B. Metcalfe on behalf of the co-authors, Response to Reviewer 3

[reviewer comments as red text in blocks]

We thank reviewer 3 for their time in reviewing our paper. However, it is unfortunate that the reviewer did not take the time to read our response to reviewer 1 as many of the same concerns have been addressed there and we have responded to those questions in detail. It is also unfortunate because it would have, hopefully, aided the reviewer in realising that this paper is about testing whether the given foraminiferal populations are statistically different so that they could potentially unravel different climatic states. Our approach is not about IFA research; IFA research is referred to because it provides an excellent sample dataset (and we highlight that it is one of the ways to understand a climate history more thoroughly). It is however not the only dataset and hence we make reference to other studies as well. We also find it disappointing that once more we are having to discuss inverse and forward modelling, as well as the statistical tests used; all aspects that where already explained in our previous response.

> *In this study, Metcalfe et al. aim to test whether the approach of using individual foraminifera analysis (IFA) can be used to assess ENSO variability. In order to accom- plish this, they use the Foraminifera as Modeled Entities (FAME) model to calculate idealized foraminifera distributions across the tropical Pacific. These results are then combined with seafloor/ CCD depth and sedimentation rate to determine which regions of the Pacific Ocean are suitable targets for IFA approaches. Modeling of foraminifera populations in order to determine if ENSO change is detectable has been done before (e.g., Thirumalai 2013, White 2018), although these studies focus on the detection of ENSO from paleoclimate proxy records. This study's novel contribution is the inclusion of the FAME model and foraminiferal growth rates to the analysis of modeled response of biological calcite to tropical variability.*

We once more state we are not testing IFA or have a model which tests IFA approaches. IFA is a suitable dataset to compare results with because it gives a lot more information than pooled analysis.

> *However, the FAME portion of the model is not validated against core-top data from the tropical Pacific, precluding assessment of its utility.*

The FAME model is validated against the whole MARGO dataset which includes the tropical Pacific. Please read Roche et. al. (2018).

> *The application of these results is likewise problematic, as it focuses on determining whether ENSO events (El Niño, La Niña) and neutral conditions have distinct distributions (forward modeling) rather than on how one could detect ENSO change (inverse modeling).*

There are a number of applications of this method, which we have outlined in response to reviewer 2. However, we consider that the reviewer comment is not an argument against what we have done – the first basic principle of understanding a proxy is 'can we detect', not 'how could we detect', as the how implies we know we can. Our manuscript is devoted to answer the question "can we detect".

> *Further, the discussion on sedimentation rate and CCD is broad-based and does not take in to consideration local changes in seafloor topography, changes in bottom-water oxygen availability that may alter bioturbation depths, and the variability characteristics of different regions with regard to the seasonal cycle, decadal-centennial variability, and ENSO change (e.g., Thirumalai 2013, Ford 2015, White 2018).*

It is true we have used broad based and conservative estimators (5 cm^{-1} kyr^{-1}) as the bioturbation mixed layer is known to vary from 1 to 35 cm depending upon the various aspects the reviewer states. However, we fail to understand how "and the variability characteristics [?] of different regions with regard to the seasonal cycle, decadal-centennial variability and ENSO change" would somehow relate to SAR and dissolution depth.

> *Finally, there are aspects of the model that are unrealistic (e.g., a 400m depth for symbiont-bearing foraminifera; assuming sample sizes of 1000 for binning) or unrealized (e.g., how many individuals were selected for generating these estimates and a lack of model-data comparison) that present significant issues to the overall utility of this model for paleoceanographic reconstruction of ENSO from IFA.*

In the paper we state we use more than one depth, we first apply a CUT-OFF value of 400 m then progressively shallower (hence the reason for multiple panels) – we know this CUT-OFF value is deeper than symbiotic species (e.g., Pracht et al., 2019, Biogeosciences) hence the use of shallower depths (the reviewer is arguing against a simple test of our model here). It should be pointed out, whilst the model can in principle run down to 400 m it will only register a value if the temperature is applicable for that species (i.e., temperatures outside of the temperature window will not give growth as highlighted by the equations in Roche et al., 2018). Second, we multiply the TOTAL BIN COUNTS by 1000 to convert it into a simple distribution to test the distributions of the various climate - this is not the picking / assumed sample size. Were we to test sample picking we would need (and would have stated) to have done a larger test (although the computation required would be enormous [samples in group * replications * resampling] *[lat * lon], i.e., an individual foraminifera picking would be [1*40*10,000]*[40*120] = ~192 million computations).

