Response to reviewer comments "Improving estimates of ecosystem metabolism from dissolved oxygen time series", by M. W. Beck, M. C. Murrell, and J. D. Hagy III

The authors wish to thank the reviewers and associate editor for providing thoughtful comments on our manuscript. We have provided our response to each of these comments below, indicated in italics. Line numbers refer to the original manuscript. The reviewer comments have been shortened for brevity.

Reviewer 1:

I hope the authors can address this difficult question: does the model work better in diurnal tidal systems because the errors between the day and night conditions on the end-members cancel each other out? This is as opposed to the mixed tides where the errors can be heavily skewed towards respiration or production during the high high tide and thus lead to the large inaccuracies noted for Elkhorn Slough. This is an important point, because if the weighted regression approach is generating offsetting errors then it is not really working.

I think that the caveats described above have been acknowledged by the authors, such as starting on line 527, but I would recommend framing the issues associated with the end member problem in a very straightforward way.

We agree this is an important issue that deserves additional discussion. As noted, we have partially addressed these concerns on line 527 but have not explicitly described the "end member" problem. We have added text to this section that elaborates these points.

'A useful approach for conceptualizing instances when weighted regression may be most appropriate can be described using an "end member" paradigm. These end members are the terminal points of the water mass that is horizontally advected past the DO sensor. One end member is typically characterized by tidal marshes during flood tide, whereas the second end member may be characterized by oxygenated oceanic waters during ebb tide. An important distinction is made between end members that are influenced by actual biological processes that occur at each terminus and those that are not, where the former case is more likely a common occurrence at locations with large tidal amplitudes. In the latter case, variation in dissolved oxygen at the DO sensor is solely related to physical advection and weighted regression should work as intended provided collinearity with the solar cycle is minimal. At sites where biological processes affect the DO at the end members, the tidal characteristics of a site and the chosen window widths are critical determinants of the ability of weighted regression to isolate the true metabolic signal. Biological signals from diurnal or semidiurnal tides that exhibit rapid pregression with the solar cycle can be statistically isolated with relatively smaller window widths as the average (i.e., detided) signal can be quantified in fewer tidal cycles. Conversely, sites with mixed tides that exhibit prolonged synchronicity with the solar cycle, as at Elkhorn Slough, will require larger window widths to characterize the true metabolic signal. Issues of collinearity at mixed sites add additional challenges as previously noted. Caution should be used with larger window widths as the potential to oversmooth and mute actual interday variation is increased. For

diurnal and semidiurnal sites, relatively small window widths that produce noticeable differences in estimates may have meaningful implications for interpreting metabolism from filtered time series. We conclude that weighted regression has the ability to describe the true metabolic signal regardless of tidal characteristcs and biological processes that vary at the end members, although chosen window widths and collinearity with the solar cycle affect its application.'

Specific questions or corrections:

1) This may be a naive question - but is tidal height an accurate proxy for advection? It is intuitive that the higher the tidal amplitude the more advection that will occur over the period. However, it is also true that the fastest currents associated with tides occur in the mid-tide range, for example on a large ebb tide the highest currents (and presumably a significant part of the advection term) occur when the tide is actually near the mid-range in tide height. Equation 13 seems to contradict this fact, where the highest advection (DOadv) is associated with the highest tide measurement (H). This is in fact slack tide when very little advection is occurring.

This point was also mentioned by reviewer 2 (p. 8, lines 121-122) and we have added text to describe this issue in more detail. We agree that tide height is not the same measure as advection but weighted regression only requires a variable that can be mapped or linked to advection. This is described in more detail in the added text, beginning on line 123. We have also changed equation 12 to indicate that tidal height is simply a function of advection and may not be proportional (i.e., high tide may not be associated with large advection).

