**General Response:**
Thank you for the revisions, which all have substantially improved our manuscript since the first submission. We felt the reviews were particularly thoughtful regarding the uses of alternative models to better understand and predict lake Anoxic Factor. The questions and criticisms raised by the reviewers led to a more careful and thorough description of our deductive modeling approach, which makes the results easier to understand and interpret. In addition, the reviewer points out a modeling situation that is easily overlooked – what we learn from the noise can be just as interesting as what we learn from the signal. The residual error from our predictions of Anoxic Factor shows an upward shift in anoxic factor in 2010, indicating either an unobserved change in drivers or an important process missing from the model at that time. While it is easy to be critical of the model for missing the shift, it perhaps is more important to use the missed shift to reflect on our knowledge of the ecosystem and to think of ways to attack this newly found problem. We are particularly grateful to the reviewers for encouraging us to pursue this line of reasoning.

The manuscript has improved substantially as an outcome of the review process. We thank the reviewers for their time and their valuable critiques. We hope the latest draft meets your expectations. Most of our response here is focused on the use of Chl-a and the description of the deductive model. Both of these have limitations, but we believe we have thoroughly addressed the reviewer's concerns. In the few cases where we do not incorporate their suggestions, we provide the details behind our decision making.

**Referee major comment:**
I think the authors should also try to use regression as an independent approach to further validate the results of the dynamic model. Normally I would prefer that the regression model ONLY use non-modeled information to make any conclusions; however, I realize that given the frequency of data collection this may be very difficult. Therefore, given that GLM-AED model simulates the physics quite well, I think the authors should run one more regression. That regression would use actual summer average Chl-a instead of GPP in the regression. This would show whether the lake actually behaves like GLM-AED says it does.

**Author response:**
We tested this approach, but unfortunately it did not yield the results we thought it might. Below we plot the relationship between the sum of measured vertical Chl-a concentration (winter to spring) against observed average Anoxic Factors (Fig R1, similar to the regression of modeled winter-spring GPP to summer Anoxic Factor). The relationship is weak. A linear model AF ~ Chl-a with the limited data from 2008-2016, returns $R^2 = 0.03$ with $p = 0.29$.

We believe three factors are at play (1) winter and under-ice measurements are rare, therefore we are potentially underestimating these values, (2) Chl-a, as a state variable, is only a rough proxy for the process, GPP, and (3) the observed Chl-a data on Lake Mendota for the period 2002-2007 is suspect due to, "an uncorrectable bias" of Chl-a data due to a change in instruments:
(see abstract information here https://doi.org/10.6073/pasta/f28e278afc34f1b7bd4f3cdc02b733a2).

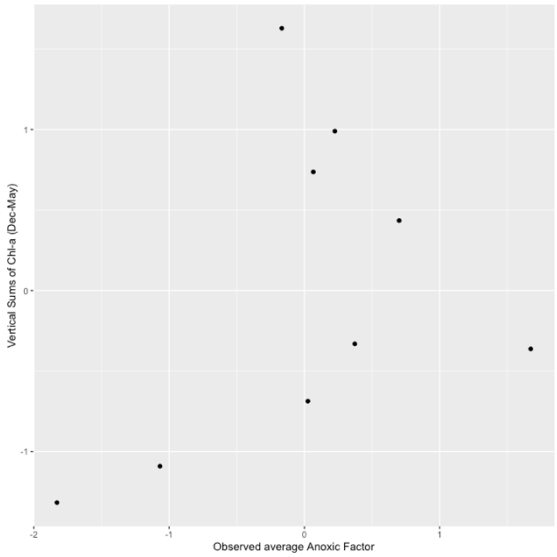


*Figure R1: Scaled observed average AF against scaled observed vertical sums of Chl-a prior to summer.*

Even though Chl-a is not a suitable predictor for AF, we agree that the influence of biology on AF merits emphasis in the manuscript. In the manuscript, the GPP prior to summer (winter to spring) was the main influential factor for summer anoxia (as stated e.g. L479: "Gross primary production (GPP) in the epilimnion prior to summer stratification is a secondary, but still important, predictor of anoxia." and in Table 1). To make this clearer, we revised all mentions of GPP as important predictor in text to "GPP prior to summer" in the manuscript. The settling of POC into deeper water layers prior to stratification is therefore the main process affecting anoxia, as stated e.g. at L480: "GPP fuels the sinking of particulate organic carbon (POC) into deeper layers before the establishment of a thermocline. In the hypolimnion, POC is readily decomposed into DOC and mineralized by bacteria in the numerical model, and reflects the dissolved oxygen volume sink."

**Referee major comment:**
My other concern is in the presentation of the results of the deductive model. First, describe how J(z) is actually computed, describe what we are seeing in Figure 3, use one year for example, and then describe what the volumetric part of this model really means. My first impression was that this model was giving very different results than the other two models. But I can see the volumetric part of the model may also represent variability in productivity and changes in stratification (although this is not described in the Discussion). I think my major confusion here is with the description of the model and Figure 3. Personally, I don't like any approach where I am interpreting the slope and intercept of noisy data. As data get noisy the slope goes closer to 0 and thus changes the overall interpretation.