The title of the article does not represent the content or main goals of the study, and the conclusions stated in the abstract are different than those in the main paper.

We disagree. If we break down the title:

Validity – is the quality or state of something being valid (valid - being logically correct or well-grounded/justifiable). We are testing whether the distributions of different climate events are statistically different (this is a fundamental test)

Foraminifera-based – we use a foraminiferal model. An alternative could include the word "populations" here, as in "foraminiferal populations".

ENSO reconstructions - The reviewer, we assume, is arguing that as we say ENSO reconstructions, and judging from their previous point ("as it focuses on determining whether ENSO events (El Niño, La Niña) and neutral conditions have distinct distributions (forward modeling) rather than on how one could detect ENSO change (inverse modeling).") that we haven't focused on the 'how one could detect', yet we are testing whether the foraminiferal distributions of different climate events are statistically different, hence we have carried out a test on a more fundamental level (i.e. foraminifera in the water, before they are incorporated in the sediment archive).

We will rephrase the abstract and conclusions for clarity.

The questions the authors raise are valid and useful, but the results as stated do not support their conclusions. In fact, the stated conclusions of the article are, in several places, contradicted within the paper itself. These contradictions are not well-explained, and thus a clear summary of the findings is difficult to parse.

We will rephrase those sections the reviewer elaborates on in the general comments.

General Comments

The study here focuses on forward modeling using FAME for IFA. However, the authors fail to prove whether existing IFA-ENSO reconstructions are valid or provide the tools for evaluating proxy data (e.g., the "inverse problem", as mentioned in other reviews, whereby foraminifera records are analyzed to infer ENSO). Thus the application of these results to the paleodata world is limited.

We are not forward modelling for IFA in sediment records, we are forward modelling foraminifera populations in the water.

The Paleodata world exists to answer questions regarding our understanding of past climate, therefore understanding if our proxies work for the period covering the observed climate record represents a fundamental test. The use of foraminiferal records to infer ENSO starts off with the prerequisite that what you are recording is ENSO, so our current research asked whether the values of the different climatic states for two proxies (calcite $\delta^{18}O$ and temperature of calcification) are statistically different. The application of these results downcore or the provision of tools for evaluating palaeoproxy data were not the stated aim of our particular research question.

The more relevant application here is in targeting locations for performing IFA studies, but this is limited as well, as the sedimentological and bioturbation properties of regions across the Pacific are much more variable captured here.

We included a rough SAR/CCD map for the benefit of the reader. However, if a reader or potential researcher in palaeoceanography has access to much better data regarding SAR/sedimentological properties, our research would still be of value because they can compare their chosen location with the FAME results that we have generated. The fact that the FAME results and FAME-SAR-DEPTH results are plotted separately allows users to pick and choose whichever plot they find necessary.

The authors use their own definition of ENSO events, despite significant previous literature and established definitions that are commonly used. The use of single month anomalies does not adequately represent the actual ENSO phenomenon, which relies on ocean-atmosphere feedbacks expressed over a period of months, and thus their analysis of differences between El Niño, La Niña, and neutral conditions may be flawed and biased toward non-ENSO SST anomalies.

We extensively discussed the definition of ENSO events in the paper and our answer to reviewer 1. In the latter we explain the rationale for our choice of definition (that foraminifera life cycle is shorter than a month and therefore several populations would exist that could conceivably have the same isotopic value as a true event).

This study does not compare the results of their FAME analysis with existing IFA reconstructions of variability from the tropical Pacific. In the eastern Pacific, Rustic 2015 used δ18O IFA on modern-era sediments to show close correspondence with calculated δ18O from reanalysis data; in the central Pacific, White 2018 showed that the distributions of Mg/Ca-based SSTs from individual foraminifera in a 4ky coretop are statistically similar to modern reanalysis data.

