Response letter

We appreciate the remarks and suggestions of the reviewers and are grateful for the effort the they have invested. Below we respond to each comment individually and indicate the revisions accordingly. For clarity, referee comments are indicated in bold and Authors' comments are indicated in italics. All the changes in the revised manuscript are highlighted with yellow.

Reply to Anonymous Referee #1:

The robustness of the transfer function could/should be test via bootstrapping or a similar analysis (this would help to determine what is the 'counting error', as in how much a few percent change in the abundance would change the transfer functions resultant oxygen value). Moreover, the fitted/predicted oxygen values give negative values, below 0 (figure 5 to 7), I am unsure whether this is possible with concentration values? The authors have considered sources of error, one such source of error is the preservation potential of certain forms.

We tried to apply a test for the transfer function using PAST software, however this was not possible using the target dataset and multiple regression. Since the target dataset is already reduced to 16 species (35 samples) from 123 species (53 samples; see also section 2.2.1) such a test has a little input. Following the bootstrapping suggestion by the reviewer, we re-analysed the living benthic foraminifera dataset (123 species from 53 samples) applying centred bootstrapping and it indeed improved our results as presented in figure 3. This did not change our interpretation however it solved the error bars issue mentioned below (figure 3).

Furthermore, species observed in living dataset are not identical to fossil assemblages in terms of abundance but similar enough to apply transfer function. Once looked closer, only one data point at one sediment core indicated a negative value (late Holocene at core M77/2-59-1 which is the northernmost core) and the all estimations are positive within the statistical uncertainty. The scale bar of these graphs show negative values since we present the results with the standard deviations and the 1sd at some data points are high, it seems like most of the estimates are below 0. The reference dataset based on living foraminifera is predominantly retrieved from stations from the OMZ centre where oxygen concentrations are really low. Therefore, abundant species considered in the key 16 species are predominantly 'low-oxygen tolerant' ones excluding others that are observed in the downcore record. This potentially results in slight shift in the quantification towards lower values. We made this more clearly discussed and improved the section 4.2. All additions and modifications are highlighted in the revised version.

For clarity though, it might be prudent of the authors to state how the agglutinated were removed from the data (pg. 5 Line 29-30), i.e., is there potential for error through a closed sum effect? If the assemblage is counted to a certain number, or split to gain a certain number of grains, by removing data (which has to be done) does this introduce some bias (when considering low abundant species – removing 3 specimens in a count of 300 means a loss of 1%, the 100% would be then based on 297, it also shifts the percentages for the remaining species which is more problematic for rare species than for dominant species)?

Comparison of living assemblages with fossil record showed that there is a distinct difference between abundance of agglutinated species. This resulted in large bias in downcore applications and thus we decided to continue our transfer function only with calcareous species. Nevertheless, for the living benthic foraminifera interpretations agglutinated foraminifera are included and these results are presented in supporting information.

Moreover, their appearance together with Miliolids at few downcore samples is in accordance with our observations on their intolerance to low-oxygen and more oxygenated bottom waters during the LGM.

In a similar vein how reliant is the transfer function on small changes for rare species – and have the authors considered a transformation of the abundance data to reduce the impact of dominance and rarity (e.g. Log the data)?

Only common species (at least three occurrences with >5 %) are considered in the reference data set, even though frequent species may also be rare in certain samples. As such, small changes in the proportion of rare species will not affect the results. A log-normal distribution is rather a character of volumetric data (i.e., population densities) rather than percentages. Therefore, a logarithmic transformation of the data was not attempted.

Have the authors considered more environmental variables (e.g. temperature, salinity, etc), whilst the approach here is to reconstruct bottom water oxygen concentrations the question is, is this the dominant control on the assemblage composition? This is especially important given how regional the dataset is especially when comparing different time periods.

We agree with the reviewer that these are really good on point questions and comments, one should always keep in mind the other factors in such investigations. We present oxygen (and rain rates) in relation with the living benthic foraminifera dataset since these are the only parameters either available or possible to calculate for all these sample locations for living benthic foraminifera dataset. Individual studies mentioned here (few of them are published already (Mallon et al., 2012; Cardich et al., 2015)) discuss the relationship between living benthic foraminifera and environmental factors. For the scope of this study we focused on oxygen concentrations which is potentially a dominant factor in such strong oxygen minimum conditions. It would be indeed an interesting study to combine such an extensive dataset with environmental parameters and statistically test their relationships as Cardich et al. (2015) reported for stations from the shelf. In case of paleo-reconstructions, there are plans for comparison work focusing on single sediment archives with application of different proxies.