**Author response:**
We revised the text in "2.3.1 Deductive Model" and added more information to it:

> L178: Using temporal and spatial linearly interpolated observed dissolved oxygen data, we applied the simple deductive oxygen depletion model according to Livingstone and Imboden (1996) in which the oxygen depletion rate *J(z)* at depth $z$ is conceptualized as
>
> $$J(z) = J_v(z) + J_A(z)\alpha(z), \quad (1)$$
>
> Where the intercept $J_V$ is the volume sink (mass per volume per time) representing organic matter mineralization processes, e.g. microbial respiration in the water column, the gradient $J_A$ is the area sink (mass per area per time) representing sediment oxygen demand, and $\alpha$ is a function for the $\alpha(z)$ratio of sediment area to water volume over the depth $z$ (Bossard and Gächter, 1981; Livingstone and Imboden, 1996):

$$\alpha(z) = -\frac{1}{A(z)}\frac{dA(z)}{dz}. \quad (2)$$

We used observed dissolved oxygen data from 1992 to 2015 (measured biweekly after ice offset) to calculate the specific oxygen depletion *J(z)* over depth for each year individually from the concentration, $[DO]_{spring}$, at the date of spring mixing offset, $t_{spring}$, to the date, $t_{2mgL}$, when oxygen concentrations, $[DO]_{2mgL}$, were below 2 mg L-1 (criterium for hypoxia):

$$J(z) = \frac{[DO]_{spring} - [DO]_{2mgL}}{t_{spring} - t_{2mgL}}. \quad (3)$$

Only dissolved oxygen data below a depth of 15 m were used. The derivatives of area to depth were approximated by using forward and backward differencing. The terms $J_V$ and $J_A$ were assumed to be constant for every year (assuming the hypolimnion to be homothermic) and were determined by using weighted linear regression.

We further revised "3.1 Oxygen Depletion Rates" to make it clearer that the volumetric and areal sinks are represented by the intercept and gradient, respectively:

L343: The derived annual oxygen depletion rates by the deductive model confirmed Lake Mendota's hypolimnetic anoxia as primarily driven by mineralization of organic matter. Observed oxygen depletion rates, J(z), and against area-volume ratios, $\alpha(z)$, were positively correlated for all years except 1993, 1997 and 2007 (Figure 3). For years with a positive relationship, the average intercept representing the volumetric sink JV as was 0.16 g m-3 d-1 and the average gradient representing the areal sink JA with was 0.04 g m-2 d-1 (adjusted R2 = 0.13, $p < 0.001$). Lake Mendota's hypolimnetic oxygen depletion was mainly driven by water column respiration mineralization processes over sediment oxygen demand. The annual volumetric depletions rate followed a normal distribution with an increase in the volumetric sink in recent years. The areal depletion rate distribution was positively skewed. An inspection of the residuals from the model fits indicates that the linear regression model may not be appropriate for some years, especially for values of the sediment to area volume ratio $\alpha(z)$ near 0.5 m2 m-3.

Regarding the description of net ecosystem production terms or physical drivers: First, we recognize that the observed data is influenced by physical as well as biogeochemical drivers. Further, the deductive model according to Livingstone and Imboden (1996) is based on the radon and phosphorus model from the Imboden and Emerson (1978) paper, in which all sink terms are described by the term *J*. But, as in the deductive oxygen model, production and vertical transport of dissolved oxygen are neglected, the volumetric sink (or the intercept in the linear regression), $J_v$, does mathematically only represent ecosystem respiration/mineralization in the water column (see also Charlton 1980, or Mathias and Barica 1980). The deductive model therefore can only describe negative aquatic ecosystem production processes, in which ecosystem respiration is higher than gross primary production. Vertical transport by i.e. turbulent eddy diffusion is neglected, therefore the volumetric processes do not represent the physics. Of course, the field data is influenced by stratification onset and the limitation of vertical fluxes, but the simple linear regression assumes that any changes in vertical fluxes are neglectable. Therefore, we decided to use the results from the deductive model as support for our sediment oxygen demand value in the process-based model, GLM-AED2. Additionally, we decided to discuss the results of the deductive model in "4.3. Biological Control over Anoxic Factor", as it can only quantify the biochemical oxygen sink terms from observed data.

Imboden, D.M., and Emerson, S. 1978. Natural radon and phosphorus as limnologic tracers: horizontal and vertical eddy diffusion in Greifensee. Limnol. Oceanogr. 23: 77–90.
Charlton, M.N. 1980. Hypolimnion oxygen consumption in lakes: discussion of productivity and morphometry effects. Can. J. Fish. Aquat. Sci. 37: 1531–1539.
Mathias, J.A., and Barica, J. 1980. Factors controlling oxygen depletion in ice-covered lakes. Can. J. Fish. Aquat. Sci. 37: 185–194.