'An important distinction between tidal height and advection is that each variable could provide different information about a tidal regime. Tidal heights at the minimum and maximum of the range may be associated with periods of low advection when water masses are not moving rapidly past the sensor, whereas tidal heights near the mean may be more likely to have greater advection. Accordingly, our use of tidal height should not be confused with a variable that directly measures advection. The model only requires a variable that indicates a particular point in the tidal cycle, such that tidal height can be mapped to advection with quantifiable periodicity that can be isolated. '

2) The symbol for uncertainty in equation (6) and (7) and in the text on page 11 is not the same as the symbol used in figure 1, and both are different than at the bottom of page 13.

We have verified all equations, text, figures, and tables show the same same symbol, ε.

3) Line 247: should this be decreased process error?

No, we found that increasing process uncertainty actually improved the ability of the model to isolate the biological signal. Process error is serially correlated and we suspect that this correlation structure is characterized well by the model since time is used as an explanatory variable. This is a minor point because this effect was most pronounced when there was no

advective component and observation error was high in the simulated DO time series. In other words, there is no reason to apply weighted regression to an actual time series where advection effects are minimal.

4) Wind speed can have a significant influence on the calculation of DO flux, and therefore should be acknowledged as another source of error not accounted for by using a constant ka value. This influence is in addition to the advection term and can be significant.

I'm not sure where this comment applies in the text. If this is in reference to the k value in equation 11, we have changed the symbol to avoid confusion because this does not refer to the the same ka used for the reaeration coefficient in the metabolism equations. We did not consider any effect of wind in the simulated time series because it is not relevant for the analysis where the primary objective was to evaluate advective effects. Wind effects could be considered a component of uncertainty, although it was not addressed explicitly. The following was added for clarification (line 186):

'For example, wind events can affect air-sea gas exchange (Ziegler and Benner 1998; Caffrey et al. 2014) such that high wind may contribute to increased process uncertainty. Although this was not an explicit focus of the simulations, wind effects could be considered an implicit component of process uncertainty in addition to the effects of other unmeasured or latent variables that influence DO in a time-dependent manner.'

5) The different scales on Fig 6 and 7 for DO are confusing.

The figures now have the same scale for DO.

6) It would be beneficial to readers if the code for the analysis was distributed with the published paper.

we created a simple R package on GitHub for implementing weighted regression and estimating ecosystem metabolism. A link was added as supplementary information: https://github.com/fawda123/WtRegDO

Reviewer 2:

Do the authors plan to make their R code available? They mention several times how useful it could be to others (a point on which I certainly agree), but it wont be useful unless its available!

We have added a link in the supplementary information for an R package to implement weighted regression and ecosystem metabolism: https://github.com/fawda123/WtRegDO

General comments:

(1) The new method offers significant advantages for the types of systems in the NERRS

network, which are the focus of their study. The authors touch on this distinction briefly in lines 98-100, but I think they need to expand on the point and make it more prominent.

Text was added to elaborate this point.

Line 100: 'The method targets the periodicity of the tidal component as an explicit variable during model fitting, which allows the ability to isolate the biological component of the dissolved oxygen time series. As a result, the weighted regression approach can preserve the true biological signal rather than risking removal of both the biological and physical components as with traditional deconvulution methods. This approach offers a distinct advantage for estuaries where the magnitude of tidal effects on water quality observations can be severe.'

(2) I was generally confused, both in the authors application of their filtering method, and in the way they calculated their metabolic estimates, about the role of the third dimension (depth). Did the authors apply the same weight vectors and half-window widths to the data from each depth in a given system? Or, perhaps the authors chose the four systems they did for their case studies because the depth at those particular stations was shallow enough such that only one series of DO data was needed? If the latter case is true (i.e., data from only one depth used at each station), how would the authors propose applying their method in a deeper water column? Could it be?

Yes, only a single depth variable was used for each station with the assumption that the water column was vertically homogenous. This is a reasonable assumption for the case studies since all NERRS sites are considered shallow, productive estuaries. The method is then of course limited to only shallow water systems, but we argue for broad appeal as these systems are widespread across the globe, in addition to the 28 reserve systems within NERRS. Additionally, extending the method to stratified systems could be possible by applying the regression and metabolism calculations to each stratum, although we suspect that data are a limiting factor. This approach would require continuous monitoring of multiple depth layers, whereas most monitoring programs deploy only a single water quality monitor at the surface or weighted to the bottom. We are aware of specific examples where near-continuous vertical casts have been collected for some time (e.g., Mobile Bay National Estuary Program, http://mondata.disl.org/mondata/station_sensor.cfm?id=middlebay), but this is not the norm. We are certainly interested in applying the method to such datasets but are concerned that this may have limited appeal given the scarcity of data.