How similar or dissimilar are the various assemblages for each time period? And how similar or dissimilar are they between? It is not that I doubt that oxygen is a dominant variable, instead knowing whether there is some variation in the assemblage due to another variable might help to put the results into better context.

Downcore distribution of benthic foraminifera at the same samples with more emphasize on the taxonomy was already published (Erdem and Schönfeld, 2017). Due to assemblage similarities observed in concerned sediment archives we continued with the transfer function approach. The only sediment core which shows relatively different abundance of certain species is the northernmost core M77/2-59-1 (e.g., Bolivina costata is not as abundant as observed in other sediment cores whereas Bolivinita minuta is more abundant). Obviously, when relative abundances of certain species show distinct changes between periods, that is reflected in the oxygen estimations. For instance, B. costata and B. minuta both have a large coefficient (table 5) that has a strong influence on the reconstructed oxygen concentrations. During some periods these species are more abundant indicating lower oxygen levels during these periods.

The discussion of the data is lengthy - what I miss is a statistical comparison between d15N, TOC and the O2 prediction of the authors – as in, figure 7 is under used. Here the authors could compare their proxy against the previous proxy values statistically (e.g., simple scatter or regression analysis) rather than descriptively (section 4.3).

We thank the reviewer for the suggestion. Any paleo-investigation, proxy based research, needs a comparison with other proxies which are potentially related to conditions aimed to be investigated. Our manuscript aims to investigate paleoxygenation that is relatively difficult with currently available proxies (which are introduced to reader in section 1) with each having certain limitations. By multi-proxy applications we aim to cover these limitations with other proxies. We here focused on only two of these proxies; $\delta^{15}N_{sed}$ and total organic carbon content, since they are available for the same cores. So that we did not extend our investigations to wider region also because it is already a long manuscript as the reviewer pointed out. Application of statistical tests to see the relation between proxies can be done but might also result in misleading interpretations since each proxy is influenced by each other (particularly oxygen and productivity related ones) and various other environmental factors. Furthermore, a scatter plot will not reveal immediate leads and lags depicting temporal relationships of the proxy records, and its statistical characters will be confined by outliers, which only can be identified and assessed in a meaningful manner if the succession of the parameters versus time is displayed.

Finally, I agree with the other reviewer that the age models should be outlined in this paper somewhere (e.g., diamond symbols with the depth in cm in the figures?), given Pg. 8 Line 18 (: 'Erosion, reworking and high energetic bottom conditions').

We followed this suggestion and gave a detailed age model description in the revised version. We modified Figure 6 and added age model tie points for each core. We made the radiocarbon dating results available on Pangaea database. All these datasets will be linked to this manuscript in the final step.

Text comments:

Pg. 6 Line 4 -7: agglutinated forms have a lower preservation potential, could this affect the splitting? Removing species from abundance counts does impact the closed sum;

We agree with the reviewer that it impact the closed sum. We still present their relative abundance in the living dataset not to mislead the reader for their occurrence. Hence once statistical tests were applied to only calcareous forms we observe large errors and variations at diversity and dominance measures of some samples (e.g., M77/2-776). However, agglutinated forms are absent in the downcore record which makes transfer function approach not applicable if we keep them in the target dataset.

Pg. 6 Line 19: Mean values – would it not be better to consider the mean with std dev. to construct the equation?

The dataset mentioned here are published (Cardich et al., 2015). It concerns revisits the same sampling locations over several years. For this study we used the same datasets but averaged values as if it is one sampling. We realized that we used different wording in different sections (2.2 and 2.2.2) when describing the datasets. We modified the concerned sentence under section 2.2.2. accordingly.

Pg. 6 Line 20: "from a synoptic compilation" what do the authors mean by synoptic (= general, vs. synoptic data)?

Schönfeld et al. (2015) presented a compilation of CTD data obtained during cruises R/V Meteor M77 legs 1 to 4 between October 2008 and February 2009.

Pg. 6 Line 25-29: How different are the 'different primary productivity values' used for the RRPOC? What would the values be if the same equation were used? This can be tested by applying each set of values used.

