B. Metcalfe on behalf of the co-authors, Response to Reviewer 1

[reviewer comments as quoted blocks]

We appreciate the effort the reviewer has taken in reviewing our paper and thank them for their interest in our work. The rapid publishing of a review on the same day as the reviewer has been nominated for review, especially given the level of detail contained within the review, must have involved great effort, and we thank the reviewer for making time for us. The reviewer raises some important points, but we believe that the reviewer may have misunderstood the goals of our paper, or that we have may not have explained them clearly enough. We would, therefore, like to take this opportunity to clear up these matters.

> *["Page 12 line 22 – What is meant by "...especially if an individual foraminiferal analysis... approach is used..." I thought the whole analysis in the paper was on whether individual foraminifera analysis can be used? Was there another method tested (for example the means analysis referred to at Page 7 line 3)? ; Page 9 line 23 – The focus of this paper is on IF analysis. Why are the Koutavas and Lynch- Stieglitz, 2003; Koutavas and Lynch-Stieglitz, 2003 etc. cited here? The whole discussion in this paragraph, lines 16-31 feels out of place."].*

We should make clear that we are principally not undertaking an evaluation of sediment-based individual foraminifera analysis, but rather an evaluation of the foraminifera populations in the water, before they have been incorporated into the sediment. We will endeavour to make this clearer in the manuscript, as we may not have done so (: we do not explicitly say 'we are modelling individual foraminifera distributions' as the reviewer assumes we are, however we also do not explicitly say 'we are not modelling individual foraminifera distributions' therefore, a clarifying sentence we will be added to a revised manuscript). Although, clearly, the question of whether foraminifera populations in the water are themselves able to record ENSO or not is important for sediment-based reconstructions, so we have included some minor discussion of sediment dynamics. Furthermore, the FAME methodology we apply does not simulate individual foraminifera, rather, it produces what the likely $\delta^{18}O_c$, T_c value for a time-step using a function that 'weights' water depths by foraminiferal growth. In other words, FAME is producing mean population values for a given time slice.

The reviewer has suggested that we carry out analysis that is outside the scope and purposes of our manuscript, for example:

> *"Furthermore, the statistical analysis focuses on a forward problem rather than the inverse problem that is the real challenge for detecting changing ENSO from individual foraminiferal analysis".*

We should make clearer, and will be glad to do so in the final version, that the purpose of our study is to determine whether or not the foraminiferal population produced under El Nino conditions are statistically different from the Neutral and La Nina conditions during the period of the observed climate record. Such a difference is an important prerequisite for any analysis: Can we detect the change we are searching for? In order to address this question, we choose a forward model as the most suitable tool. In the following reply to the reviewer, we will expand upon why the inverse method may not be suitable for our application/research question; discuss the use of the statistical methods in our paper which will hopefully address the reviewers concerns about palaeo-applications; discuss the reviewer's comments regarding validation; and answer specific questions.

Inverse Problem

> *[Focus in the inverse problem. It is really the inverse problem of detecting a change in ENSO from a change in the distribution of foraminifera d18Oc or T that is the focus of IF ENSO reconstructions. The analysis in this paper basically asks the question: are the distributions from El Niño months different from neutral or La Niña months? This is a useful first step in the inverse problem but it doesn't really answer the question stated in the title about the validity of foraminifera-based ENSO reconstructions.]*

While interesting, we believe that the inverse approach is not suitable to our particular research question. We will make this clearer in the manuscript. The inverse problem, as its very name suggests, flips the question: “we have this data what variables must have occurred to produce them”. One is inverting the scientific method to explain causal factors from observations rather than explaining observations with causal factors. The reviewer points out several papers that have done this approach, but there is a lack of large-scale analysis beyond single cores or forcing a climate model with these boundary conditions to explain inter basin variability. So, why did we not use an inverse problem, well, firstly, it has been done before, as the reviewer suggests:

> *(“This type of analysis has been done before (Thirumalai, Ford, White), with a focus on the inverse problem of estimating ENSO change from individual foraminifera distributions. Here the novelty is the inclusion of a forward model of foraminifera growth rate.”)*

And, secondly, it would not address the central question we are asking. Using the inverse problem would change our fundamental question from ‘*determining whether the foraminiferal population produced under El Nino conditions are statistically different from the Neutral and La Nina conditions*’ to an entirely different question, namely, ‘*with this dataset what magnitude and frequency of ENSO would have to have occurred to produce these observations*’. The reviewer seems to partly realise that these are not the same question:

> *(“The forward problem is whether El Nino, neutral, and La Nina months have different distributions and requires that each individual d18Oc or T value be assigned beforehand to one of those three states.“)*