We have added some text to summarize the above points:

Line 259: 'NERRS is a network of 28 shallow, productive estuary reserves...'

Line 285: 'NERRS sites are typically shallow and vertically-mixed such that one water quality monitor is adequate for the entire water column, including the benthos.'

Line 445: 'The regression method and open-water technique may also have broad appeal for

application in estuaries that are shallow and vertically-mixed, such as those within NERRS. Extension to stratified systems is possible, provided time series across depth strata are available. The open-water technique assumes the water column is vertically mixed or that estimates from a surface sensor only apply to the upper layer in a stratified system (Staehr et al. 2010; Kemp and Testa 2012). Time series from different depths would be required to estimate metabolic rates that are vertically-integrated, followed by additional validation to ensure rates agree with expectations. In particular, application of the open-water method to each depth stratum would require an estimate of gas exchange across the pycnocline, whereas the technique implemented herein relied on a model for air-sea gas exchange at the surface. Meaningful metabolic estimates could be obtained using such an approach, following the application of weighted regression to remove tidal influences.'

Further:

(a) Unless I missed it, what was the method of integration by which the authors got from their volumetric estimates for each time step in equation (14) (specified in g O2 m-3 hr-1) to the depth- integrated (areal) estimates (in g O2 m-2 t-1) that appear in all of their figures and tables?

Text was added for clarification on line 307: 'Finally, volumetric rates were converted to depth-integrated (areal) estimates (g O_2 m^{-2} d^{-1}) by dividing by the mean water-column depth from the water quality sensor. A half-meter was also added to account for approximate placement of the sensor slightly off of the bottom.'

(b) I am concerned (or perhaps just confused) about the method the authors chose for deriving daily estimates of P and R from the hourly fluxes calculated using equation (14). The description of the method is not clear. Did the authors (1) calculate the individual hourly fluxes, then group them by day or night, then average them, and then multiply each by the duration of the day or night, or (2) calculate the individual hourly fluxes, then divide them by day or night, then add them together within their respective groups? Both methods have been used in the DO time- series metabolism literature; the second may be better since it doesnt assume the rate of primary production is constant over the course of a given day.

We used an approach identical to Caffrey et al. 2014 to average DO flux in each day/night period, which is then multiplied by different time periods to estimate production and respiration. In honesty, the comment is slightly confusing because the two approaches as described would produce an identical result. The first describes averaging DO flux within day/night groups then multiplying by the respective time in each day/night period, while the second describes a summation of DO flux within each day/night group. We did a simple spreadsheet calculation using both approaches to verify an identical result. For the purpose of the comment, we have revised the text to clarify the method:

Line 300: 'The diffusion-corrected DO flux estimates as hourly rates of DO change were first averaged during day and night periods for each 24 hour 'metabolic day' in the time

series. The 'metabolic day' was considered the approximate 24 hour period between sunsets on two adjacent calendar days. Hourly DO flux was considered respiration during night hours and net production during day hours. Total respiration (Rt) rates were assumed constant during day and night such that daily rates were calculated as the average DO flux during night hours multiplied by 24. Daily gross production (Pg) was the average DO flux during day hours minus the average DO flux during night hours, multiplied by total sunlight time. Net ecosystem metabolism was gross production (positive) plus total respiration (negative).'

(3) Notation and terminology:

(a) The use of Pg to denote gross primary production is very confusing, especially since the units of Pg are in g (and Pg also being the SI unit for petagrams). Perhaps something else would be more appropriate? In addition, what is the t in Rt as the notation for respiration? Total? If so, the authors should specify when they introduce their variable (unless they did, and I missed it?).