We only calculated rain rates for sampling locations when the information is not available. Otherwise we used the same equation to calculate the rain rates for most of the sampling locations (Martin et al., 1987). Stations from shallower than 100 m and deeper than 1000 m is not covered unless it was already reported in Dale et al. (2015). Different values mentioned here are the primary production estimates from Pennington et al. (2006) and Martin et al. (1987) concerning latitudinal differences. In our calculations, it is not possible to use the same primary productivity estimations since they were reported showing distinct latitudinal differences (i.e., Equatorial upwelling, 13 mg C m^{-3} d^{-1} vs Peruvian coastal upwelling 145 mg C m^{-3} d^{-1} (Pennington et al., 2006)). We were able to compare the results from Dale et al. (2015) and our calculations for latitudes 11°S to 12°S. Overall results from Martin curve are slightly higher than observations of Dale et al. (2015). Nevertheless, the offset is consistent and it would not impact the observations presented in figure 2b for instance. Meanwhile, we realized a mistake at the figure caption that is now corrected in the revised version.

Pg. 10 Line 25 – 29: Have you considered placing the various species into comparable niche occupations? The table is a good reference guide for readers, but it would be interesting whether the different species regionally/globally occupy different niches or similar ones.

The aim of the table is to provide an overview of the abundant species observed at the Peruvian margin sediments (both modern and downcore). To some extend we tried to bring out their regional or global occurrences in relation with specific environmental factors. Some species such as Epistominella exigua have relatively high amount of records in publications that can be implemented to certain environmental conditions. However, we should be careful while doing that, especially when there is not much information available, as in case of Bolivina costata. For this reason, we wanted to give a small review on availability of information on these species. We will learn more as the genetic information become available over time.

Pg. 12 Line 14-16: "Moreover, we are confident in the [O2]BW differences in each time interval considered, even though the absolute estimates for each sample might be biased because of the dominance of the low oxygen samples in the reference dataset." – maybe elaborate why you have confidence despite the absolute estimates being biased? And how does the absolute estimates being biased fit with research question 2 and 3?

This part of the discussion (section 4.2) is improved accordingly as mentioned earlier under the first referee comment.

Section 2.2.2: what is the sensitivity of the CTD and equipment used for oxygen, is there not some lower limit (5 umol/kg) below which the data is not accurately measured? Or at least the reliability is not the best.

The oxygen sensor that was used on the CTD was a electrochemical "Seabird" sensor (Clark type) that has a detection limit of 1-2 µmol/kg (Revsbech et al., 2009). In a comparison study to other sensors (STOX vs Clark-type (Clark JR et al., 1953; Revsbech et al., 2009) it has been reported that the actual oxygen concentrations at the Peruvian OMZ can be much less when the seabird sensors reach this limit (up to 2 µmol/kg (Kalvelage et al., 2013). Sometimes it can be at the lower nmol/kg range in this region (Revsbech et al., 2009). The maximum error of the oxygen data is constrained to +/- 0.5 µmol/kg. Therefore, almost zero oxygen conditions are recordable. However, values below 2 µmol/kg should be treated with care when Clark-type Seabird sensor was used. In case of our living dataset

this is a concern for three stations out of 53. The oxygen concentrations at these stations were likely even lower according to the Revsbech et al. study. A downward correction of 1 µmol/kg would not influence our transfer function in a statistically significant way.

Section header 1.1 Benthic foraminifera as oxygen proxy -> 'as an oxygen proxy' or reword as 'as a proxy for oxygen'?

Corrected in the revised version

Figure comments:

Figure 1a: scale bar missing – if the authors (as implied by the caption) are trying to demonstrate the low oxygen values how about a single contour around the purple? Figure 1 Caption: should it read as two units? "<0.5 ml/l to <20 mol/kg" the 'to' implies a 'sliding scale'

We added a contour line to F.1a as suggested and modified the caption in the revised version.

Figure 2b: perhaps color the symbols to show the different rain rates?

In the revised version the sampling locations with similar rain rates were grouped using different colours.

Figure 3: (capitalise R in relative abundance). (Bottom panel) The last datapoint (sample M77/2-776) is forcing the plot's yaxis to be skewed toward higher fisher alpha values so that the values of the other samples are condensed. Consider, perhaps using a logscale for the yaxis of the Fisher Alpha panel, alternatively the authors could exclude from this plot sample M77/2-776 and with a big red arrow just tell the reader the values of this 'outlier'. (Bottom and Middle panel) I assume the bars are 'errorbars' – some seem to be not symmetrical around the datapoint (possible depending on the statistic used) but more importantly Site 830, 1004P1 the error bars are below the datapoint.

As mentioned earlier we improved this figure using PAST and applying bootstrapping to living foraminifera dataset. We also modified the scale bar of Fisher alpha.

Figure 5 to 7: Is it possible to have a negative value for oxygen concentration?

