model achieved an RMSE of 8.04 d, an NSE of -31.99, an KGE of 0.21 and r of 0.62. Here, the model is biased as the observed Anoxic Factor is higher in all years except 2013. Still, the interannual variability expressed by the correlation coefficient r was captured very well by the model. The p-value for the pre-2010 period of the correlation coefficient was p=0.0591. For the post-2010 period, the p-value = 0.19, reducing our confidence in the model for this shorter time period. The visual inspection of these plots (10c and 10d) highlights that they represent different ecosystem states, as there is step-change in the Anoxic Factor starting in 2010.

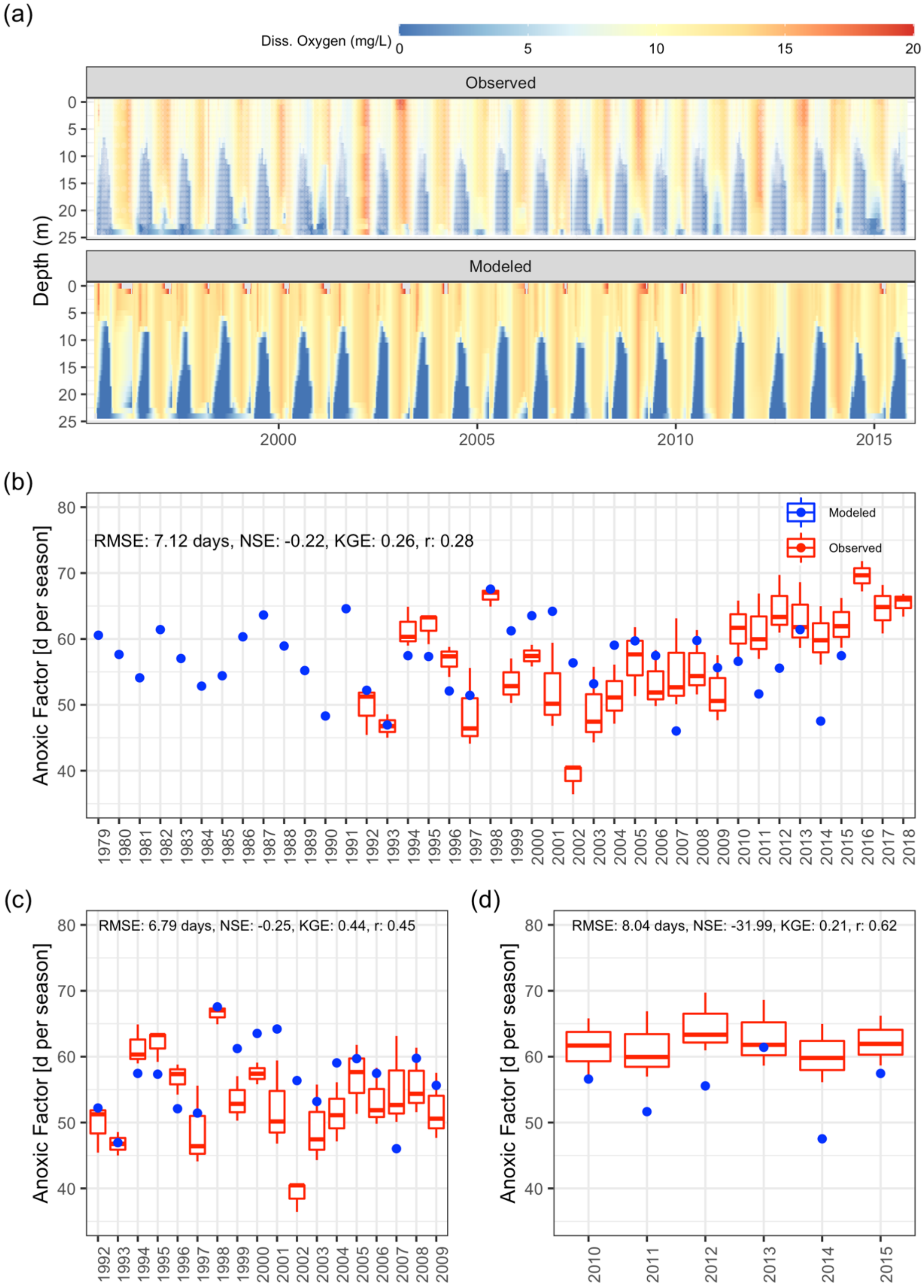


Figure 1 Comparison of observed to modeled dissolved oxygen concentrations and ecosystem response. A **Contour plot of observed (upper figure, white dots mark sample events) and simulated dissolved oxygen concentrations.** B **Comparison of**

simulated Anoxic Factor (red dots) against interpolated range of Anoxic Factor derived from observed data (box-whisker plots) over the period 1979 to 2018. C **Comparison of simulated Anoxic Factor (red dots) against interpolated range of Anoxic Factor derived from observed data (box-whisker plots) over the period 1992 to 2009.** B **Comparison of simulated Anoxic Factor (red dots) against interpolated range of Anoxic Factor derived from observed data (box-whisker plots) over the period 2010 to 2015.**

Therefore, we focused our regression analysis on the pre-2010 period. First, we inspected if the distributions of the observed and modeled Anoxic Factors were similar by investigating the null hypothesis that they are identical populations as determined by the Wilcoxon test (see attached figure below). This test achieved a non-significant p-value of 0.13, indicating strong overlap in populations, and therefore are comparable. We added this figure as Figure A9 to the Appendix A of the manuscript. Also, a similar comparison of the Anoxic Factors for the post-2010 period revealed that observed and modeled distributions were significantly different with a p-value of 0.032. This effectively highlights that we can talk about "two different lakes here". We added these sentences to the main text in the results and in the discussion:

> L412-421: The simulated Anoxic Factor over the total time period averaged 56.7 ± 5.2 days with an RMSE of 7 days, an NSE of -0.22, and an KGE of 0.26 (correlation coefficient r = 0.28). The model's underestimation of the recent positive trend of Anoxic Factors starting in 2010 was investigated by quantifying the fits during two periods: 1992-2009 (Figure 10C) and 2010-2005 (Figure 10D). In the pre-2010 period (1992-2009), the model achieved an RMSE of 6.79 days, an NSE of -0.25, an KGE of 0.44 and r of 0.45 for Anoxic Factor predictions. In the post-2010 period (2010-2015), the model achieved an RMSE of 8.04 days, an NSE of -31.99, an KGE of 0.21 and r of 0.62. A subsequent Wilcoxon signed-rank test highlighted, that the observed average and modelled Anoxic Factors from the pre-2010 period showed no significant differences between the two distributions, suggesting they belong to the same population (p-value = 0.13, Appendix Figure A9), whereas the distributions of observed mean Anoxic Factors and modeled ones after 2010 were significantly different (p-value = 0.032, Appendix Figure A9).