**Minor Comments**

**Referee comment:**
1. Line- 21. Remove the word "evolutionary".
**Author response:**
Technically, the CMA-ES algorithm belongs to the group of evolutionary optimization algorithms that mimic biological evolution to find a global optimum for a given function. To avoid confusion, we agree that eliminating all mentions of "evolutionary" in the manuscript is warranted.

**Referee comment:**
2. Line 25 and later. I think the real strength in a regression model is to provide independent information that the dynamic model is simulating reality. See suggestion above.
**Author response:**
Please see our reply to the referee's first major comment above, in which we regress observational data per the referee's suggestion. Based on the limitations discussed in our first reply (data scarcity, potential bias), we used linear regression on modeled data as previously done in Snortheim (2017), Ward (2020) and Weng (2020). Here, all assumptions of the manuscript were done in model space and we recognize the constraints of the model, although we aimed to minimize potential bias by calibrating it to the best of our knowledge and data availability.

**Referee comment:**
3. Line 30. Make it read "a measured step upward".
**Author response:**
Agreed, we revised the text accordingly:

> L31: A measured step change upward in summer anoxia in 2010 was unexplained by the GLM-AED2 model.

**Referee comment:**
4. Line 48. There is a decadal shift in anoxia in Lake Mendota, and this should be brought into the final discussion a little better. This may be a major difference in what Snortheim described (line 60).
**Author response:**
Agreed. We revised the text in "4.3 Biological Control over Anoxic Factor" to discuss Snortheim et al. more and highlight the connection to spiny water flea invasion:

> L510: 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 (Supplement Figure A10), which was not captured by the model. A similar increase in observed Anoxic Factors starting in 2010 was also visualized in the study by Snortheim et al. (2017), but possible causes were not discussed. This The increased spatial as well as temporal extent of summer anoxia 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 ($p=0.032$) (Supplement Figure A9). Similarly, the pre-2010 observed Anoxic Factors were significantly different than the post-2010 observed Anoxic Factors ($p=0.0049$). 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 detection of this decadal shift in summer anoxia post-2010 highlights a hidden biological process that was not considered in the process-based model and 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 of Lake Mendota that focus entirely on ecosystem shifts associated with the invasion of spiny water flea in 2009 and the exponential growth of zebra mussels from 2015-2018 (Spear, 2020).

We added two sentences regarding the decadal change to "5 Conclusions":

L625: Further, our modelling framework detected a decadal shift in the Anoxic Factor starting in 2010, which was not replicated by our process-based model and therefore probably not driven by physical or chemical drivers, but related to an ecosystem shift caused by the invasive Bythothrephes longimanus.

**Referee comment:**

5. Line 83. Need to be careful here. Just because the model has high frequency output, it may not represent what is really happening in the lake. Empirically evaluating results of dynamic models may describe the mathematical equations in the model, but not how this particular lake actually works.

**Author response:**

We agree, but in this study we decided to evaluate emergent ecosystem characteristics by working in model space and calibrating the process-based model to best of our knowledge and data. We revised that line to:

L83: Results from deterministic lake models can be analysed using statistical models to derive general relationships of cause and effect in the model space.

**Referee comment:**

6. Line 92. You state that you are going to use data driven empirical models to evaluate observed data, that is really a good idea, and I think you need to do this more. Maybe by using Chl-a, you can get to this.

**Author response:**

Please see our reply to the referee's first major comment above.

**Referee comment:**

7. Line 96. I think you should add something about the decadal changes in Lake Mendota. Also state this in your Conclusions. Because the models don't capture it, it suggests something outside of the physics and chemistry is driving it. This is a strength of the overall approach.

**Author response:**

Agreed, we added a sentence regarding the decadal change to "5 Conclusions":

L625: Further, our modelling framework detected a decadal shift in the Anoxic Factor starting in 2010, which was not driven by physical or chemical drivers, but probably related to an ecosystem shift caused by the invasive Bythothrephes longimanus.

**Referee comment:**

8. Line 117. I still have a problem with PIHM-Lake never really being presented in this paper or published elsewhere.

**Author response:**

Unfortunately, a related publication about PIHM-Lake is still undergoing the review process. Therefore, we added multiple paragraphs to the supplement that explain PIHM-Lake in more details.