As per our comment to a similar question by reviewer 2, we will address this in a revised form of the MS.

Specific Comments

The authors focus on δ18O proxies for IFA, and discount Mg/Ca reconstruction and the modeling efforts done with those (White 2018, Ford 2015). To discount Mg/Ca ratios as a paleoproxy without the kind of analysis provided for δ18O seems premature. While changes in carbonate concentration, salinity, and preservation environment can indeed alter Mg/Ca ratios, significant study has been done and is underway to understand these roles. Species-specific calibrations and various corrections exist that are well quantified. Not using Mg/Ca for the Tc seems rather limited.

This point was discussed in our answer to reviewer 1. We use integrated temperature as a pseudo proxy for Mg/Ca – we did not attempt (though it would be interesting to perform such an analysis in another dedicated paper) to convert the input temperature into proxy values, i.e., a pseudo equilibrium Mg/Ca, like how the oxygen isotope values are calculated. Most ocean reanalysis and model datasets do not include the full variables required, although some models do (e.g., Grey and Evans 2019, Paleoocean. Paleoclim.).

The section of the paper that deals with and discusses Mg/Ca, is not "to discount" the proxy but was written to preempt comment(s) regarding our paper about why we did not attempt to model the full variables of this proxy (as we state in the paper why it would be beyond the remit of the paper).

Nor have we, as the reviewer states, 'discounted' the modelling efforts done with Mg/Ca (e.g. White et al., 2018; Ford et al., 2015) – if the reviewer would like to clarify this point, we would gratefully alter the text. However, this would seem to be the reviewer's own projection on to our paper and not something we categorically stated. We will try to make this clearer.

The number of foraminifera picked from a given sediment interval is an important component of IFA. Increasing bin counts to 1000 artificially (Page 6) does not represent the numbers typically used in such analyses; the numbers used for other analyses (Page 7) are not specified.

They are not specified because we didn't used the number of foraminifera picked, we artificially convert the distribution into testable values by multiplication. However, as we have stated throughout we are testing the distribution of the *population* not a *sample*.

In the results, the first statistical test is to test whether the means of the FPen and FP-neu δ18O distributions are different and use this to determine whether ENSO events can be detected. Comparison of the population means does not necessarily reflect differences in the population distributions, and only provides a measure of mean conditions that may or may not be related to ENSO variability. The use of the Anderson-Darling test to assess differences in distribution is used later. It is unclear how these two different tests were related, and how the mean δ18O FPen/neu was utilized.

As per a similar comment as reviewer 1, we will reword this sentence for clarity. We agree it does not necessarily reflect differences hence why we used the AD test.

The author's use of the Anderson-Darling test to assess differences in distributions is novel, but results of this test are not compared to those that have been used to assess IFA results in previous studies (e.g., std dev (Thirumalai 2013, Koutavas and Joanides 2012, Rustic 2015) or Q-Q (White 2018, Ford 2015)). Is this more sensitive, less sensitive, or does it measure different aspects of the distribution change NOT captured in the other analyses? Without such comparison, the ability to assess the validity of IFA reconstruction (the purported goal of this paper) is limited.

Different statistical techniques plot, test or validate different aspects of a sample or dataset being used. There seems to be some confusion by all three reviewers as to why the Anderson-Darling is being used. As we have already stated this before (in our reply to reviewer 1), a statistical test should be chosen by a study's author that can be used to test the research question devised by that author. We seek to investigate the (dis)similarity of distributions, therefore the Anderson-Darling test is the suitable test.

The specifics of sedimentation rate and bioturbation vary greatly across the tropical Pacific and rely on multiple processes. The role that oxygen plays in bioturbation is important, especially as bottom-water oxygen levels vary across the tropical Pacific. Likewise, seafloor topography is highly variable, with ridges and sea mounts that are not apparent at the resolution used.