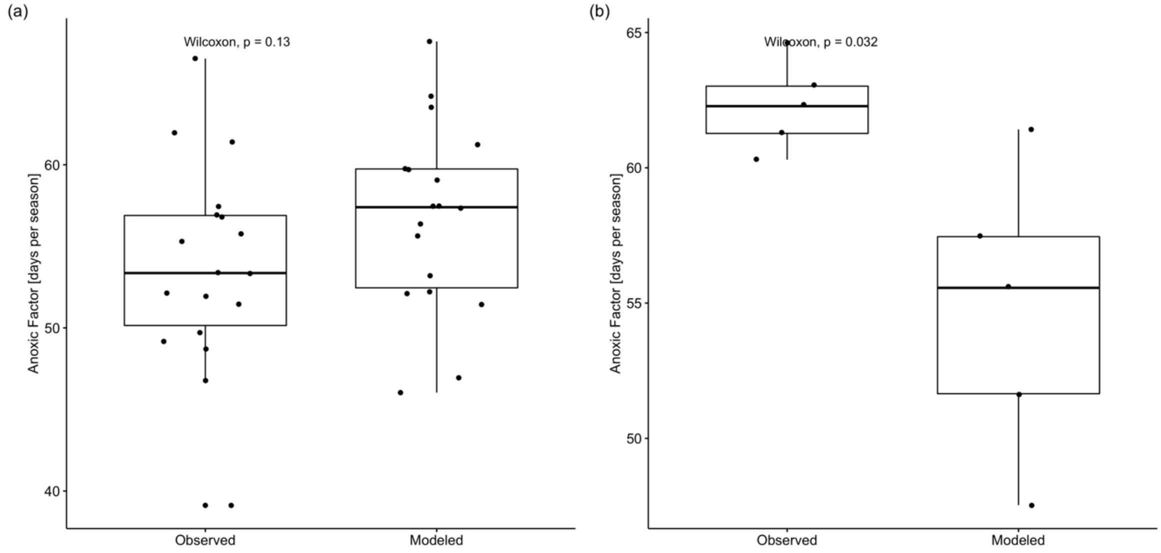


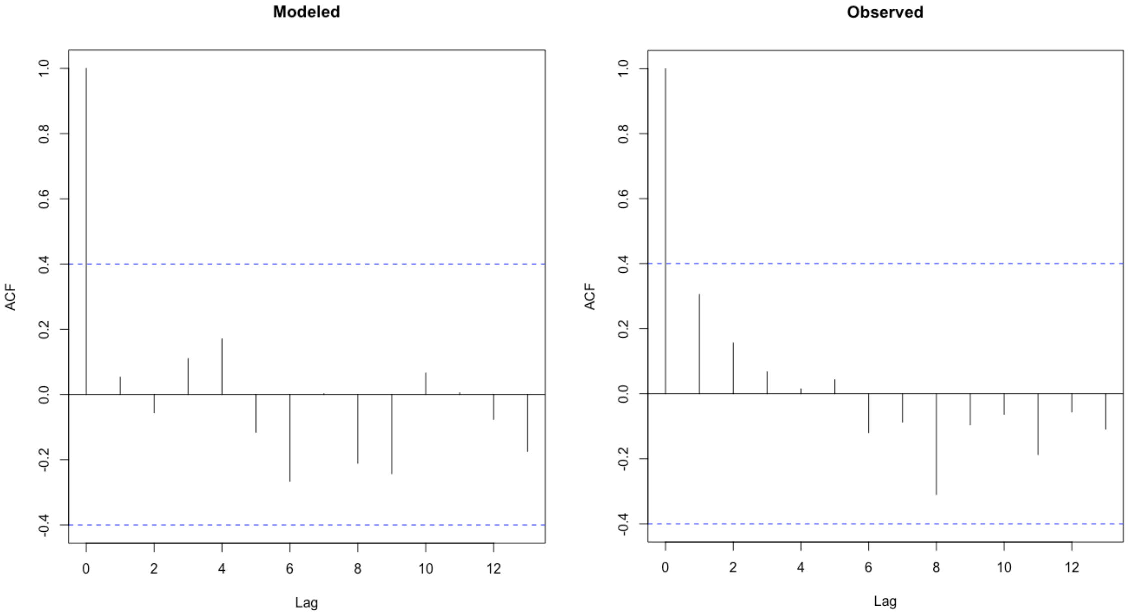
Figure A9 Box-whisker plots of observed to modeled Anoxic Factor for (a) the period 1992-2009 and (b) for the period 2010-2015.

We discussed novel insights into these two distinct periods by expanding this paragraph

> L498-518: The model replicated the maximum anoxia event in 1998 but struggled to replicate the minimum in 2002. The discrepancies of 5-10 days between the simulated and observed range of the Anoxic Factor beginning in 2010 are related to an increased spatial as well as temporal extent of summer anoxia (Appendix Figure A10), which was not captured by the model. This was highlighted by the statistical analysis of the pre-2010 (1992-2009) and post-2010 (2010-2015) Anoxic Factors. Prior to 2010, there were no significant differences between observed and modeled distributions (p=0.13); whereas,

> after 2010, the observed distribution was significantly higher than the modeled distribution (p=0.032) (Appendix Figure A9). For simplicity and due to limitations in Lake Mendota monitoring data post-2010, we focused the regression analysis of the Anoxic Factor in this study only on the pre-2010 period.
>
> The change in Anoxic Factor post-2010 may be due to an ecosystem shift in Lake Mendota that began in 2009, when the invasive spiny water flea (*Bythothrephes longimanus*) was detected in surprisingly high densities in the lake (Walsh et al., 2016b, 2018). Spiny water flea effectively became the dominant *Daphnia* grazer, causing historically low *Daphnia* biomass in 2010, 2014 and 2015 (Walsh et al., 2016a) and reducing water clarity. The spiny water flea may have increased organic matter supply to the hypolimnion by grazing down certain phytoplankton. Mendota's Daphnia population historically consisted of *Daphnia pulicaria* and the smaller-bodied *Daphnia galeata mendotae,* who compete differently with spiny water flea*. D. mendotae* biomass increased in spring after the spiny water flea invasion (Walsh et al., 2017), grazing on phytoplankton and probably accelerating organic matter mineralization before stratification onset. This could be one potential cause that contributed to the increase in hypolimnetic oxygen depletion after 2010. Our GLM-AED2 model could not replicate this food web change, and subsequent shift in anoxia dynamics, due to limitations of the numerical model, i.e., GLM-AED2 had constant ecological parameters over the entire modeling period and did not have zooplankton dynamics instantiated. We envision future monitoring and modeling studies that focus entirely on ecosystem differences and shifts between the pre-2010 and post-2010 periods of Lake Mendota.