Supplement text:

**PIHM-Lake description**

PIHM-Lake is built upon a physically-based spatially distributed hydrologic model—PIHM (Penn State Integrated Hydrologic Model) (Qu and Duffy, 2007)—with the capability of simulating surface, subsurface, and channel water exchange between a catchment and a lake, as well as the water level change of the lake. As illustrated in Supplement Figure A11, PIHM-Lake model uses a finite volume numerical scheme and unstructured triangular mesh to represent the domain. It tracks the changes of surface and subsurface water storage on a 3D catchment and 1D lake as a function of precipitation, evapotranspiration, recharge, surface and groundwater flow, channelized flow, and snow melt. The spatial variation of overland flow and groundwater flow between the catchment and the lake is characterized by the water flows through the edges of each triangular mesh. Specifically, based on the conservation of mass of water, the generic form of the governing equations for PIHM-Lake is

$$\frac{dS_{canopy}}{dt} = vFrac * (1 - sFrac) * P - E_c$$
$$\frac{dS_{snow}}{dt} = sFrac * P - SM$$
$$\frac{\partial S_{surf}}{\partial t} = TF - \nabla q_{sw} - I - E_s$$
$$\frac{dS_{unsat}}{dt} = I - R - E_g - E_{gt}$$
$$\frac{\partial S_{sat}}{\partial t} = -\nabla q_{gw} + R - E_{sat} - E_{tsat}$$

where $\frac{dS_{canopy}}{dt}$ = the time rate of change of the canopy water storage, $S_{canopy}$ (m), due to canopy evaporation $E_c$ (m/day) and canopy interception $vFrac * (1 - sFrac) * P$ (m/day). $vFrac$ and $sFrac$ are the vegetation fraction and snow fraction, respectively. $P$ = precipitation (m/day). $\frac{dS_{snow}}{dt}$ = the time rate of change of snow storage $S_{snow}$ (m) due to $sFrac * P$:snow formation from precipitation when temperature is below 0 $^{o}$C (m/day) and *SM*, snow melt (m/day), which is a function of degree-day factor of ice and snow melt. $\frac{\partial S_{surf}}{\partial t}$ =the time rate of change of surface water storage, $S_{surf}$ (m), due to *TF*= throughfall (m/day), $\nabla q_{sw}$= net overland flow (m/day), *I*: infiltration (m/day), and $E_s$: surface water evaporation (m/day). $\nabla q_{sw}$ is modeled by the diffusion wave approximation of St. Venant's equation assuming shallow surface water depth and negligible influence of inertia force on overland flow, which is equivalent to Manning's equation. The estimation of infiltration rate is a function of the gradient of the surface and subsurface hydraulic head. $\frac{dS_{unsat}}{dt}$ represents the time rate of change of unsaturated water storage (m) due to *I*: infiltration (m/day), *R*: recharge (m/day), $E_g$: soil evaporation (m/day), and $E_{gt}$: transpiration (m/day). The recharge is calculated using Richard's equation assuming a vertical exchange of water across a moving water table interface. $\frac{\partial S_{sat}}{\partial t}$= the time rate of change of $S_{sat}$: the saturated water storage (m). $\nabla q_{gw}$ = net groundwater lateral movement between adjacent cells (m/day) is

represented by the Darcy-type flow proportional to groundwater gradient. $E_c$, $E_s$, $E_g$ and $E_{sat}$ are the evaporation (m/day) from the vegetation canopy, surface water, unsaturated and saturated soil zone, respectively. The potential evaporation rate is estimated by the Penman equation. The transpiration (m/day) is described by $E_{gt}$ or $E_{tsat}$, depending upon the vegetation coverage, the rooting depth and the groundwater table. If the groundwater table is higher than rooting depth, plants uptake water from the saturated zone, and $E_{tsat}$ applies. Otherwise, water uptake occurs at the unsaturated soil zone, and $E_{gt}$ applies.

For the hydrodynamics of the 1-D lake, we consider a two-layer system: a surface water layer and an aquifer layer. Surface water flow between the catchment boundary cells directly affects the water storage of surface water layer. Meanwhile, subsurface water flows through the aquifer layer and indirectly contributes to surface water through negative recharge. Likewise, based on the conservation of mass of water, the governing equation for the 1D lake component is

$$\begin{cases} \dfrac{\partial S_{lake\_surf}}{\partial t} = P - E_s - R + q_{sw} \\ \dfrac{\partial S_{lake\_gw}}{\partial t} = R + q_{gw} \end{cases}$$

where $\frac{\partial S_{lake\_surf}}{\partial t}$ = the time rate of change of lake surface water. $\frac{\partial S_{lake\_gw}}{\partial t}$ = the time rate of change of water storage in lake bottom aquifer. $P$ =precipitation (m/day); $E_s$ = surface water evaporation (m/day); $R$ = recharge (m/day). A positive value of R indicates downward lake surface water, while a negative value indicates an upward groundwater recharge to surface water; $q_{sw}$ and $q_{gw}$ are surface and groundwater flow through the edges of the lake boundary, respectively.

Details of the model processes and code is referred to the model repository: https://github.com/hydro-geomorph-zhang/PIHM-Lake.