These are standard terms that describe gross production and total respiration (see Caffrey et al. 2014, references therein). The above revision in response to the previous comment defines the acronyms accordingly.

(b) The word aggregation is used throughout the text. I gather from my reading of the ms that the authors mean binning or averaging, over some time or spatial scale. Aggregation is not a particularly precise term, because it could mean many different things. Perhaps the authors could consider replacing this word in the ms with a word or phrase that describes what they mean, e.g. time averaging.

We have noted all instances in the text where 'aggregation' is used and revised for clarity, mostly replacing with 'time averaging' or greater detail as needed.

(4) Duplication of presentation. If space is an issue, it seems as though Figs. 3 and 4 present the same data as Tables 1 and 2. I know heatmaps are all the rage right now, but I honestly felt the specific correlation coefficients in the tables told me a lot more than the heatmaps.

We have retained the figures because they do not show identical information, although similar. The figures provide an overview of the correlation and error associated with each unique combination of simulation conditions (Fig. 3) and window widths (Fig. 4). The tables provide a distribution of the correlation and error for each unique parameter where a criteria applied. For example, the first row in table 1 shows the distribution of the correlation/error for all simulation results when the diel DO component was set at zero. This information is a summary whereas all values are shown in Fig. 3 as 27 tiles where the diel DO component was zero.

Text was added to figure/table captions to emphasize these differences. For example, figure

3 caption now includes 'See Table 1 for a summary of combined results for each unique parameter', whereas Table 1 now includes 'See Fig. 3 for results of all parameter combinations.'

Specific comments:

Abstract generally: If the authors could work in a sentence on the significance of their method vis--vis traditional time-series filtering techniques (see my comments above), I think they could increase the impact of their work. Additionally, I would have liked to see at least a few numbers or figures in the abstract (e.g., by what % on average did the method reduce variance in the NEM estimate relative to those calculated from the unfiltered data, or how much did it reduce RMSE?).

The following was added to the abstract: 'The method targets the periodicity of the tidal component while preserving the true biological signal, offering a distinct advantage over traditional deconvulution methods.'

Percentage reductions were also included: 'Variability from advection was reduced on average by 70.2% for production and 74.3% for respiration.'

p. 3, line 12: integrated by what dimension and over what interval? Depth? Time?

This term was removed as the details are now available in the text.

p. 3, line 22: useful is nonspecific and does the authors work a disservice; can they think of something more specific and impactful?

The above response includes percentage reductions to increase specificity/impact.

p. 3, lines 23-24: See my comment above regarding the methods performance when the solar cycle and period of tidal height change were highly collinear. The method yielded estimates at the Elkhorn and Padilla sites which were significantly different from those obtained with the unfiltered (traditional) data. By the authors own admission, neither the filtered nor unfiltered data probably accurately represent what is truly happening at those sites. A sentence specifying those situations in which the method did not perform particularly well would be more honest, rather than simply saying when it generally did work well.

The sentence was modified as follows: 'The model was especially effective when the magnitude of tidal influence was high and correlations between tidal change and the solar cycle were low, whereas collinearity in the latter instance limited model performance.'

p. 3, line 26: timescales

Changed.

p. 4, lines 32-33: Reference to Gaarder and Gran (1927) missing here.

Citation added.

p. 4, lines 42-43: This simply isnt true. There are dozens of reasons bottle incubations are more effective in certain situations, depending on the research objective. Open-water measurements may be more effective for evaluating certain kinds of events or trends, but if the authors are going to make this claim, they should specify what they mean.

The sentence was revised: 'Open-water estimates also provide a more accessible means of tracking ecosystem change over time as compared to discrete sampling events with bottle-based approaches.'

p. 4, line 46: Reliable estimates

Changed.

p. 5, line 67: advection, leading to

Comma added.

p. 6, line 81: example illustrates

Changed.

p. 7, lines 98-100: See my comments above about the methods significance vis--vis traditional detrending methods. An expanded discussion of the difference here is warranted.