Further, by analyzing the autocorrelation function (ACF) of the observed mean Anoxic Factors and the modeled ones (see figure below), we concluded that there is no autocorrelation between annual Anoxic Factors. It may be the case that the interannual variation in the Anoxic Factor (investigated by ACF) is effectively random, which does not mean that the Anoxic Factor is necessarily random, but that the variation in external drivers may be random. Still, our model's simulated Anoxic Factors are from the same distribution as the observed mean values highlighting the model's ability to capture the overall distribution of anoxia. Further, the fit metrics (highlighted in revised Figure 10) show that the model can capture inter-annual variability significantly prior to 2010, even if the average value is off by about a week.



We therefore followed the reviewer's suggestion to only include the pre-2010 period for the regression analysis, which is discussed in the next comment and response block:

> L314-317: Only model output and model driver data from the period 1980-2009 were used in the regression analysis. The first year, 1979, was dropped from the investigations due to a lack of prior winter information. The years 2010-2015 were dropped due to an apparent ecosystem shift (see Section '3.4 Oxygen Dynamics').

Referee comment:
Regression model. I think there are four flaws in the approach used here: 1. Not including loading and in-lake variables that would potentially describe interannual variability in productivity. 2) Including modeling results in a regression analysis. Given that the model does not simulate AF, it appears that using modeling results in the regression may just add noise to the regression or reinforce parameters that are in the model. 3) Using one correlation and one regression to simulate two very different types of lakes, and 4) Using way too many variables in a single multiple regression equation. Even though it appears based on stepwise regression all of the variables are significant, I think it is way over parameterized. Several studies have shown that with regressions using very few observations, many variables can look significant – with each variable coming in to describe one or a few unique observations. A good rule of thumb is to keep only 1 variable in a multiple regression for each 8-10 observations. So for this regression with 37 (and actually only 28 monitored years) observations, there should only be maybe 3 independent variables.
Referee suggestion: 1) include variables like actual loading rather than concentrations, include variables that describe inlake productivity (total phosphorus, chlorophyll, Secchi). I am not sure what GPP actually represents. If GPP does describe the changes in chlorophyll, it should be stated. I also do not think it is a good idea to include things describing DO (like maximum height of anoxia) when you are trying to predict AF (this can get to circular reasoning) 2) Only use the 28 actual observations in the correlations and regressions. 3) Look at the correlations for each part of the record (different biological conditions) separately. 4) Stick to correlations and not use regressions. Or if you do look at regressions start simple and add variables only significant when you consider the change in AIC.
Author response:
Thank you for your very thoughtful explanation of the regression analysis' flaws and your very helpful suggestions how to overcome these.

1) We changed the inflow variables, total phosphorus inflow concentration and total nitrogen inflow concentration (both in g per m2), to total phosphorus inflow loading and total nitrogen inflow loading (both now in g per d per m2 of lake area). These loading variables were included in the model to assess the importance of external hydrological drivers for the extent of anoxia. To capture in-lake productivity variables, our regression included the cumulative gross primary production in the surface and bottom lake that represent the total sum of photosynthesis, hence expressed as carbon uptake, of each functional phytoplankton group, and scales directly with in-lake Chl-a concentrations. Further, our regression also includes the temporal change of dissolved as well as particulate organic carbon in the bottom layer from stratification onset to fall mixing onset. To make it clearer what GPP represents, we added this sentence to the main text:

    L300-308: Here, GPP represents the sum of all functional phytoplankton group's photosynthesis rates parameterized as the total carbon uptake:

    $$f_{uptake}^{PHY_C} = R_{growth}^{PHY}(1 - k_{pr}^{PHY})\ \phi_{temp}^{PHY}(T)\ \phi_{stress}^{PHY}(X)\ min\{\phi_{light}^{PHY}(I)\ \phi_N^{PHY}(NO_3, NH_4 PHY_N)\ \phi_P^{PHY}(PO_4, PHY_P)\ \phi_{Si}^{PHY}(Rsi)\}[PHY] \quad (7)$$

    where the carbon uptake $f_{uptake}^{PHY_C}$ of an individual group *PHY* depends on the growth rate R_{growth}^{PHY}, the photorespiratory loss $(1 - k_{pr}^{PHY})$, temperature scaling $\phi_{temp}^{PHY}(T)$, metabolic stress $\phi_{stress}^{PHY}(X)$, and a minimum function taking into account limitations by light $\phi_{light}^{PHY}(I)$, nitrogen $\phi_N^{PHY}(NO_3, NH_4 PHY_N)$, phosphorus $\phi_P^{PHY}(PO_4, PHY_P)$ and silica $\phi_{Si}^{PHY}(Rsi)\}$ (Hipsey et al., 2017; adapted from Hipsey and Hamilton, 2008). As the GPP is the main model output variable for phytoplankton dynamics, it scales directly with biomass and Chl-a concentrations.

    Following the reviewer's suggestion, we removed the maximum height of anoxia in the regression analysis. The variable was removed from all paragraphs (2.3.4 Regression Model, Table 1)

2) For the regression we only used modeled results and no actual observed data. This was done to identify internal connections in the numerical model and its mathematical equations. Similar analyses of modeled output and model driver data were done in Farrell et al., 2020; Snortheim et al., 2017; Ward et al., 2020. We added these sentences to the Methods section:

    L279-282: All candidate predictors were either modeled output or boundary data for the model. This enabled the regression analysis to identify internal connections in the numerical model itself (similar analyses of modeled output and driver data were done in Snortheim et al., 2017; Ward et al., 2020; Weng et al., 2020).

3) Following our reasoning in the first comment and response section, and the suggestions by the reviewer we revised our regression analysis by only using model data from 1980-2009. This excludes the first year as warm-up period and the post-2010 period due to different ecosystem conditions (probably spiny water flea invasion). We added additional discussions regarding the ecosystem shift (see response to first comment).
4) Following the suggestion of the reviewer, we re-did the regression analysis with 21 candidate predictors using model output and model drivers from 1980-2009 (we removed the anoxia height from the sediment) using the Boruta algorithm (random forest classifier). This analysis identified 10 variables as important. Subsequently, we did a step-wise analysis of the AIC of each model. This resulted in the identification of seven predictors: HBR ratio during spring, HBR ratio during summer, Birgean Work in spring, epilimnetic GPP, Schmidt Stability in summer, Birgean Work in summer, and onset date of stratification. The AICs of each model with any of these variables removed did not result in significant changes (this table was added to the manuscript as Table A3):

Table A3 Step-wise model-selection by removing predictors of the multiple linear regression model using seven predictors.