As previously discussed, no it is not possible to have negative concentrations. This is potentially the artefacts of the transfer function where the estimated values are biased towards lower values. We think that it is because majority of the samples are from really low oxygen concentration depths. Nevertheless, our estimations are positive within the statistical uncertainty. This is the reason we keep our discussion with changes over time but not exact values for specific time. We added a short paragraph under section 4.2 where relevant.

Figure 5: give a 1:1 line

1:1 line for both graphs is now shown in the revised version

Reply to Anonymous Referee #2:

1. In the current format, the authors have not demonstrated that the live population are identical to the dead population in the core tops, and without this evidence down-core reconstructions are not scientifically scrutinized.

We thank the reviewer for addressing this important point, and we take the opportunity to emphasize again that living faunas and dead assemblages are generally different in species composition, and that standing stocks are not mirrored in the concentrations of empty tests in near-surface sediments in most environments. These fundamental differences are intrinsic. Living faunas represent the conditions during the weeks before sampling, whereas dead assemblages are a product of many generations added over an unconstrainable time period. Therefore, species richness and density of the dead assemblages are generally higher than the respective values of the living fauna. Another biasing factor is the taphonomic processes altering the composition of the dead assemblages through time, in particular during successive burial under the influence of different redox and pH conditions. The fossil assemblage in sediment cores thereby differs markedly from the dead assemblage at the sediment surface. Therefore, an identical composition of living faunas and dead assemblages in the topmost layers of sediment cores is impossible to be found, and this should not be imposed as a prerequisite for downcore applications of foraminiferal proxies.

The very reason why benthic foraminifera are proven reliable paleoindicators is because they live in equilibrium with the ecological conditions in their immediate environment. Due to their short generation times of usually less than a year, they respond quickly to changes in the setting of abiotic or biotic environmental factors. Once a foraminifera reproduces, an empty test is conveyed to the sedimentary record. If the species is reduced in abundance due to environmental changes, a lower number of empty tests is produced per unit of generation time. Conversely, a species benefiting from the change and increasing in abundance will deliver more tests to the fossil record. A transfer function, as it has been applied in the present study, relates the relative abundances of those species to the change in environmental conditions and accommodates for taphonomic alterations. The reviewer is referred to the interesting textbook of Fisher and Wefer on "Use of Proxies in Paleoceanography" for further reading.

In summary: our approach is scientifically valid and verifiable through comparison with other proxies for past oxygen conditions.

2. Information concerning age models of the different cores is missing. The age model needs fully discussed and shown in the article as it is crucial to consider the context and interpretations of the reconstructions.

We followed this suggestion and gave a detailed age model description in the revised version. We modified Figure 6 and added age model tie points for each core. We made the radiocarbon dating results available on Pangaea database. All these datasets will be linked to this manuscript in the final step.

3. The authors should have a good look at their data and critically reflect whether their conclusions really reflect the data. The main Figure 6, I presume, shows reconstructed O2 plus error. Main changes seem to occur during deglaciation. There does not appear to be any differences between LGM and core tops/late Holocene (the authors suggest a 30 uM change from

the LGM to Holocene at the lower OMZ boundary): -The first site at 626 m shows (within error!) similar O2 values during the LGM as core top; e.g. no statistically significant increase in LGM oxygenation. -The second core at 1013 m: all reconstructed values are below present day values: no significant increase in LGM oxygenation here. -Third core site at 1249 m: LGM oxygen concentrations are lower compared with core top; so no significant LGM increase in oxygenation here. –Fourth core at 997 m: perhaps H1, early deglacial higher O2 values; but no reconstructions for the LGM. So none of the cores show that the Peruvian margin, at the water depths investigated, was better oxygenated during the LGM compared to today.

For modern oxygen values we used the CTD data collected during each expedition at the same time when living benthic foraminifera samples were collected. The stars shown on the figures are indicating the values when the sediment archives were collected (during R/V Meteor M77 expeditions). The LGM estimations are indeed either really close or below the actual measurements which is seemingly a concern. We mention our concerns about absolute values and potential bias toward lower oxygen value in quantification in section 4.2. Nevertheless, it is possible that bottom waters became more oxic after the late Holocene as reported for the shelf during the last 100-150 years (Cardich et al., 2019). However, we cannot comment further for the rest of the Holocene trend on the basis of currently available information. As this circumstance is apparently not sufficiently addressed in our Discussion, we improved section 4.2. accordingly.