On P8: "Similarly, the individual characters of El Niño events, which are very short in duration, become lost in the bioturbated sediment record " The purpose of IFA is not to discern the properties of an individual event. Change in frequency or amplitude of events over a period of time can be statistically detected using various means to compare the distribution of integrated conditions over the period of sedimentation. Bioturbation serves, then, to extend that integrated time and the range of conditions experienced.

Bioturbation causes, in the case of low SAR sediment record, thousands of years of time to become mixed into a single interval of sediment core. Hence, this may serve to mix values associated with the long-term climate signal and the ENSO signal.

With respect to the comment regarding IFA, we understand (e.g. Ganssen et al., 2011) that IFA is not a way to deduce a particular event (e.g. monsoon) but a way to characterize the samples. However, our sentence on page 8 is not saying that individual events are to be reconstructed, but that the characteristics of single event get muddled up in time. Now this would not be a problem (fundamentally) if $\delta^{18}O$ did not have a $\delta^{18}O_{sw}$ component. But $\delta^{18}O$ does have a $\delta^{18}O_{sw}$ component, thus, shells that are anomalous in one time period (e.g. LGM) may - with a change in the ice volume effect on the $\delta^{18}O_{sw}$ - have a value that is similar to 'background signal' in another time period (e.g. Holocene). Hence, the use of the word 'lost'.

> *Bioturbation will also not remove anomalous values (page 9) – rather, such values may be present as part of a distribution representing more integrated time. Likewise, bioturbation has the effect of smoothing the signal, but the "signal" is a function of all sources of variability (ENSO, annual, decadal, centennial). The relative expression of these forms of variability along with the amount of time integrated by a sample are both important in terms of the ability to capture ENSO signals.*

Anomalous values are only anomalous in relation to the rest of the dataset. As our answer to the comment above explains, if a samples anomalous values are moved into a sample with similar values then they will no longer be anomalous. We will clarify what we meant by smoothing the signal in a revised form of the MS.

> *On P.10, Cole and Tudhope (corals) and White et al (IFA) are cited in error when discussing lake colour intensity and precipitation-driven records.*

We were referring to some of the analysis within those papers (as per our comment to a similar point of reviewer 2 we will alter this sentence).

> *Also on P10, the authors claim: "If the number and magnitude of ENSO events were reduced, the relatively low downcore resolution of marine records may not accurately capture the dynamics of such lower amplitude ENSO events using existing methods."*
> *– Which methods? Q-Q, std. dev, event counting, others? It's not entirely clear this is even referring to IFA reconstructions, as the records discussed previous are sedimentary, coral, and IFA (but noted as "precipitation driven", see above).*

As our comment to reviewer 1 very clearly explained that we are not modelling IFA it is a shame that reviewer 3 did not have the time to read our replies. Here, "methods" is generally referring to proxies – we will elaborate in the revised version to reduce the confusion.

> *P.10* ***line 5****: "****The possibility of a marine sediment archive being able to reconstruct ENSO dynamics comes down to several fundamentals: the time-period captured by the sediment intervals (a combination of SAR and bioturbation), the frequency and intensity of ENSO events, as well as the foraminiferal abundance during ENSO and non-ENSO conditions.****" This statement leaves out other key elements, including the relative expression of ENSO events, the seasonal cycle, and decade-and-longer variability. These elements are (arguably) more important for inverse modeling, where the ability to disentangle growth rates from other sources of variability is impossible, and thus the signatures of ENSO in such records need to be discerned.*

[Bold is to clarify what part of the reviewers comment is a quote] – The reviewer suggests we have left out key elements. We agree that we missed out '*seasonal cycle, and decade-and-longer variability*' which we will add. But the '*relative expression of ENSO events*' is already included within the reviewer's chosen quote (underlined).

> *A key point in the paper (P10) says "The results presented here imply that much of the Pacific Ocean is not suitable for reconstructing ENSO studies using paleoceanography, yet several studies have exposed shifts within std dev(δ18Oc) of surface and thermocline dwelling foraminifera. One can, therefore, question what is being reconstructed in such studies.". This study has, at this point, not tested whether the Std.dev of δ18Oc from individual foraminifera have reconstructed ENSO (also, the wording of this sentence is odd).*

If the two populations are statistically similar (as in La Nina and El Nino have statistically indistinguishable distributions) then it is logical to question what the measure of dispersion (std dev) of the measured sample is linked to (i.e. it is seasonality, or species depth habitat change). We will rephrase in the revised version of the MS to improve readability.