Predictor	AIC
HBR ratio during spring *(Spring.HBR)*	-61.820
HBR ratio during summer *(Summer.HBR)*	-60.529
Birgean Work during spring *(Spring.Birgean)*	-60.189
Gross primary production in the epilimnion (*Epi.GPP*)	-58.952
Schmidt Stability during summer *(Summer.St)*	-51.829
Birgean Work during summer *(Summer.B)*	-50.848
Onset of stratification *(Onset.Strat)*	-42.900

We reduced the final model to only three predictors (as the reviewer suggested) including onset date of stratification, Schmidt Stability in summer (as the AIC was similar to Birgean but the concept of Schmidt Stability is more generally known) and epilimnetic GPP. The text in "2.3.4 Regression Model" was accordingly changed to:

> L321-330: This multiple linear regression model to predict Anoxic Factor included seven variables: HBR ratio during spring, HBR ratio during summer, Birgean Work in spring, Schmidt Stability in spring, epilimnetic GPP, Schmidt Stability in summer, Birgean Work in summer, and onset date of stratification. We reduced the complexity of the final multiple linear regression model to only three predictors of Anoxic Factor: onset date of stratification, Schmidt Stability in summer, and epilimnetic GPP. Schmidt Stability was included instead of Birgean Work as the resulting AIC of both models were similar, but the concept of Schmidt Stability is more commonly used in the limnological research community (Appendix Table A3). The final multiple linear regression model was configured as (scaled predictors, adjusted R^2 = 0.84, p < 0.001 Appendix Table A4).
>
> $$\hat{y} = 0.24 Epi.GPP + 0.54 Summer.St - 0.46 Onset.Strat - 5.44 * 10^{-17} + \hat{\epsilon}, \quad (8)$$
>
> where $\hat{\epsilon}\, N(0,38^2)$.

The results text in "3.5 Regression Model" was changed to:

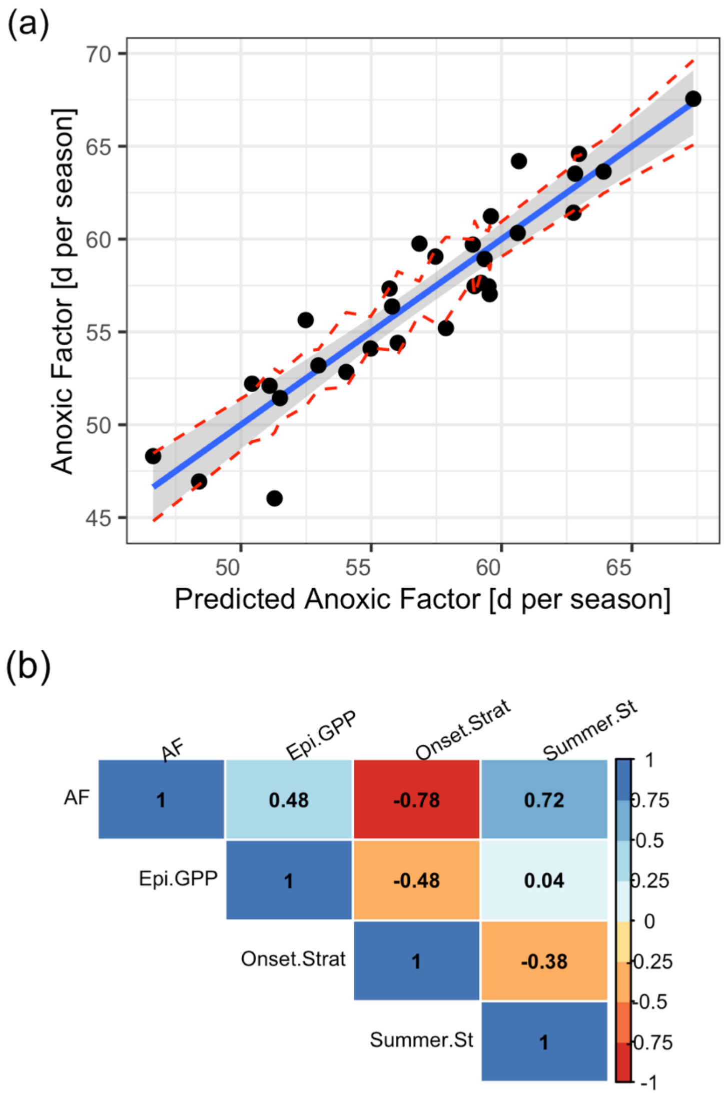
> L423-430: We included in total 3 predictors in our final multiple linear regression which were deemed important by the Boruta algorithm and stepwise linear model investigations using AIC for the period 1980-2009: Schmidt Stability during summer, the onset date of stratification, and gross primary production in the epilimnion (Appendix Table A4).

The linear model showed a good agreement between simulated and predicted Anoxic Factor (Figure 11 A, Appendix Table A4). The Anoxic Factor was positively correlated to the summer Schmidt Stability (r = 0.72, Figure 11 B) and the gross primary production in the epilimnion (r = 0.48). It was negatively correlated to the onset of stratification (r = -0.78, Figure 11 B).

We changed Appendix Table A3 (formerly A2) and Figure 11 accordingly:

Table A3 Most parsimonious multiple linear regression model (adjusted $R^2 = 0.84$, $p < 0.001$) explaining the summer Anoxic Factor.

	Estimate	Std. Error	t value	Pr(>|t|)	Rel. importance [%]
Intercept	-1.04e-15	5.70e-2	0.00	1.00	
Schmidt Stability during summer *(Summer.St)*	5.386e-1	7.920e-2	6.800	3.23e-7	43
Onset of stratification *(Onset.Strat)*	-4.581-1	9.006e-2	-5.086	2.68e-5	42
Gross primary production in the epilimnion (*Epi.GPP*)	2.436e-1	8.327e-2	2.926	0.00704	15



	AF	Epi.GPP	Onset.Strat	Summer.St
AF	1	0.48	-0.78	0.72
Epi.GPP		1	-0.48	0.04
Onset.Strat			1	-0.38
Summer.St				1

Figure 2 Predicted against simulated summer Anoxic Factor. A Linear model with a prediction which was done using a multiple linear regression model of the form: $\hat{y} = 0.24\boldsymbol{Epi.GPP} + 0.54\boldsymbol{Summer.St} - 0.46\boldsymbol{Onset.Strat} - 5.44 * 10^{-17} + \epsilon$,where ϵ $N(0, 38^2)$. The red lines represent confidence intervals. B Correlogram of the input data using Pearson correlation coefficients

We changed the following sentences in the main text to reflect these changes:

> L434-438: The Schmidt Stability during summer (rel. importance of 43 %) as well as the timing of stratification (rel. importance of 42 %) all influence Anoxic Factor, and are all driven mainly by atmospheric drivers and heat convection throughout the water column. The most important predictor of Anoxic Factor directly related to biological processes is gross primary production in the epilimnion (rel. importance of 15 %), Appendix Table A4).
>
> L596-599: Physical metrics – summer Schmidt Stability and onset date of stratification – were the most important predictors driving the summer Anoxic Factor. Although the gross primary production was still influential in affecting year-to-year variability of hypolimnetic anoxia, biological control over the Anoxic Factor was limited in our study period.