See the response to the first general comment, i.e., the additions to the final paragraph of the introduction.

p. 7, lines 101-102: evaluate the ability

Changed.

pp. 7-8: Appropriate place for clarifying role of depth, or why it is being ignored (my comment above).

Yes, see the response to the second general comment. Also, the response to the next comment addresses the role of tidal height/depth vs. advection as a model variable.

p. 8, lines 121-122: I buy the idea, but is there some justification in the literature for using tidal height as a proxy for advection?

See our response to reviewer 1 about the use of tidal height as a proxy for advection, i.e.,

the expansion of the paragraph on page 8... 'An important distinction between tidal height and advection...'

p. 9, lines 139-140: I was generally confused about why different time vectors were used to represent days and hours (or is it hour of day) separately. Maybe just some clarification of terminology would help.

The weighted regression model evaluates weights associated with distance from the center of the window in days as well as hours within the day. For example, if we care about modelling oxygen at 3 pm, we would weight all other observations close to 3pm in other days with high importance, with further diminishing importance as window distance increases in days from the center. The weights widget in the supplementary information can be used to view the influence of both, in addition to the tidal weights. The citation at the end of the paragraph for the multimedia section should help interpretation. We also revised the line... 'Windows for time (continuous throughout the day) and hour (within each day) are used...'

p. 12, line 196: What is the justification for using this particular range of O2 concentrations to generate the simulated data set?

The values were chosen based on a general approximation of the observed data from the case studies, text was added... 'as a rough approximation from the case studies below'

p. 12, line 204: lunar

Changed.

p. 13 ff: Equations appear to employ the wrong del for the partial derivative, i.e. versus

This was changed to summation.

p. 14, line 237: The authors are characterizing a corr. coefficient of 0.63 as similar? Or did I miss something? On what specific basis in Tables 1 and 2 is the claim of similarity made?

'Similar' may have been too strong of a description for these results. Instead, we suggest that the filtered time series from simulation results produced a 'reasonable representation' of the true biological signal for most scenarioes. Poor results were only obtained for scenarios when advection was minimal and process uncertainty was high, which was previously noted on lines 450-451.

p. 17, lines 299-300: Related to my comment about depth, above. To get from a piston velocity to a volumetric reaeration coefficient (isnt this just a depth-corrected piston velocity - why the need for even more jargon?), one must invoke some mixed layer depth. What was the depth (i.e., H in Thbault et als equation A11) used in each case? The Thbault paper seems to contain a not-so-unique method of calculating gas transfer velocity (and air-sea flux), gussied up with new terminology for some of the parameters.

We have not changed the terminology to reduce confusion between our methods and those we cite. This work builds on the metabolic approach described in Caffrey et al. 2014. Ongoing work here at EPA and elsewhere will further develop these methods. We anticipate that interested readers will consult our references for more details on the exact techniques. Further, we have clarified the role of depth in the response to the previous comment about areal rates.

p. 17, lines 300-307: The precise method of obtaining the daily ecosystem metabolism estimates was not clear. (See my general comment 2, above.)

See our response to the second general comment.

p. 18, lines 319-321: I think this assumption is valid, but can the authors provide some justification or precedent? One citation would be sufficient.

We have unpublished data that generally support this claim, unpublished citation added. To our knowledge, no studies have examined these relationships at the level of detail described in our manuscript so we cannot provide a peer-reviewed citation.

p. 19, lines 350-351: I disagree. These percentages from the filtered data do represent a decrease in the number of anomalous values by a factor of 3, but they are not near zero. In one case, 7.12% of values were still anomalous.

Sentence was revised: 'Anomalous values were substantially reduced for all case studies...'

p. 23, line 427: Particularly confusing use of the word aggregated. Aggregated or averaged by what, and over what time or spatial interval?

Sentence was revised: 'Results for Sapelo Island suggested that time averaged estimates, either as monthly or annual means, were comparatively unchanged by filtering...'

Discussion generally: The discussion read less like a discussion than a simple summary of the rest of the paper. Perhaps a separate discussion isnt really necessary?

