Roman & Levin et al. updates the current knowledge of biological effects of ocean deoxygenation and reviews indicators that could be useful if implemented into monitoring programs designed to detect and track ecological shifts as a direct consequence of low oxygen stress.

General comments

There is a great and timely review of the biological response to hypoxia and the large-scale impacts of deoxygenation. Literature examples span multiple ecosystems and are global in representation. The authors target a wide audience and bridges this with examples that span multiple biological levels of organization perspective (and thus interdisciplinary among the biological sciences). As a consequence, they brought to my attention new papers on a topic that I am somewhat familiar with which is an indication that the literature review component alone is immensely valuable. The review furthers the discussion by touching on where deoxygenation may have been overlooked by major research and global science initiatives. This will be a useful guidance document for those who wish to direct resources towards 'oxygen' as a key variable in their monitoring and management efforts, and hopefully cite this review as a result.

The majority of my comments focuses on a meat of the review – the summary of indicators of deoxygenation stress. At the onset, the authors introduce key terms like 'indicator', 'monitoring programs', and cite the Essential Ocean Variables (EOVs) as background context. Readers that already have some buy-in and are tasked with monitoring efforts would predictably ask *"what do I measure and how do I measure it?"*. These are the start-up questions to address when developing indicators for monitoring programs (See Yoccoz et al. 2001, Reynolds et al. 2016 for more). The term 'indicator' could be better defined and applied in the manuscript. In general, ideal indicators have specific criteria when being evalualted for monitoring programs – they should be: quantifiable (*measurable with units*), sensitive (*small change in the system pressure incudes a big change in indicator*), responsive (*indicators responds in the same time frame as system pressure*), specific (*not influenced by covariates*), and operationally feasible (*i.e. costs associated with acquiring enough data points so that a signal can be detected is reasonable*)…and be able to detect & track trajectories of change over time (or space) as a consequence of the system pressure (here – deoxygenation).

Several of the examples as presented are uncoupled from this main manuscript theme which sandwiches the beginning and end of the manuscript. Not all the reviewed biological response examples are presented as indicators in the current text but still exist as a recap of their original citation. Additionally, a science evaluation of the criteria for indicator assessment and associated 'challenges' of using them in monitoring programs could be useful – this is loosely touched upon in Table 1 (e.g. the *specific* criteria relates to the table's confounding factors), but is inconsistent. Some structured text that considers this for all presented examples could be useful in a minor revision. The translation of the expansive theoretical knowledge into operational guidance is often the missing piece that results in the failure to launch of many monitoring programs.

Refs:

Yoccoz et al. Monitoring of biological diversity in space and time (2001) Trends in Ecology & Evolution, 16: 446-453.

Reynolds, J.H., Knutson, M.G., Newman, K.B., Silverman, E.D. and Thompson, W.L., (2016). A road map for designing and implementing a biological monitoring program. *Environmental Monitoring and Assessment*, *188*, pp.1-25.

Specific Line comments - ***Many of my specific line comments were written before my general comments section; they will mostly focus on whether the biological response examples could meaningfully be used in a monitoring program designed to track deoxygenation impacts.***

L40 – is 'indicators' an appropriate term here? Biological response seems a better fit.

L82-L85 – Do the authors suggest that 'only' measuring biotic indicators here? Or in companion with existing oxygen monitoring? If the former and with the rest of the manuscript read – it would be challenging to conclude on the causal relationship of any signal in the biotic parameters back to deoxygenation without having the ambient oxygen having been measured in parallel as well.

L85-L89 – it would strengthen the messaging to the intended target audience listed in the previous sentence if a guiding definition of 'indicator' is presented.

L89-L92 – subjective and arguable given the evaluation criteria of what makes a good indicator for monitoring programs hasn't been presented. Changes in animal behaviour, unless specific to the focal question/objective of the monitoring program could be considered poor indicators?

E.g. if the cost of deploying and recovering an oxygen sensor is less than the cost and risks of acquiring and maintaining a population of animals and measuring live animal behaviour in a controlled experimental setting…than the latter would not be a good indicator. How would one then apply any lab-derived data back into the natural system state influenced by deoxygenation stress?

L98-102 – Just a thought about the idea behind the EOVs and how the relate to this review. Core traits of the EOVs are that they are measureable and inclusive, but also 'non-specific' to just one system pressure/stessor. The challenge with developing deoxygenation indicators is to have them be 'exclusive' – not correlated to other pressures in that they are informative to one very specific stressor.

L107 – « *Garcon et al. 2019 examined…* » What did the authors conclude ?

L124 – The covariate and scaling issue. Given these are established challenges with interpreting the ecological consequences of deoxygenation/hypoxia – can the authors reflect on the presented list of their indicators and suggest which ones are best?