Referee comment:
My other main concern is that the deductive model seems to say that it is the inlake productivity that is driving the interannual variability in AF, and the other models seem to be saying it is driven by physics and sediment oxygen demand. Maybe with further analysis the models will come to more similar conclusions. If I am wrong with this interpretation, it should be explained better.
Author response:
The deductive model itself can only determine between two sources of depletion, either a volumetric one or an area sink. It cannot distinguish between biological or physical drivers of these depletion causes. Although the deductive model states that the volumetric sink is higher than the area sink, this is only of importance for the in-lake biological drivers (as the area sink depends on in-situ biogeochemical conditions). In the manuscript we state that the anoxia variability over a summer season is mainly driven by changes in the physical drivers, whereas we acknowledge that oxygen depletion itself (as shown in the regression model) is a function of biological and chemical activity. The deductive model itself does not consider any physical drivers, even diffusion is neglected. We added these lines to the main text to clarify our message:

> L529-532: We note that the simple deductive model itself can only differentiate between two sources of depletion and neglects any physical transport drivers of oxygen, e.g., diffusion. Therefore, the results of the deductive model only add direct information to the actual depletion process of dissolved oxygen, but not of the dominant drivers.

Referee comment:
1. Line-125. Very little information is given on the actual loading. Can these estimates be compared with others?
Author response:
Thank you. We compared our loadings with literature values, especially regarding phosphorus. Previous estimates range from about 15-67 t of total phosphorus (TP) per year (Kara 2011). Our estimates are at the higher end of this range. There is a concern that previous estimates did not fully account for loads of adsorbed phosphorus (hence, phosphate bound on sediment), because of the importance of extreme storm events on particulate loads (Carpenter 2017). To accommodate for a potential underestimation of TP loads, we added to the inflow boundary condition the adsorbed phosphate variables, which was set roughly equal in magnitude to non-adsorbed phosphorus. This puts our estimates of total P load near the upper range of previous estimates. Bennett (1999) estimated the long-term average annual TP input with 34 t P. Our Yahara inflow had an average annual TP load of about 25.3 t/y and ranged between 2.69 to 73.09 t/y over the period 1979-2015. Due to the use of a hydrological model, our inflows accounted for a closed water balance and included near-lake groundwater/spring inflows. Our average annual load of 25.3 t/y is slightly higher than the loadings by of Lathrop (2009). We added these lines to the main text:

> L136-144: To provide information regarding adsorbed soluble reactive phosphate, we doubled measured total phosphorus (TP) concentrations and applied specific ratios to individual phosphorus forms (Farrell et al., 2020; Snortheim et al., 2017; Weng et al., 2020). This put our estimates of TP near the upper range of previous load estimates. Bennett et al., (1999) estimated the long-term average annual TP load to be about 34 t, whereas our average annual TP load (with adsorbed phosphate) was about 50.6 t and ranged between 5.3 to 146.1 t (1979-2015). Our average annual TP load (without adsorbed phosphate) was about 25.3 t and ranged between 2.7 to 73.1 t (1979-2015), which is similar to previous estimates between 15 to 67 t (Kara et al., 2012). By doubling our TP by adding adsorbed phosphate, we accommodate a potential TP load underestimation due to the importance of extreme storm events on particulate loads (Carpenter et al., 2018).

Further, we checked our derived annual TP loadings using the Vollenweider model by assuming winter TP concentrations, TP_{lake}, of 140 ug/L, a residence time, RT, of 4 years, P retention, σ, of 0.7, and a mean depth, z_{mean}, of 12.8 m:

$$TP_{lake} = \frac{L}{z_{mean}\left(\frac{1}{RT} + \sigma\right)}$$

$$L = 0.14\ g/m3\ (12.8\ m\ (0.25\ y^{-1}\ +\ 0.7\ y^{-1})) = 1.70\ g/m2/y$$

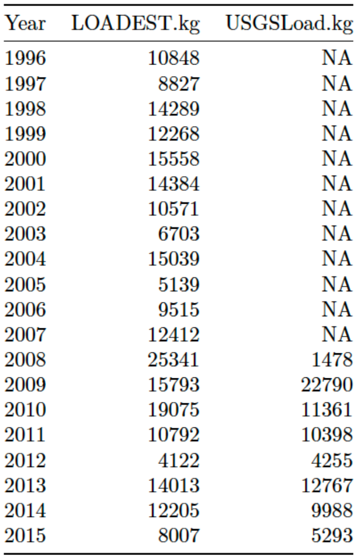
By multiplying L with the lake area of Lake Mendota (approx. 39.61 km2), the Vollenweider model quantifies the annual load for steady-state conditions with 67 t/y, which is slightly above our average annual TP load (with adsorbed phosphate) of 50.6 t/y.

Referee comment:

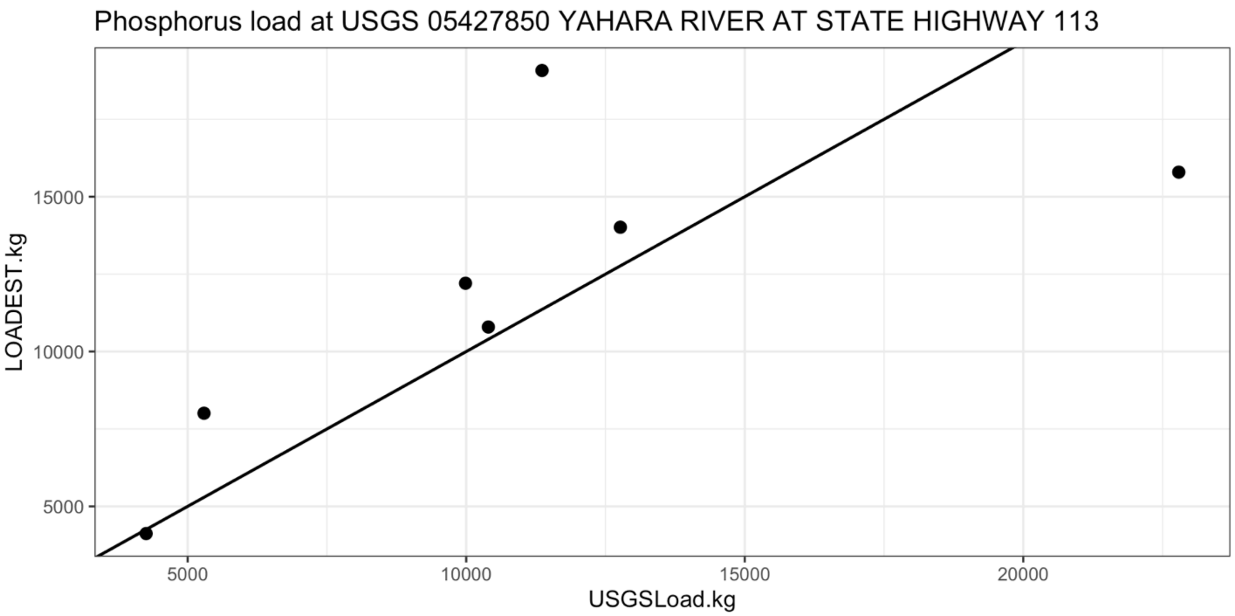
2. Line 128 – It says here to look at Weng et al. 2020 for a description of the loading regression, but when I look at that paper, I don't see any more than they used a regression, with no statistics either for the monitored sites or the watershed modeling.