Concerning the potential bias towards lower oxygen values, we restrain ourselves making comments such as; 'during the LGM at 1000 m water depth oxygen was 50 µmol/kg'. We rather focus on the absolute changes between periods and sediment archives. Still we present all the quantification results for each data point in Supplementary information Table 1. The average values for each time period were calculated according to estimations presented here. Unfortunately not all time periods are covered in every core, therefore we emphasize that the results are stacked (in the revised version sentence starting at P. 11 Line: 29). We are aware that the approximation of 30 µmol/kg is predominantly influenced by the results of core M77/2-52-2 (LGM ranging between 52 and 61 µmol/kg vs. late Holocene ranging between 23 and 33 µmol/kg), since it is the only core which covers all of the concerned periods. Once again we primarily focus on the change in oxygenation rather than reporting absolute values as given facts for these cores.

Concerning the results of core M77/2-47-2 from (626 m), which does not indicate any change during the LGM and deglaciation, we conceded that these results are puzzling in the first glance, Nonetheless, this record also shows that when the OMZ intensifies (or diminishes) the change is profound around its borders and the conditions are rather stable close to its centre. This regional dynamics has also been disclosed with other proxy based approaches (as discussed in section 4.2; in the revised version P.12 Lines: 8-11). Moreover, during the LGM this core location was at least 100 m shallower which was potentially within the OMZ core. We modified and made additions under section 4.2. All these changes are highlighted in the revised version.

References

Cardich, J., Gutiérrez, D., Romero, D., Pérez, A., Quipúzcoa, L., Marquina, R., Yupanqui, W., Solís, J., Carhuapoma, W., Sifeddine, A., and Rathburn, A.: Calcareous benthic foraminifera from the upper central Peruvian margin: control of the assemblage by pore water redox and sedimentary organic matter, Marine Ecology Progress Series, 535, 63-87, 10.3354/meps11409, 2015.

Cardich, J., Sifeddine, A., Salvatteci, R., Romero, D., Briceño-Zuluaga, F., Graco, M., Anculle, T., Almeida, C., and Gutiérrez, D.: Multidecadal Changes in Marine Subsurface Oxygenation Off Central Peru During the Last ca. 170 Years, Frontiers in Marine Science, 6, 10.3389/fmars.2019.00270, 2019.

Clark JR, L. C., Wolf, R., Granger, D., and Taylor, Z.: Continuous recording of blood oxygen tensions by polarography, Journal of applied physiology, 6, 189-193, 1953.

Dale, A. W., Sommer, S., Lomnitz, U., Montes, I., Treude, T., Liebetrau, V., Gier, J., Hensen, C., Dengler, M., Stolpovsky, K., Bryant, L. D., and Wallmann, K.: Organic carbon production, mineralisation and preservation on the Peruvian margin, Biogeosciences, 12, 1537-1559, 10.5194/bg-12-1537-2015, 2015.

Erdem, Z., and Schönfeld, J.: Pleistocene to Holocene benthic foraminiferal assemblages from the Peruvian continental margin, Palaeontologia Electronica, 20.2.35A, 1-32, 2017.

Kalvelage, T., Lavik, G., Lam, P., Contreras, S., Arteaga, L., Löscher, C. R., Oschlies, A., Paulmier, A., Stramma, L., and Kuypers, M. M.: Nitrogen cycling driven by organic matter export in the South Pacific oxygen minimum zone, Nature geoscience, 6, 228, 2013.

Mallon, J., Glock, N., and Schönfeld, J.: The response of benthic foraminifera to low-oxygen conditions of the Peruvian oxygen minimum zone, in: Anoxia, Springer, Dordrecht, 305-321, 2012.

Martin, J. H., Knauer, G. A., Karl, D. M., and Broenkow, W. W.: VERTEX: carbon cycling in the northeast Pacific, Deep Sea Research Part A. Oceanographic Research Papers, 34, 267-285, 1987.

Pennington, J. T., Mahoney, K. L., Kuwahara, V. S., Kolber, D. D., Calienes, R., and Chavez, F. P.: Primary production in the eastern tropical Pacific: A review, Progress in Oceanography, 69, 285-317, 2006.

Revsbech, N. P., Larsen, L. H., Gundersen, J., Dalsgaard, T., Ulloa, O., and Thamdrup, B.: Determination of ultra-low oxygen concentrations in oxygen minimum zones by the STOX sensor, Limnology and Oceanography-Methods, 7, 371-381, 2009.

Schönfeld, J., Kuhnt, W., Erdem, Z., Flögel, S., Glock, N., Aquit, M., Frank, M., and Holbourn, A.: Records of past mid-depth ventilation: Cretaceous ocean anoxic event 2 vs. Recent oxygen minimum zones, Biogeosciences, 12, 1169-1189, 2015.