The author guidelines for L&O methods indicate that the discussion should describe the degree to which the method meets defined needs, potential for new insight, comparison with alternative approaches, and new questions that may have been raised (http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1541-5856/homepage/ForAuthors.html#6). We argue that the current discussion adequately covers these points so we have not removed it from the manuscript.

p. 23, lines 440-441: remove some variation; the method doesnt remove all variation

Changed, though we have used 'a significant amount' instead of 'some'.

p. 26, lines 500-503: This seemed like it (or some similar text) belonged in the methods section, to describe why the sites were selected. In addition, the grammar of this sentence was confusing.

Sentence was revised and moved to line 269 in the assessment section: 'Many daily metabolism estimates were also not interpretable (i.e., 'anomalous') from these sites using standard open water methods.'

p. 27, lines 527-535: This paragraph belongs in the discussion section. Perhaps after the paragraph ending on line 479.

Moved.

p. 28: Will code be available publicly somewhere?

Yes, see the comment above and in response to reviewer 1.

Comments on figures and tables:

Figure 1 ff: One of my few questions about the simulation datasets centers on the way biology was parameterized. I appreciate that modulation in the DO signal due to biological processes was modeled as a purely sinusoidal function for purposes of evaluating the method, but even the purest biological signal will not be sinusoidal. Major differences are (1) that the peak of the biological curve will usually lag each days peak insolation by some number of hours (evident in the authors own data in Fig. 6) and that (2) there is a point of photoinhibition, where the perfect shape of the curve is dented by a decrease in photosynthetic activity before it begins its actual decline. Ive not put either of these very eloquently, but I assume the authors know what Im speaking of. Some brief acknowledgement in the methods section about the representation of the biological signal as a perfect curve like this would be instructive.

Text was added in the methods to clarity this point (line 196): 'Although the representation of DO_{die} as a simple sine wave does not account for lags with the solar cycle or periods of photoinhibition that may be observed in actual time series, we consider the approach sufficient for the simulations.'

Figure 6: What was the threshold PAR intensity used for defining the PAR shading? What are the units of PAR? Millmoles of what? Photons? Microeinsteins (E)? Or W m-2?

Information added to figure legends.

Figure 8 ff: One effect of filtering with the authors method is that it appears to dramatically attenuate the interday variation in metabolic parameters evident in unfiltered data. Remember that the discovery of this temporal heterogeneity has been one of most widely celebrated features to emerge from analysis of high-frequency time series (of, e.g.,

D.O.). What is the implication of the authors method for the existence (or not) of this heterogeneity? Does this mean there really isnt as much temporal heterogeneity in ecosystem respiration and in photosynthesis as this method has suggested? Some discussion would be instructive (perhaps replacing some of the repetitive text in the current discussion); the authors wont have to look far into the literature to find many studies that have remarked on the apparent heterogeneity in unfiltered data.

This is a valid point that we discussed at length during manuscript preparation. The degree of smoothing using filtered results to estimate metabolism is affected by the choice of window width, such that larger widths have the potential to oversmooth. However, there is a tradeoff between the need for larger window widths and the characteristics of the dataset. Smaller window widths are effective for sites with rapid pregression of the solar cycle with tidal variation, such that reduction of true biological variability (ie.g., interday variation) is less likely. Conversely, larger window widths are needed for sites with mixed tides, although there is greater potential for oversmoothing and reducing interday variation that is biologically relevant. We provide cautionary recommendations in the final section, including implications for resolving interday heterogeneity. Text that precedes the addition provides additional context (see the response to the first general comment from reviewer 1).

'Caution should be used with larger window widths as the potential to oversmooth and mute actual interday variation is increased. For diurnal and semidiurnal sites, relatively small window widths that produce noticeable differences in estimates may have meaningful implications for interpreting metabolism from filtered time series. '

Table 3: Were the metabolic rates given here those calculated using the authors method?

These were based on our methods, without filtering. Text was added to the footnote for clarity: '...estimated using methods described herein with observed data'