Author response:

Thank you for pointing this out. Yes, there are no previous publications describing the regression fit analysis. We described in the previous response that our TP loads were near the upper range of previous estimates due to our addition of adsorbed phosphate (due to extreme storm events and land erosion). For the regression analysis between discharges and nutrient concentrations, we used the state-of-the-art loadflex R-package (https://github.com/USGS-R/loadflex, Appling et al., 2015). As monitored TP estimates are rare, a comprehensive statistical analysis is challenging. The attached figure visualizes the fit for 8 years (2008-2015), which was satisfactory for most years. USGS monitoring began Oct 2008, so a comparison cannot be made for the entire year. Overall, our model tended to overestimate nutrient loads into the lake.



Year	LOADEST.kg	USGSLoad.kg
1996	10848	NA
1997	8827	NA
1998	14289	NA
1999	12268	NA
2000	15558	NA
2001	14384	NA
2002	10571	NA
2003	6703	NA
2004	15039	NA
2005	5139	NA
2006	9515	NA
2007	12412	NA
2008	25341	1478
2009	15793	22790
2010	19075	11361
2011	10792	10398
2012	4122	4255
2013	14013	12767
2014	12205	9988
2015	8007	5293



Referee comment:
3. Line 136 – You mention other data earlier years, who collected that?
Author response:
The additional data points were measured by Patricia Soranno for her thesis. We added that information in the sentence:

> L149: The dissolved oxygen data set was complemented with historical measured dissolved oxygen data from 1992 to 1994.

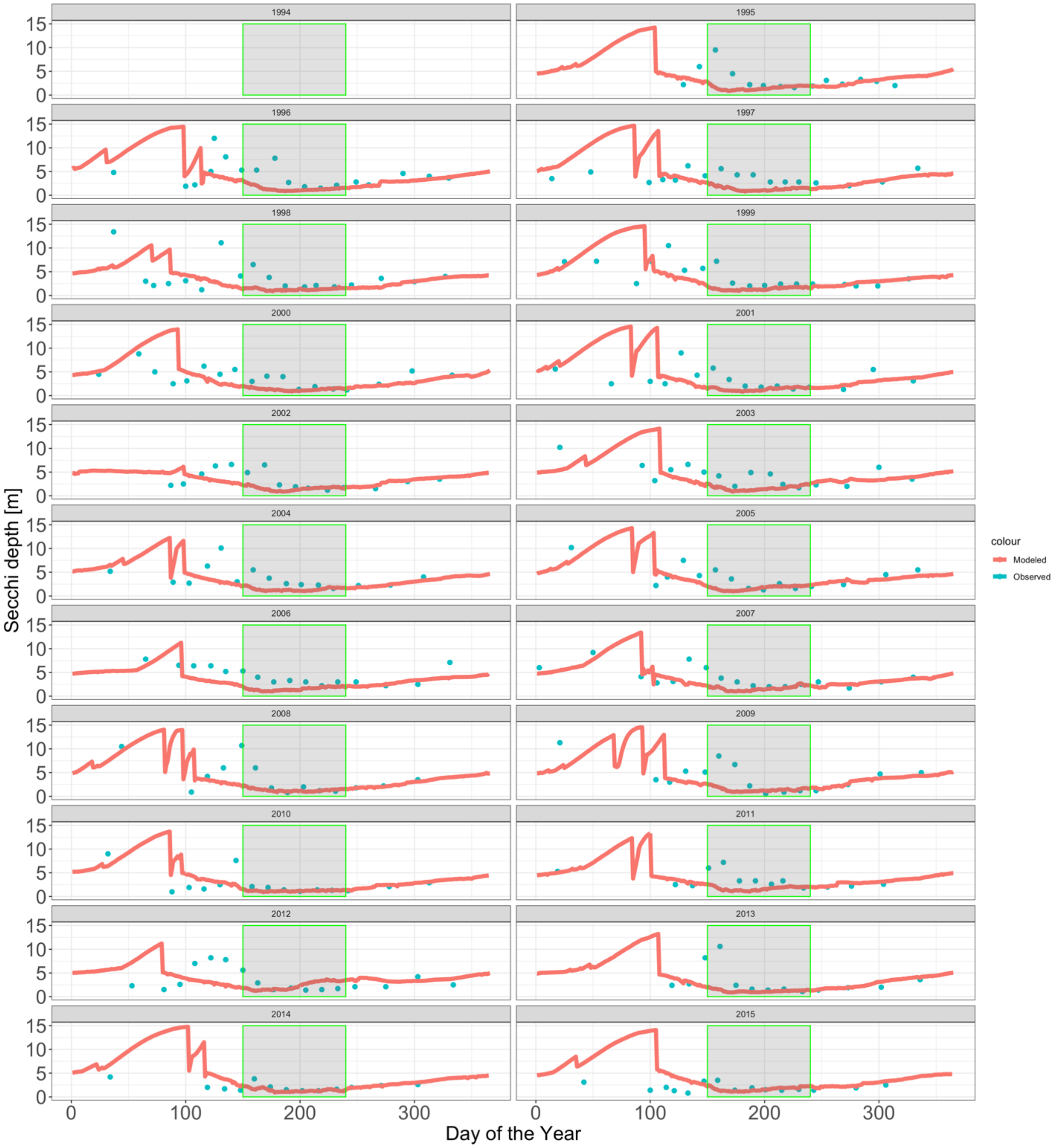
And we acknowledged her in the Acknowledgement section "We are thankful for supplementary dissolved oxygen field data from 1992-1994 by Patricia Soranno." Her data does not officially belong to the NTL-LTER monitoring data set of Lake Mendota, but it gave us valuable early spring-summer information regarding oxygen dynamics.

Referee comment:
4. Line 159 – See comments above about mixing real observations with modeled data. 5. Line 190 – There are lots and lots of parameters in AED, how did you narrow it down to the ones to start with, you need to start somewhere?
Author response:
We used the Morris Sensitivity Method to identify crucial parameters for the calibration. For this analysis we included the main model parameters regarding sediment flux and in-water biogeochemical reactions, mainly, of the main nutrient modules: oxygen, carbon, silica, nitrogen and phosphate. For the initial values, we chose starting values either from the AED2 webpage (https://aed.see.uwa.edu.au/research/models/aed/modules.html, default values, or values inside the typical range) or from previous modeling work on Lake Mendota (see Snortheim et al. 2017). We added this sentence to the main text for clarification:

> L231: Initial model parameter values were taken from default parameter values and ranges, as well as literature values (Hipsey et al., 2017; Snortheim et al., 2017).