> *The first paragraph of the discussion (p9) purports to be about paleoclimatological archives that "have been used to indirectly and directly study past ENSO". However, the discussion is on mean-state reconstructions (Koutavas 2003, Dubois 2009). Koutavas 2003 is non-IFA mean-state reconstruction; likewise, the Dubois 2009 paper notes that "we prefer not to invoke any ENSO-like state for the glacial EEP based solely on our UK'37 SST." While it may be true that this result and Koutavas 2003 are at*

odds, this is not an issue of IFA or ENSO reconstruction, but rather aggregate analysis and mean -state reconstruction. Discussion of std.dev ENSO studies (modeled by Thirumalai, Koutavas 2006, Koutavas and Joanides 2012, Leduc 2009, Sadekov 2013, Rustic 2015) is not found, yet the following paragraph (see above) is largely about this approach. Further, significant discussion and analysis of IFA reconstructions of ENSO during the LGM is found in Ford 2015, which is not discussed here.

The quote says "to indirectly and directly study past ENSO", one could pool papers that are non-IFA mean state reconstructions into the 'indirectly' and studies that utilise IFA into 'directly'. However, in this paragraph we are cataloguing changes ("The resultant data of such studies have been used to infer") around the Pacific. In the example the reviewer gives of Dubois, we are referring to upwelling intensification. Here is the section the reviewer is referring to in its entirety: "The resultant data of such studies have been used to infer a relatively weaker Walker circulation, a displaced ITCZ and equatorial cooling (Koutavas and LynchStieglitz, 2003); both a reduction (Koutavas and Lynch-Stieglitz, 2003) and **intensification (Dubois et al., 2009) in eastern equatorial Pacific upwelling**; and both weakened (Leduc et al., 2009) and strengthened ENSO variability (Koutavas and Joanides, 2012; Sadekov et al., 2013) during the LGM. A number of these results are contentious, for instance **the reduction in upwelling in this region (Koutavas and Lynch-Stieglitz, 2003) is contradicted by Dubois et al. (2009), who used alkenones (i.e., $U37K'$ ratios) to suggest an upwelling intensification**." We can include a section in this paragraph discussing only the std dev (reviewer comment: *Discussion of std.dev ENSO studies (modeled by Thirumalai, Koutavas 2006, Koutavas and Joanides 2012, Leduc 2009, Sadekov 2013, Rustic 2015) is not found, yet the following paragraph (see above) is largely about this approach*) although it is mentioned on page 10 lines 19-26.

The main analysis uses an unrealistic mixed-layer depth of 400m for the models foraminifera. Symbiont -bearing forams (G. ruber and G. sacculifer) live in the photic zone, and thus modeling and analysis of these organisms should be constrained to these depths.

There isn't really a 'main analysis' – we just chose the 400 m cut-off to start with. Irrespective, we know that symbiont species don't live so deep, however, the computation of FAME is such that if the temperature is appropriate a growth rate will be calculated (see Roche et al., 2018). Therefore, as we state in the paper, we varied the cut-off depths to see how these would alter the distribution. This is intended as a sensitivity study.

The model results using the shallower depths and specific, photic zone depths (Figure 4, figure 5, Figure 6) show that much of the tropical Pacific is suitable for such analyses, provided adequate carbonate preservation. This is very much in contrast with the point made previously in the paper that much of the tropical Pacific is unsuitable.

Here is the crux of our comments – we state that it is possible but, as the reviewer points out (in Bold hereafter), provided the signal can be recovered (i.e. SAR/Depth). *"The model results […] show that much of the tropical Pacific is suitable for such analyses,* ***provided adequate carbonate preservation.****"*. We will make this clearer in the revised version

In these figures, confusingly, some figures show significant areas in white while others use gray for no discernable reason. The figures are also improperly labeled, according to the captions – in each figure, G. sacculifer is on the left, G. ruber is in the middle, and this is reversed in the caption. Which is which?