L143 – The focus of this review

Reading through – are the levels of biological organization intended to present the 'level' of organization at which (1) the indicator is measured, or (2) the level of organization at which 'the biological response' manifests? The manuscript isn't quite consistent or clear on this.

L154 – Consider "Individual-level indicators"?

Individual-level indicators – these are described more as biological responses – there lacks a discussion on how these can be translated into monitoring programs?

E.g. in research fisheries surveys – individual specimens are brought up in catch. Length (cm), weight (kg), age (otoliths), are measurable, individual-level data. I can envision tissue & blood subsamples being taken if some of the presented indicators can be quantified post-hoc (e.g. protein & enzyme levels present at sampling), but the behavioral ones, at least as presented, would be extremely challenging to implement in a realized monitoring program.

L163 – The indicator "Amount of HIF-alpha protein"

- Indicator sensitivity is species-dependent

L182 – Not clear why HIF may be more difficult to measure? Is it the cost and challenges of the live-animal cultivation requirements to get a data point?

L186-L213 – For sensory systems indicators and the suite of variables that require manipulative lab-based experiments to quantify – would these really be 'monitoring indicators'?

L198 - Would monitoring programs, which infer natural/field systems…be tracking vision loss or sensory loss in situ? How would fisheries actually measure this within their standard equipment? Table 1 suggests 'control' specimens would be required to be brought into the lab, and manipulated via factoral experiments . Unless these biological responses can be calibrated back to a field-based indicator (can behavioural responses like these be calibrated to a fisheries metric like CPUE?) – these feel a bit uncoupled from being useful monitoring program indicators without additional thought.

L186 – Vision – perhaps mention this metric only applies and is limited to mobile animals with eyes.

L302-L306 – Pcrit, oxyregulation/oxyconformation and the theory comes from quite a long history of fish physiology research – suggest citing at least one reference here. (e.g. Fry 1933).

L320-L329 – the theory underpinning Pcrit and oxyregulatory ability also relates to ontogeny; this nuance is often lost with the generalization towards 'size'. The ability to oxyregulate is not present throughout developmental life-stages in many marine ecotherms – the ability to oxyregulate manifests in later life history stages (e.g. fish) which also correlates with larger body sizes.

Pcrit seems out of place here but also is inconsistent with what is / isn't an indicator – which goes back to the comment on what is actually being measured and how it can be measured as part of a monitoring program. Compare this with the units of the other indicators presented which is quantified through a measurement of a biological parameter / response variable.

A Pcrit is just a single environment oxygen level but determined to be a biological relevant threshold using live-animal respirometry experiments. Time-series of the threshold itself is not what would be generated by a monitoring program as it's not the parameter that is responding to the deoxygenation system pressure nor would time-series be generated of Pcrits. A realized deoxygenation monitoring program that implements Pcrits into its general framework would realistically be generating standard oceanographic oxygen time-series using CTDs with the

interpretation/early warning signal being the where and when those oxygen data show environmental oxygen levels to be at Pcrit levels of interest.

L377 – *Lethal hypoxia has also been estimated from field measurements with organism presence/absence.*

I would interpret these as more a sublethal hypoxia threshold being estimated from organism presence/absence for at least mobile species; since most will swim away before at levels higher than the actual lethal levels are reached unless the oxygen loss was rapid and companion data on carcasses was also available

L421 – Metabolic Index –high degree of extrapolation, not only from individual-level lab-based experiments but to entire theoretical niches (based on only a few dimensions…O2 and temperature primarily), but also across entire phyla, coupled with the uncertainties of climate-model forecasts.

The primary discussions coming from this metric have been theoretical hypotheses presented as predictions – as noted, the real-world validation through testing the accuracy of M.I. predictions at the regional scenarios – the extant fisheries perspective remain limited.

When considering the metabolic index under the monitoring / indicator science context – the metabolic index is a derivative of Pcrit theory and so the same comments above apply here as well. The metabolic index in its current formulation doesn't quite seem like an indicator that could be used to track and detect system-state changes.

L489 – Suggest clarifying that indicator species is not a hypoxia-specific idea, and they are bioindicators of a set of environmental conditions or state.

L536 – *Indicator species presence may be a straightforward way to detect oxygen changes and is easy to interpret.*

Potentially – but cited examples are metazoans only.

Empirical evaluation of how well an organism can act as an indicator species is determined by fidelity and specificity. The hypothetical perfect indicator taxa of 'deoxygenation waters' would have high fidelity (the species is present at all sites with deoxygenated conditions) and highly specific (that species is only present at sites with deoxygenated conditions). Since there are no obligate anaerobic water-breathing animals – using any metazoan as an indicator species of deoxygenation conditions would always require additional context and would not be straight forward. I.e., all metazoans can inhabit normoxic waters…if they are absent, it would be another dimension of their niche that is restricting them from occupying oxygenated waters.