Referee comment:
6. Line 215-Can you expect to capture interannual variability in productivity without having the phytoplankton simulate things specific to Lake Mendota?
Author response:
This is a good point, thank you for raising this. Although we did not calibrate the functional phytoplankton groups specifically to Lake Mendota, we still checked simulated Chl-a and Secchi depth values, as well as timings of phytoplankton bloom peaks. In general, the model did replicate the seasonal succession well. We've attached the

following figure that compares the observed to modeled Secchi depths for the reviewer to inspect. The summer Secchi depths from the model are similar to the observed ones, highlighting that the ecosystem dynamics during anoxia are similar. The gray boxes highlight the time period from day of the year 150 to day of the year 180 (June to end of August): for the majority of years the model can replicate the Secchi depth dynamics of the June period., whereas generally it underestimates the initial summer Secchi depth.



Referee comment:

7. Line 260 – My bet is that anoxia does occur under the ice, but you can't get that from one measurement during the winter.

Author response:

Yes, measurements and the monitoring have shown that there is anoxia under the ice in Lake Mendota, but it varies a lot. We agree that we cannot determine the full anoxia extent under ice with only one or two measurements per season. We changed sentence the sentence accordingly to:

> L276: We quantified the seasonal Anoxic Factor only for the summer season.

Referee comment:
8. Line 267 – Loads would be better than concentrations. Concentrations generally do not vary much from year to year. If you did really use loads, you should state that. But you should describe this better.
Author response:
We changed the inflow variables, total phosphorus inflow concentration and total nitrogen inflow concentration (both in g per m2), to total phosphorus inflow loading and total nitrogen inflow loading (both now in g per d per m2). These loading variables were included in the model to assess the importance of external hydrological drivers for the extent of anoxia. We changed the information in Table 1 accordingly:

Total phosphorus inflow loading	Winter/spring/summer of year n-1 and n	Extracted from driver data	g P per day per m^2
Total nitrogen inflow loading	Winter/spring/summer of year n-1 and n	Extracted from driver data	g N per day per m^2

Referee comment:
9. Line 273 – See comments above.
Author response:
Please see our response above.

Referee comment:
10. Line 278- Since Gross primary productivity (GPP) is your only in-lake productivity term, you should describe this in more detail. If this is directly related to chlorophyll, maybe this addresses some of my concerns.
Author response:
Thank for raising this point. GPP (gross primary productivity) in the lake is the cumulative photosynthesis, represented by cumulative carbon uptake per time step, of all functional phytoplankton groups. Therefore, it scales directly with the simulated Chl-a output. We clarified this in the main text:

> L300-308: Here, GPP represents the sum of all functional phytoplankton group's photosynthesis rates parameterized as the total carbon uptake:
>
> $$f_{uptake}^{PHY_C} = R_{growth}^{PHY}(1 - k_{pr}^{PHY})\ \phi_{temp}^{PHY}(T)\ \phi_{stress}^{PHY}(X)\ min\{\phi_{light}^{PHY}(I)\ \phi_{N}^{PHY}(NO_3, NH_4PHY_N)\ \phi_{P}^{PHY}(PO_4, PHY_P)\ \phi_{Si}^{PHY}(Rsi)\}[PHY] \quad (7)$$
>
> where the carbon uptake $f_{uptake}^{PHY_C}$ of an individual group *PHY* depends on the growth rate R_{growth}^{PHY}, the photorespiratory loss $(1 - k_{pr}^{PHY})$, temperature scaling $\phi_{temp}^{PHY}(T)$, metabolic stress $\phi_{stress}^{PHY}(X)$, and a minimum function taking into account limitations by light $\phi_{light}^{PHY}(I)$, nitrogen $\phi_{N}^{PHY}(NO_3, NH_4PHY_N)$, phosphorus $\phi_{P}^{PHY}(PO_4, PHY_P)$ and silica $\phi_{Si}^{PHY}(Rsi)\}$ (Hipsey et al., 2017; adapted from Hipsey and Hamilton, 2008). As the GPP is the main model output variable for phytoplankton dynamics, it scales directly with biomass and Chl-a concentrations.

Referee comment:
11. Line 281 – Consider dropping this whole paragraph.
Author response:
Although we understand the reasoning behind dropping this paragraph as the same results could probably be achieved by either starting with a simple linear regression model and extending it, or by step-wise analysis of AIC, we decided to keep the Boruta algorithm and analysis in the manuscript. This method allows us to analyze 21

potential predictors in a comprehensive framework before reducing the final number of important predictors by step-wise analysis.

Referee comment:
12. Line 306 – The major conclusion of the deductive model says that water column respiration controls oxygen depletion, yet everything else seems to point to physics. Am I missing something here?? Is water column respiration the cause and physics drives the variability in this? More explanation is needed.
Author response:
Please see our response above regarding the limitations of the deductive models and its incapability to acknowledge physical drivers.

Referee comment:
13. Line 322 – Please give the stats for DO. This is really what matters in this paper, especially in the part that varies from year to year.
Author response:
We agree, thank you. We added these sentences to the main text:

> L356-360: The simulated dissolved oxygen concentrations in the whole water column achieved an RMSE of 3.22 mg L-1, an NSE of 0.56, and an KGE of 0.77. Here, the average fits were better in the surface layer (RMSE of 2.77 mg L-1) compared to the bottom layers (RMSE of 3.31 mg L-1), whereas the temporal dynamics (as expressed in NSE and KGE) were slightly better in the bottom layer (an NSE of 0.64, KGE of 0.81) compared to the surface layer (NSE of -0.36, KGE of 0.47).

Further, the discussion of the oxygen fit has its own subparagraph in “3.4 Oxygen Dynamics” where we state that:

> L405-409: Dissolved oxygen dynamics, including the spatial extent of oxygen depletion in the water column, and the timing of summer anoxia periods, were replicated by the GLM-AED model (Figure 9A-B); although the model overestimated spring and summer time surface oxygen concentrations due to a higher net ecosystem production. The depth-averaged fit criteria of dissolved oxygen concentrations were similar to a recent study from Farrell et al. (2020) in which the RMSE were 1.88 mg/L and 2.49 mg/L in the epilimnion and hypolimnion, respectively, of a GLM-AED model calibrated for Lake Mendota.

Referee comment:
14. Line 333 – Reorder this paragraph to put the peaks later when you talk about summer.
Author response:
We agree. We moved the sentence to a later paragraph and combined it with the description of the annual course of Schmidt Stability:

> L395: Schmidt Stability peaked on average in August at approx. 720 J m-2 (Figure 6), followed by a peak in the Birgean Work at approx. 1250 J m-2.