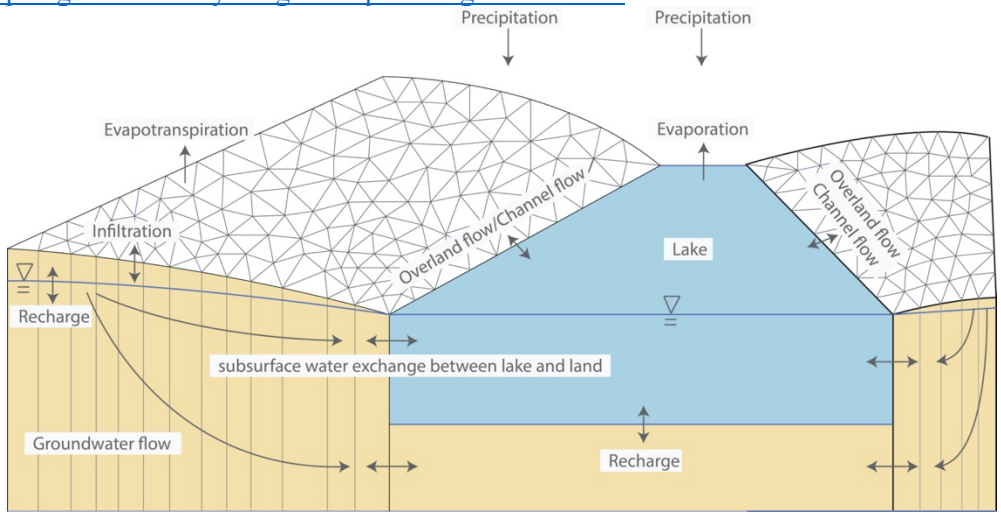


*Figure A11 Conceptual framework of PIHM-Lake.*

**Referee comment:**

9. Line 134. You have all kinds of nutrient components. I think you need to describe the assumptions you made to not only go from TP and TN to all of them. Why would you double the only thing that you actually measured (Line 137)?

**Author response:**

We used measured concentrations without any assumptions for the inflow loading regressions of all water quality variables except phosphorus (as described in the manuscript), refractory organic matter, dissolved inorganic carbon and silica. For these variables, inflow concentrations were not available, except for phosphorus, and so we used constant value for the loadings similar to the long-term averages measured in the lake. For phosphorus, we doubled the measured TP concentration to account for adsorbed phosphate, which is easily underestimated in manual sampling programs. Most studies underestimate these loads which become increasingly more important due to extreme storm events (see Carpenter et al. 2018).

**Referee comment:**
10. Line 139. Didn't Lathrop present measured/actual loading to Lake Mendota in several papers? Seems funny that those estimates are still not mentioned.
**Author response:**
Thank you for mentioning this point. We had the calculations by Lathrop on hand here but decided to focus on the estimates by Bennett as well as Kara. We added estimates of annual TP loads by Lathrop and Carpenter to the text:

> L141: 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 annual TP load estimates between of 15 to 67 t (Kara et al., 2012) and 10 to 80 t (Lathrop and Carpenter, 2014).

**Referee comment:**
11. Line 150. I think a reference is needed for the 1992-1994 data.
**Author response:**
The additional oxygen data were sampled during the graduate work of Patricia Soranno, but were not part of any publication to the best of our knowledge. We acknowledged her work and data, and added her Doctoral Thesis from 1995 as reference here:

> L149: The dissolved oxygen data set was complemented with historical measured dissolved oxygen data from 1992 to 1994 (Soranno, 1995).

We hope that in the near future we can add her data to the NTL-LTER data repository.

**Referee comment:**
12. Line 180. Describe how J(z) is actually computed and how Jv and Ja are estimated from the slope and intercept of the relation between J(z) and alpha. So does each point in Figure 3 represent a different depth?
**Author response:**
We revised the text in "2.3.1 Deductive Model":

> L178: Using temporal and spatial linearly interpolated observed dissolved oxygen data, we applied the simple deductive oxygen depletion model according to Livingstone and Imboden (1996) in which the oxygen depletion rate *J(z)* at depth *z* is conceptualized as
>
> $$J(z) = J_v(z) + J_A(z)\alpha(z), \quad (1)$$
>
> Where the intercept $J_V$ is the volume sink (mass per volume per time) representing organic matter mineralization processes, e.g. microbial respiration in the water column, the gradient $J_A$ is the area sink (mass per area per time) representing sediment oxygen demand, and $\alpha$ is a function for the ratio of sediment area to water volume over the depth *z* (Bossard and Gächter, 1981; Livingstone and Imboden, 1996):
>
> $$\alpha(z) = -\frac{1}{A(z)}\frac{dA(z)}{dz}. \quad (2)$$
>
> We used observed dissolved oxygen data from 1992 to 2015 (measured biweekly after ice offset) to calculate the specific oxygen depletion *J(z)* over depth for each year individually from the concentration, $[DO]_{spring}$, at the date of spring mixing offset, $t_{spring}$, to the date, $t_{2mgL}$, when oxygen concentrations, $[DO]_{2mgL}$, were below 2 mg L-1 (criterium for hypoxia):

$$J(z) = \frac{[DO]_{spring} - [DO]_{2mgL}}{t_{spring} - t_{2mgL}}. \quad (3)$$

Only dissolved oxygen data below a depth of 15 m were used. The derivatives of area to depth were approximated by using forward and backward differencing. The terms $J_V$ and $J_A$ were assumed to be constant for every year (assuming the hypolimnion to be homothermic) and were determined by using weighted linear regression.

