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 at a minimum. For other 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, mixed tides that exhibit prolonged periods of synchronicity with the solar cycle, as was observed for 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. 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 choosing appropriate window widths and evaluating collinearity with the solar cycle are critical factors that must be considered in 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 some additional text to describe this issue in more detail. We agree that tide height is not the same measure as advection but all that is needed for weighted regression is 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 the two variables 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 is directly proportional to 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 the model can isolate. '

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 somewhat of a minor point though because this was most pronounced when there was no advective component in the simulated DO time series and observation error was high. 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 either process or observation 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 simulation analyses, wind effects could be considered an implicit component of process uncertainty in addition to the effects of other unmeasured 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 that provides code and examples 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 that provides instructions and examples for installing a simple R package to implement weighted regression: 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 a separate parameter 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 in the output 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). The authors adapted the Hirsch et al. WRTDS method for a laminar or channel flow case (rivers and streams) for use in estuaries, where the regime is certainly not channelized or laminar. How was the method used to filter data from multiple depths simultaneously in the four case study systems? 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 this was the case, are the authors making the assumption that the water column in which the series of DO measurements was made at each of these stations was vertically homogeneous with respect to the processes and properties of interest over the entire period of measurement? Do the authors assume that the parcels of water flowing by the sensor in each case were vertically homogenous with respect to DO and also with respect to the strength of the biological processes (photosynthesis, respiration) that determine DO, even as the tide changed? I can envision reasonable justifications for each of these, but the authors must address them. 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? The method has a serious limitation if it cant be used for a water column deeper than a few meters.

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 were chosen specifically as shallow-water, productive estuaries. The method is then of course limited to only shallow water systems, but we argue that it has 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 strata, 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-continous vertical casts have been recorded 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 256: 'NERRS is a network of 28 shallow, productive estuary reserves...'

Line 295: 'NERRS sites are typically shallow and vertically-mixed such that one water quality monitor is considered to represent the entire water column, including the benthos.'

Line 465: 'The method may also have broad appeal for application in estuaries that are shallow and vertically-mixed. Extension to stratified systems could be possible, provided water quality time series across depth strata are available.'

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?

(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.

(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?).

(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.

(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.

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?).

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

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

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.

p. 3, line 26: timescales

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

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.

p. 4, line 46: Reliable estimates

p. 5, line 67: advection, leading to

p. 6, line 81: example illustrates

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.

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

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

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.

p. 9, lines 139-140 ff: 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.

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

p. 12, line 204: lunar

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

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?

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.

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

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.

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.

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

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?

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

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.

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

p. 28: Will code be available publicly somewhere?

Yes, see the comment above.

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.

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?

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.

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

References cited in review:

Gaarder, T. and H. H. Gran. 1927. Investigations of the production of plankton in the Oslo Fjord. Rapports et Procs-Verbaux des Runions du Conseil Permanent International pour l'Exploration de la Mer 42: 1-48.

Odum, H. T. 1956. Primary production in flowing waters. Limnology and Oceanography 1(2): 102-117