Crucially and in contrast to the inverse methodology, our research question does not exclude the possibility of there being no detectable change, while the reviewer’s proposed question forces the ENSO parameters to contort into those that generate a particular dataset. In conclusion (of this point), the inverse problem approach would not give us an answer to the hypothesis and/or research question that we have chosen: **with a chosen set of input parameters (temperature, salinity) using an ecological model what would the theoretical observations of T_C or $\delta^{18}O$ be? And would the populations of El Nino, Neutral and La Nina climate be similar or different?**

In addition, for the purposes of carrying out the inverse problem, there is a lack of sediment-based data to do a large, basin-wide inverse-analysis. As we have already shown in our study (namely in the SAR and water depth/CCD plots), the seafloor of the Pacific is not conducive to providing samples with which to perform an inverse analysis on a basin-wide scale, and there is also a sampling bias, as the reviewer correctly alludes to: (“Page 10 line 28 – The discussion of model limitations does not ask what would seem to be the most important questions: Does the modeled growth rate actually reflect the real ocean (<u>and the sampling bias for what is recorded in sediments</u>)?”).

A further problem is that we would need to vary temperature AND salinity to realistically produce an inverse model, and not just temperature, as is the case when producing an accurate $\delta^{18}O_C$.

> *[Furthermore, the statistical analysis focuses on a forward problem rather than the inverse problem that is the real challenge for detecting changing ENSO from individual foraminiferal analysis. The forward problem is whether El Nino, neutral, and La Nina months have different distributions and requires that each individual d18Oc or T value be assigned beforehand to one of those three states. The inverse problem is to determine from comparison of two different d18Oc or T distributions (as would be measured in two sediment samples) whether any change in their distributions occurred and whether it can be ascribed to changes in the statistics of ENSO events (frequency, magnitude).]*

We would like to point out that neither in our response to the reviewer, nor in the paper, are we being critical of inverse modelling. We believe that it can be an appropriate and valuable technique, but it is not the suitable technique for answering our central research question. There is a fundamental difference in what the reviewer would like us to produce and what we have done (*“With different analyses the authors could address the questions they pose. However, it could be very different from the manuscript in its current form and in my opinion would need to independently evaluated and reviewed”),* namely that our paper sets out to use FAME to produce distributions using an input of temperature and salinity. The reviewer would like us to produce temperature and salinity from the distributions. However, how we should create these distributions without the input temperature parameter required for FAME is not clear. Indeed, as should be clear from the reading of the FAME methodology already published (Roche et al., 2018), there is no simple bijective relationship between the $\delta^{18}O_c$ and the oceanic variables (T, $\delta^{18}O_{sw}$).

Statistics

> *[Apply statistical tests on parameters used on paleo-IF distributions . The author's use Anderson-Darling tests for differences in distributions. They should demonstrate how this might be useful for paleo-IF analysis. It would also be greatly to their advantage to test the approaches actually used for paleo-IF analysis (1-sigma, quantiles) to see how they perform in this framework. A welcome contribution would be demonstration that a new/different type analysis from those typically applied to paleo-IF distributions is better. As it stands, the focus on the forward problem and on statistical approaches not used for paleo-IF analysis make the manuscript in its present form not a good evaluation of the IF approach for ENSO reconstruction.*
>
> *[Page 3 line 23 – Here the authors introduce the 1-sigma d18Oc parameter than has been used in some studies to look at changes in ENSO variance. But, they never really address whether this parameter is useful and can detect changes in ENSO. Thirumalai et al. (2013) took this question on already. More discussion of what has been done previously is needed. Also, why not test the actual way that IF analysis is used (e.g. 1sigma, quantiles etc.) rather than a new method as introduced here (Anderson-Darling test)?]*

To reiterate, the reason we use the long-established Anderson-Darling (AD) (1954; doi:10.2307/2281537) approach is because it is the most suitable method for the computer modelling study that we are carrying out, for the following reasons: The FAME model, coupled to the observational climate data that we have inputted, can produce high-resolution probability density functions (PDF) associated with El Nino, La Nina and neutral conditions. An Anderson-Darling test allows us to directly test if these PDFs are significantly different from one another or not. Obviously, an Anderson-Darling test may be more difficult to apply to foraminifera sampled from natural archives, where workers are limited in data resolution by the number of foraminifera that can be picked for analysis, by bioturbation of the natural archive, etc. Subsequently, in those cases, it might indeed make sense to use simpler statistics such as standard deviation and quantiles. Since we are not analysing natural archives, but rather data produced by a model for which we control what is generated, it makes more sense, in our case, to use a more powerful method such as the Anderson-Darling test.

Nonetheless, we appreciate that workers analysing natural archives are accustomed to using more straightforward statistical analyses, and would also like to see the 1 sigma and quantile intervals, so we will additionally report those for comparison in a revised version of the manuscript. Of course, these statistical parameters may not answer our research question and will not impact the answer to our research question (as AD is the appropriate test). These tests may however have flaws, given that the standard deviation is not the best descriptor of non -normal data and outside of the realms of statistics its usage assumes that the data will significantly impact the standard deviation or that standard deviation can be used as a measure of ENSO. If for instance one follows the reasoning of Mix (1987), a species may actually calcify solely in the anomaly regions (the la nina or el nino), and such species may not have a standard deviation that due to ENSO. This argument can be used for quantile-quantile, if the species does not calcify for the full year, an assumption of such an approach, it will mean that the data does not reflect the full year but a subset.