**Referee comment:**
13. Line 225. Remove the word evolutionary.
**Author response:**
Agreed.

**Referee comment:**
14. Line 273. Is there any way to describe how you really combined simulated DO and measured DO to get the AF? Was the real data always used and then interpolated with simulated data?
**Author response:**
Yes, real data is always used and interpolated to determine the temporal and spatial extent of summer anoxia.
Method: Our observed data was bi-weekly during the ice-free period, therefore we needed to apply interpolation techniques to approximate DO values on a daily grid with a higher vertical resolution (as this matters for the determination of AF). Therefore, we used three different interpolation techniques, namely linear, constant and spline. In the final Fig. 10, modeled data were visualized as point values, whereas observed Anoxic Factors needed to be visualized as box-plots. We revised the text accordingly:

> L276: Observed Anoxic Factors were calculated by temporally and spatially interpolating bi-weekly monitored field data, using an ensemble of approaches (linear, constant and spline interpolation between neighboring data points). We quantified the seasonal Anoxic Factor only for the summer season, respectively for the modeled and observed data. We then compared the modeled Anoxic Factor (quantified by using modeled daily dissolved oxygen data profiles) against a set of observed Anoxic Factors (here, the bi-weekly data were temporally and spatially interpolated to get daily estimates over a finer vertical resolution) that were obtained by the application of three interpolation techniques.

**Referee comment:**
15. Line 278 and 324 and 421. Can you make this into two regression models? One the way you did it and one with Chl-a?
**Author response:**
Please see our reply to the referee's first major comment above.

**Referee comment:**
16. Line 317. Something to consider for the future. In the regression model add a variable to represent the change in time: 0 for the first half and 1 for the second half. Then you can see if the change was significant.
**Author response:**
Thank you very much for this very helpful suggestion!

**Referee comment:**
17. Line 337. See comments above about explaining Figure 3. It would help to state that 0.16 is the average intercept and 0.04 is the average slope from all of the figures. Remove the word respiration, this is what gets confusing. By removing the word respiration, then physics is still in this part.

**Author response:**
Thank you. We changed the text accordingly:

> L344: Observed oxygen depletion rates, J(z), against area-volume ratios, $\alpha(z)$, were positively correlated for all years except 1993, 1997 and 2007 (Figure 3). For years with a positive relationship, the average intercept representing the volumetric sink $J_V$ as was 0.16 g m-3 d-1 and the average gradient representing the areal sink $J_A$ with was 0.04 g m-2 d-1 (adjusted R2 = 0.13, p < 0.001). Lake Mendota's hypolimnetic oxygen depletion was mainly driven by water column respiration mineralization processes over sediment oxygen demand.

But vertical transport by i.e. turbulent eddy diffusion is neglected, therefore the volumetric processes do not represent the physics (see response to general comment at the beginning). Although we do recognize that the field data is influenced strongly by physical processes, the regression model mathematically does not incorporate these considerations.

**Referee comment:**
18. Line 343. Why would you add both pieces to get an estimate of SOD, shouldn't you only use the 0.04?
**Author response:**
In the GLM-AED2 model, the DO equation is mainly based on atmospheric exchange plus the sediment oxygen demand, which represents, in a conceptual way, the total oxygen sink over the water column. As bacterial mineralization in AED2 is based on temperature- and oxygen-dependence, we decided – conceptually – to use the SOD value of the model as the sink for oxygen, hence the main model compartment for the oxygen depletion rate. Therefore, we applied the total depletion rate, quantified by the deductive model, as the model's sediment oxygen demand, as "internal fluxes of organic carbon from the sediment back into the water column would drive additional oxygen depletion." (L355 in the main manuscript).

**Referee comment:**
19. Line 405. Should reference Table 2. I don't think your RMSEs are similar to that referenced – they are bit higher.
**Author response:**
Thank you for pointing this out. We revised the text accordingly:

> L413: 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, Table 2); 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 but slightly higher to a recent study from Farrell et al. (2020) […].