Considering the above, consider adding context and specifying obligate anaerobes as potentially more ideal use case for 'indicator taxa' theory.

L577–L624 are behavioral responses described not actually individual-level responses when being measured?

This section repeats the measurable parameter described in the individual-level response associated with sensory systems.

E.g. L589 – this paragraph describes the consequence of low oxygen exposure to individual fish rather than presenting as an actual measurable indicator. For a measurable parameter indicator a population-level indicator (or biological response); the population-level response here when describing the compression into shallow waters (shoaling) is the decrease in the average depth distribution of the population.

L635 – Population size

As presented – the measurable parameters presented for population size/growth rate/recruitment are standard ones measured by traditional stock assessment monitoring programs.

Can you provide some discussion on whether this would be a 'good' indicator given it relies on having a foundation of another species/individual-specific indicator (lethal O2, L641) and how in reality, if this measurable parameter (standardized counts for a species') is confounded by so many other variables. In general, consider adding some text that clarifies when indicators may or may not be useful in a realized monitoring program.

L711 - Ecosystem Indicators: L713-L782 – Most of the quantifiable parameters only inform on the community-level of biological organization (e.g. species assemblages metrics); suggest clarifying this in the section subheading.

L752 – there are additional 'diversity indices' not mentioned that integrate some level of functioning into their formulation by including simple traits / functional levels into the calculations. E.g. ITI (infauna trophic index) and the AMBI – AZTI indeces are often used in aquaculture-impact assessments marine biotic index that includes simple traits/trophic level categories into the calculations. On the topic of and linking back to anthropogenic over-enrichment; the benthic infauna/macrofauna diversity indeces do include ones that integrate a level of functional. While this doesn't explicitly link back to hypoxia sensitivity, they are inversely correlated to hypoxia (as the are correlated to high sulphide levels), and integrate trophic traits into their formulation, thus may be slightly better at representing the ecosystem level of organization than pure species assemblage derived diversitiy metrics.

L799-L814, L838-851 – Abundance and Biomass

As written – this is discussion individual population-level abundances without a quantifiable metric that captures the interspecies component of the community re-organization in mind. E.g. Given hypoxia sensitive differs among species – there can be interspecific shifts in abundance levels (some increase, some decrease) as some moderate hypoxia thresholds are crossed, but community-wide biomass decreases as oxygen approaches zero

Analogous indices have been developed for infauna to monitor for system degradation as a consequence of anthropogenic enrichment (e.g. poylchaete/amphipod ratios).

L853 – Taxonomic shifts and ratios:

I wrote the above comment before reading this section – which might work better if it shifts before the abundance/biomass paragraphs.

In a system experiencing gradual deoxygenation – the interspecific shift in species abundances usually happens first (among mobile species - sensitive ones leave, tolerant ones invade) before the community-wide decline in abundance occurs. This would be a useful to suggest as

the ‘early-warning’ signal, given the shift in relative abundances among species would theoretically happen first before community-wide abundance decline.

Taxonomic shifts could also go by different names discussed earlier in the paper (i.e., beta diversity), so there is some intellectual overlap here with the earlier section as well.

L930-934 – The metric presented sems to be behavioural response and not a measurable indicator of a community-level component. Given this is the ecosystem-level + ecosystem function discussion section, as a reader I would be looking towards quantifiable biological processes capturing the energetics of the system (e.g. biological rates) – this paragraph could use a retool and link the shift in bioturbation behaviour to the measurable ecosystem indicator.

L1020 – Scaling of Indicators

Might be useful to provide an operational definition on what ‘scaling’ means; the types of ‘scaling’ aren’t clearly defined for the reader. The common theme across the summarized ‘scaling’ approaches appears to be ‘multi-level integration across biological responses to deoxygenation’ but not clearly defined in a way that makes the differences obvious. Given ‘spatial-temporal’ is also discussed (L1023) - I was looking for the traditional concepts/definitions of ecological scale, resolution (grain) and extent, to be presented in a discussion that draws analogies to the biological organization structure theme; wasn’t sure if that was the intent. I feel just a slight refocusing of the wording could help with the clarity in this section.

L1085 – “oxygen-stress”, perhaps reword to deoxygenation stress. Review doesn’t discuss hyperoxia effects (oxygen being too high).

L1186-L1188 - *Many of the biological indicators of oxygen stress described in this*
1187 paper, if tied to specific DO response thresholds,

From the perspective of, “what is needed in the practical launch and implementation of a realized deoxygenation monitoring program” – this might be the key statement to emphasize. Implementation of any of the parameters to produce longitudinal data (time-series) to track biological responses to deoxygenation will require the natural oxygen levels to be monitored in tandem.

Technical corrections:

L433 – Howard et al. (2020) missing from References

L1079 – ROMS-BEC – could you define this acronym?

L1087 – Table 1. Difficult to read with some text cut off within the excel table format.