Referee comment:
15. Line 345 – This paragraph could probably be deleted.
Author response:
As the main take-away message of our manuscript is related to physical drivers, we decided to keep this short paragraph describing the deep-water stagnancy in the manuscript. By comparing the additional energy demands of Lake Mendota with other similar sized lakes, the reader gets valuable information regarding the lake’s energy budget, and potential conclusions to the anoxia drivers of similar lake systems.

Referee comment:
16. Line 370 – It says the model captured annual anoxia events. Yes it described the annual development, but right now it does not seem to have any interannual capabilities??
Author response:
We quantified the correlation coefficient for the Anoxic Factor with r = 0.28 (total period), r = 0.45 (pre-2010) and r = 0.62 (post-2010), see also Figure 10. Especially for the pre-2010 period the p-value of the correlation coefficient was p=0.0591, which was slightly above the significance level. Overall, this highlights the model’s overall ability to predict interannual changes and dynamics.

Referee comment:
17. Line 374 – See above.
Author response:
See response above.

Referee comment:
18. Discussion – Need to tie all three model results together better. Right now two say physics and one says productivity.
Author response:
We disagree that two models point to physical drivers and one to biological ones. The deductive model distinguished the main oxygen consumption as either being a volumetric or an area sink term. This information was used to set up the sediment oxygen demand in the GLM-AED2 model. The results of the calibrated GLM-AED2 model were then used in a regression analysis to identify internal connections of the numerical model and its mathematical equations. This confirmed that in the process-based GLM-AED2 model three variables were important predictors of anoxia and its interannual variability. The deductive model itself does not consider any physical drivers (see responses above please).

Referee comment:
19. Line 394 – Although I completely agree with you, I am not sure where this comes from given the model results.
Author response:
The statement regarding that […] Biology matters but its interannual dynamics are not that influential […]" originates from the regression analysis. This analysis highlighted GPP as one of the most influential terms in projecting the variability of anoxia in Lake Mendota, but not as influential as physical variables (GPP only explained 15 % of the interannual variance of the Anoxic Factor).

Referee comment:
20. Line 420 – Again I agree with you, but other than one variable in seven in the regression, I don't know where this comes from. Need to describe this variables importance.
Author response:
As GPP is an ecosystem-scale metric that represents phytoplankton carbon uptake, net aquatic primary production as well as ecosystem respiration it surely highlights the biological control over Anoxic Factor, even if the regression deemed physical variables as more important.

Referee comment:
21. Line 425 – Maybe the lack of relations is due to using loading concentrations rather than actual loads. This is what I think the methods say.
Author response:
We changed the inflow parameters of total phosphorus and total nitrogen from concentrations to loadings and still the effect of anoxia is low. This is probably due to Lake Mendota's long water residence time of approx. 4 years.

Referee comment:
22. Line 433 – Is it loads or concentrations. If it is concentrations, that wouldn't surprise me at all. It is not the annual variations in concentrations that drive things, it is the difference in loads.
Author response:
See point above.

Referee comment:
23. Line 440 - This could be an important point, maybe there is so much oxygen consumption in the bottom, that it dwarfs any water column consumption. But this disagrees with findings of the other models.
Author response:
We discussed the sediment oxygen demand in the main text:

> L519-529: The simple deductive model established that the volumetric oxygen sink (i.e. water column oxygen demand) is consistently higher (on average about four times higher) than the sediment oxygen sink. The volumetric sink in lakes has been found to be strongly dependent on the trophic state of the lake, whereas the sediment sink is not (Rippey and McSorley, 2009). Eutrophic lakes tend to have high volume sinks that reach maxima of about 0.23 g m-3 d-1 (Rippey and McSorley, 2009) similar to the average volume sink of 0.16 g m-3 d-1 quantified by the deductive model for Lake Mendota. This finding is confirmed by the works of Conway (1972) who found that the high hypolimnetic oxygen demand of Lake Mendota was driven by algae decomposition, originating from the surface layer. Although eutrophic lakes tend to have a high sediment oxygen demand, the specific values can range from 0.3 g m-2 d-1 (Romero et al., 2004; Steinsberger et al., 2019) to extreme values of 80 g m-2 d-1 (Cross and Summerfelt, 1987), most studies measured or applied a value between 1 to 4 g m-2 d-1 (Mi et al., 2020; Veenstra and Nolen, 1991). The sediment oxygen demand calculated by our deductive model of 0.04 g m-2 d-1 was closer to the average value of approx. 0.08 g m-2 d-1 measured by Rippey and McSorley (2009) on 32 lakes.

In our numerical model the sediment oxygen demand (SOD) is replicating the volumetric and area sink as explained in the "Methods" section. Also, the model SOD is represented over the whole vertical axis (sediment area per volume for each grid cell) instead of a stagnant bottom layer only near lake's bottom. The results of the deductive model did not confirm a very high SOD compared to other eutrophic lakes, see extreme values in Cross and Summerfelt (1987) of up to 80 g per m2 per d.

Referee comment:
24. Line 445. The apparent changes caused by the Spiny water flea may be totally confounding any correlations, regressions, and your GLM-AED2 modeling. You may have to stick to one of the periods to really describe the effects of physics vs internal. Or have two different models.
Author response:
See our initial response please. We described the Anoxic Factors for the pre-2010 and post-2010 periods in more details and focused our regression analysis only on the pre-2010 period.

Referee comment:
25. Line 472. Rather than implementing a different type of dynamic model, maybe better capturing change in productivity and clarity, will help in describing the physics.
Author response:
We agree that a better replication of changes in ecosystem-scale metrics like productivity or even water clarity would improve the simulations a lot. Still, water quality models are generally way overparameterized and have problems regarding equifinality. The occurrence of tiny water flea has proven that ecosystem changes will have strong effects on other ecoysystem characteristics like anoxia. Therefore, even the best calibrated fixed water quality model will have problems replicating a dynamic ecosystem. Further, our monitoring campaigns do not capture important water quality variables on a high temporal scale, e.g. daily, which generates further uncertainty. Therefore, in our opinion an improvement of the hydrodynamic calculations for example by using a state-of-the-art turbulence closure scheme is the most applicable approach to improve the simulations in the near future.

Referee comment:
26. Line 481 – you didn't calibrate the biological parameters, so this should be rewrit-ten.
Author response:
We calibrated physical as well as chemical parameters in GLM-AED2 but did not modify the biological parameters of the functional phytoplankton blooms. As these functional variables represent multiple phytoplankton species, a direct calibration would potentially result in an over-calibration of the model for specific time periods, which we tried to avoid. We changed the sentence accordingly to:

> L551: Our GLM-AED2 model overestimated spring phytoplankton biomass, which resulted in an overestimation of surface dissolved oxygen concentrations.

Referee comment:
27. Line 497 – Rather than thinking the deductive model is biased, maybe it is the only approach capturing the effects of the biology.

Author response:
Please see statement above regarding the limitations of the deductive model, and the lines that were revised to better formulate this in the main text.