**Referee comment:**
20. Line 416. Rather than saying the AF has no significant differences, use the model not capturing things after 2010 as a strength and that there are decadal changes occurring in the lake.
**Author response:**
Thank you, we highlighted the detection of the decadal shift in anoxia in the discussion and conclusions (see other replies). Here in "3.4 Oxygen Dynamics" we highlight that the model's simulated Anoxic Factors were similar to the ones observed pre-2010, but significantly different in post-2010. We revised the text:

> L426: 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, Supplement Figure A9 A), whereas the distributions of observed mean Anoxic Factors and modeled ones after 2010 were significantly different (p-value = 0.032, Supplement Figure A9 B), highlighting a potential decadal shift in oxygen depletion patterns.. On the contrary, the modeled Anoxic Factor

distributions of the pre- and post-2010 period were not significantly different (p-value = 0.49, Supplement Figure A9 C), whereas the distributions of the observed Anoxic Factors were significantly different (p-value = 0.0049, Supplement Figure A9 D).

Additionally, we revised Supplement Fig. A9 to also highlight the differences between pre- and post-2010 modeled and observed Anoxic Factors, respectively:

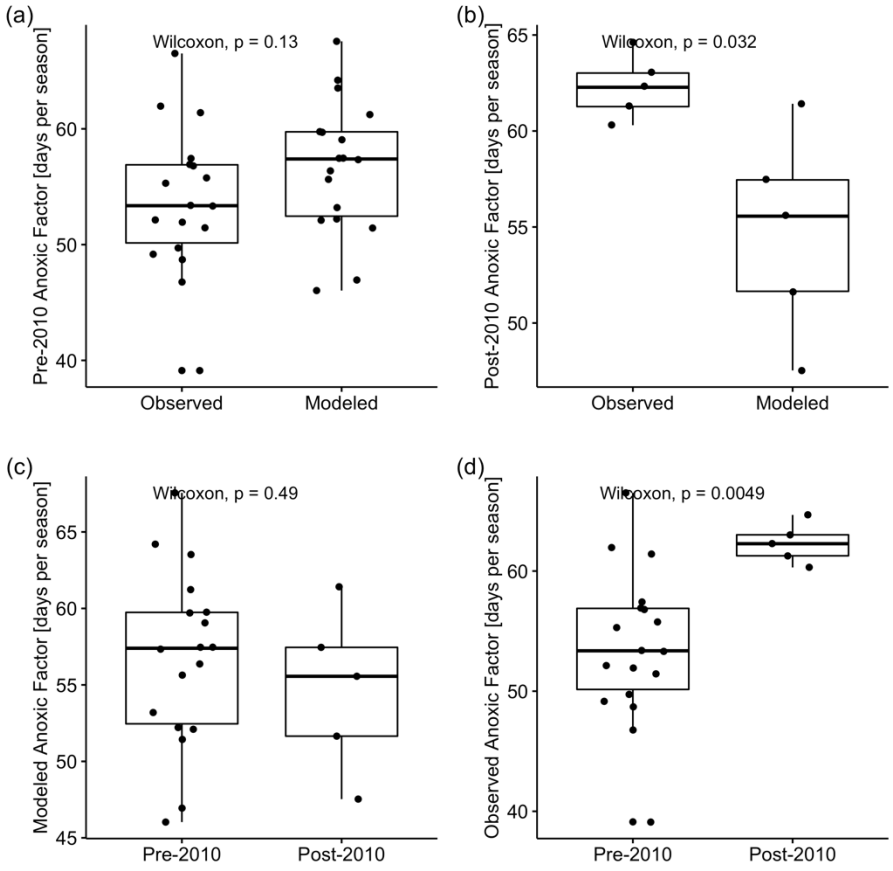


Fig. A9:

*Figure A9 Box-whisker plots of (a) observed to modeled Anoxic Factors for the pre-2010 period 1992-2009, (b) observed to modeled Anoxic Factors for the post-2010 period 2010-2015, (c) pre- to post-2010 modeled Anoxic Factors, and (d) pre- to post-2010 observed Anoxic Factors.*

**Referee comment:**

21. Line 432. Rather than ignoring the results of the deductive model, add a line here about it representing all volumetric processes including the physics.

**Author response:**

The deductive model according to Livingstone and Imboden (1996) is based on the radon and phosphorus model from the Imboden and Emerson (1978) paper, in which all sink terms are described by the term *J*. As in the oxygen model, production and vertical transport of dissolved oxygen are neglected, the volumetric sink (or the intercept in the linear regression), $J_v$, does only represent ecosystem respiration/mineralization in the water column (see also Charlton 1980, or Mathias and Barica 1980). Vertical transport by i.e. turbulent eddy diffusion is neglected, therefore the volumetric processes do not represent the physics. Still, we do recognize that physical changes in the system influence the relationships and measured concentrations of the observed data, and therefore i.e. stratification onset, plays an important role, although any physical transport is mathematically neglected by the simple linear regression approach. We use the

results from the deductive model as support for our sediment oxygen demand value in the process-based model, GLM-AED2. Additionally, we discuss the results of the deductive model in "4.3. Biological Control over Anoxic Factor", as it can only quantify the biochemical oxygen sink terms from observed data.