We will correct this – the label on the figure is correct (left *G. sacculifer*; mid *G. ruber*; right *N. dutertrei*), we will correct the figure captions. The reviewer is correct that some are white and some grey when we originally made the figure everything was white, black and hashed. Unfortunately, these hashes draw the eye (and can hide some small locations) away from the white only locations we decided to make it grey to highlight this (The top panel of Figure 5 'G. sacculifer 60 m' is missing the grey which is our mistake but demonstrates the drawing of the eye). The N. dutertrei dataset does not have the hashing so we decided to make it white only.

The conclusions are at odds with what is presented at various points in the paper. Specifically: "Overall, our results suggest that foraminiferal δ18O for a large part of the Pacific Ocean can be used to reconstruct ENSO, especially in an individual foraminifera Analysis approach is used, contrary to previous analysis (Thirumalai et al. 2013). This conclusion is contradicted in the abstract, and in various parts of the study (e.g., P10
– "the results presented here imply that much of the Pacific Ocean is not suitable for reconstruction ENSO studies with paleoceanography. . ") Which is it?

We will rephrase these sections for clarity: reconstruction is possible when considering foraminifera only, but not necessarily possible if SAR/depth is taken into account.

Again, Koutavas 2003 is cited here, but that is not an IFA study. In general, clearly noting which studies are IFA/ENSO and which are mean state / aggregate / non-IFA

studies will clarify the discussion surrounding the use of IFA and IFA techniques to identify ENSO signals.

Again we are not doing an IFA paper, so our citations are based upon a mixture of analyses. We will consider the reviewers suggestion of stating which studies are and are not based upon IFA more clearly in the revised text.

This study does not directly address the Thirumalai 2013 study, as presented. The role of seasonality does not appear to be well addressed in this study (a key factor of Thirumalai 2013), the questionable definition of ENSO events confounds direct comparison, and the lack of clarity on sampling rates and other facts precludes a direct comparison. If this was a goal in this analysis, the Thirumalai study should be discussed in detail at the beginning (and should be discussed, in any case, earlier when discussing approaches for quantifying the suitability of locations for ENSO reconstruction), and the differences between their approaches (e.g., forward vs. inverse modeling). Suitable criteria for comparison should be noted (e.g., std. dev. Vs A-D tests).

Our goal as the research question states was not to compare our results with those of Thirumalai et al. (2013) but to *test whether the foraminiferal ($\delta^{18}O$ and Tc) distributions for climate state A and B are statistically different*. If they are different then it is theoretically possible to discern A and B when the two are pooled into the same 'bag'. However, a secondary consideration is whether at any location (irrespective of picking etc) would the signal likely be preserved, hence the SAR and water depth are two simple characteristics that have been known to palaeoceanographers. We are using a forward model, because (as explained in our answer to reviewer 1) it is not a bijective function (i.e., reversing the equation does not give you one solution) which makes it hard to do inverse modelling. We left out a discussion of inverse and forward modelling techniques because it's not really an appropriate discussion – in fact it would only be appropriate if the study had the ability to do both. With this goal in mind we decided to use a statistical method that *tests whether the distributions for climate state A and B are statistically different*. We could have tested the dataset with various other statistical tests but those would not directly answer our research question hence they would fail the reviewers "suitable criteria for comparison". Standard deviation and Anderson-Darling test different things. (AD makes a comparison of two distributions and whether they can be considered the same, whereas std dev. is a value for the spread of data around a mean). Regarding confounding and precluding comparisons: The definition of ENSO events is Oceanic Nino Index except we decided to reduce it to one that is appropriate and compatible with the foraminiferal lifecycle (as stated in our comment to reviewer 1) because of the short lifecycle they have (~1 month) and the impossibility to discern ENSO-lite (i.e. those near El Nino or La Nina events – as discussed in the reply to the reviewers comment about ENSO definition) from ENSO (this will bias both our results and any work using foraminifera). As we are testing the population, we did not therefore include a sampling/picking monte carlo style parameter, as that would be testing a sample.