Imboden, D.M., and Emerson, S. 1978. Natural radon and phosphorus as limnologic tracers: horizontal and vertical eddy diffusion in Greifensee. Limnol. Oceanogr. 23: 77–90.
Charlton, M.N. 1980. Hypolimnion oxygen consumption in lakes: discussion of productivity and morphometry effects. Can. J. Fish. Aquat. Sci. 37: 1531–1539.
Mathias, J.A., and Barica, J. 1980. Factors controlling oxygen depletion in ice-covered lakes. Can. J. Fish. Aquat. Sci. 37: 185–194.

**Referee comment:**
22. Line 444. Change to timing and strength of stratification.
**Author response:**
Thank you, we changed the text accordingly:

> L457: Our work demonstrates that oxygen dynamics in Lake Mendota are strongly governed by the stratification strength and timing in the water column.

**Referee comment:**
23. Line 469. Hopefully Chl-a will show the same results.
**Author response:**
Please see our reply to the referee's first major comment above.

**Referee comment:**
24. Line 504. Add But this does show a decadal shift in the extent of AF.
**Author response:**
Thank you, we revised the text using the suggestion by the referee:

> L518: 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 detection of this decadal shift in summer anoxia post-2010 highlights a hidden biological process that was not considered in the process-based model and 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).

**Referee comment:**
25. Line 517. I really think the volume part of this model includes much of the physics associated with the volume of the hypolimnion and the length of stratification. This should be included. If you don't it really looks like this model gives a completely different interpretation.
**Author response:**
Please see our discussion of the deductive models' representation of physical processes at the beginning.

**Referee comment:**
26. Line 533. I really think you are being too hard on GLM. If you calibrated it better you should not have a consistent hypolimnetic bias. It has been shown to work well on many lakes, so I would not criticize it so hard. I really think the biggest problem was not calibrating the phytoplankton, by not doing that it affected many things. I think that is the number one thing for future model development. And the second thing would be trying to simulate the change in phytoplankton that occurred in 2010.
**Author response:**
We agree, and after investing seemingly years of our combined lives in modeling phytoplankton in Lake Mendota using GLM, we can say with certainty that it is very difficult.

We discuss several of these points in the manuscript, i.e. “improving the representation of phytoplankton and zooplankton dynamics in numerical models.” (L575), “[…] numerical representations of phytoplankton life cycles (Hense, 2010; Shimoda and Arhonditsis, 2016), and/or allometric scaling (Shimoda et al., 2016) could significantly improve numerical phytoplankton predictions” (L579). Our statement regarding GLM-AED2’s simulated discrepancies of hypolimnetic temperatures is rooted in a discussion of boundary conditions (“proximity of the atmospheric forcing boundary condition to the surface layers” (L556)) as well as the deep water mixing algorithm based on a vertical diffusivity approach instead of solving for turbulent diffusivity over the water column (like in a turbulence-closure scheme). Both points are essential part of GLM’s design philosophy and should not be interpreted as critic. We agree that your neglection of a thorough phytoplankton calibration is an important shortcoming on our site that hopefully follow-up studies will focus on. We revised the text to reflect that: (a) shortcoming of calibration, and (b) we would need more data to even do a calibration:

> L573: Discrepancies between simulated and observed Anoxic Factors, therefore, could be rooted in our simplifications of the phytoplankton dynamics and its model parameter calibration, and the related organic matter fluxes, and highlight the importance of improving the representation of phytoplankton and zooplankton dynamics in numerical models. Simulating a magnitude of individual species rather than functional phytoplankton groups has been shown to improve numerical water quality and ecosystem predictions (Hellweger, 2017), though it is unclear if it could improve spring bloom predictions in Lake Mendota. This depends also on a more extensive monitoring program that measures and specifies specific phytoplankton species over the vertical gradient on a regular basis.

**Referee comment:**
27. Line 578. I don’t see any reason why earlier stratification would cause as shallower thermocline. However, a warmer epilimnion could cause a shallower thermocline.
**Author response:**
This statement is grounded in Fig. 8c. We changed the text accordingly:

> L596: Further, a warmer epilimnion can cause the thermocline to become more shallow during the course of summer, which would cause the anoxia height to be spatially limited by a layer that is closer to the surface, hence more lake area would be anoxic. Increased oxygen depletion rates may also cause the anoxia height to be spatially limited by an earlier, and therefore lower, thermocline depth.

**Referee comment:**
28. Conclusions. Earlier you mention decadal shifts in the Abstract and Introduction. You found one using your models. This is a strength and should mention that by using GLM-AED you can say it was not driven by the physics, and it is probably driven by the changes in the biology.
**Author response:**
Thank you, we added a sentence to “5 Conclusions”:

> L625: Further, our modelling framework detected a decadal shift in the Anoxic Factor starting in 2010, which was not driven by physical or chemical drivers, but probably related to an ecosystem shift caused by the invasive Bythothrephes longimanus.