Validation of the forward model

> *[Here the novelty is the inclusion of a forward model of foraminifera growth rate. This model is used to estimate the biased sampling in depth and time that different foraminiferal species have, and how this contributes to the analysis of the ENSO signal.*

> *However, that part of the model is not really validated and it is unclear how much it adds to the analysis… Validate the growth rate calculation through comparison with sedimentary relative abundances. This was done to some extent in the paper cited for the foraminifera model (Roche et al., 2018) but in that paper no clear assessment of the errors was presented. The model in Roche et al., 2018 is a simplification of earlier growth rate modeling of foraminifera. In Roche 2017 all parameters besides temperature are discarded. How well then does the model work?]*

The model is indeed a simplification of the earlier FORAMCLIM model, because light, food etc. are not easily parameterised in models or validated with proxies. However, to claim that Roche et al. has "no clear assessment of the errors [was] presented" is not correct in our opinion. For the question relating on how the model works, the reviewer is referred to the initial publication where all the equations are described in detail; the code is itself also made available in the supplementary online material so that the reviewer can even try the model itself. While the reviewer's comments on Roche et al. (2018) relate to another publication, in the interest of discussion let us focus on how the reviewer would validate the model.

> *"I think the authors should use the modeled growth rate for the species they are targeting and calculate the relative abundance of those three species in a sediment sample. This can be compared to the measured relative abundance of those three species (summing to one) recalculated from their relative abundance amongst all species counted in coretop datasets. This should be shown as a scatterplot of observed vs. predicted on x- and y-axes rather than on a map as is shown in the supplement to Roche et al., 2018."*

Unfortunately, this would not work due to the closed sum problem. Relative abundance between species is based upon a closed sum calculation, i.e. not the true abundance as a fraction of the total foraminifera flux. Therefore, variation caused by other species not being considered/modelled has the potential to alter the relative abundance of the species being considered. In other words, if you take the relative abundance of *G. ruber, G. sacculifer* and *N. dutertrei* and sum them to 1 that would not get rid of the closed sum problem, because it fails to consider other species which are not being modelled. In fact, you would not only magnify the counting error, but you would also be basing your data on a small percentage of the total foraminifera flux. Additionally, as clearly stated in Roche et al., 2018, the FAME model is constructed so as to produce a $\delta^{18}O_c$ value where the given species of foraminifera is assumed to be able to grow: "It should be clearly understood that this approach is not able to and does not attempt to determine the relative abundances of the different species. Instead FAME provides a simplified approach to compute the δ^{18}Oc of a generic population of foraminifers if environmental conditions permit its growth. From a model–data perspective, this approach enables one to compute the calcite δ^{18}O for a given species, were it to exist in the sedimentary record ". A further useful reference in this instance is the actual formulation of the model in equation 8, page 3590 of Roche et al. (2018).

> *Page 10 line 28 – The discussion of model limitations does not ask what would seem to be the most important questions: Does the modeled growth rate actually reflect the real ocean (and the sampling bias for what is recorded in sediments)?*

As the reviewer states in this question, there is a sampling bias within sediments hence the selection of a forward model and why in this instance it is more logical than an inverse model.

> *Do the modeled growth-rate weighted d18O distributions match actual measured individual foraminifera d18O distributions (such as in Koutavas and Joanides or Rustic)? If no growth-rate weighting is applied are the results better or worse?*

As discussed earlier, FAME is not attempting to produce the measured IF distributions from natural archives.

> *Clearly separate the role that the growth model and (T,S) timeseries play in identifying ENSO change.*

> *To what degree are the outcomes and conclusions of this paper depending upon the modeled growth rates versus the sea water properties (T,S,d18Ow)? Many prior workers have analyzed in different ways the reconstruction of ENSO from IF analysis. These approaches include summary statistics like the standard deviation (Thirumalai; Koutavas; Leduc; Sadekov; Rustic), as well as examination of changes in the quantiles of IF distributions (Ford; White). What is added here is the foraminifera growth rate weighting. What effect does this have? From the histograms in Roche et al. (2018) it appears that the growth-rate weighting does not have major consequences for the mean d18Oc value of a sediment sample. It may have consequences for the IF variability though. The authors could show a map that quantifies the growth-rate weighting effect with respect to the non-weighted results (ratio, difference).*

Thank you for the comment. Figure 4 and supplementary figure S3 are already aiming at this, and we will endeavour to make this clearer. We will expand upon the section that is already included in the paper: **“The model-driven results were assessed with the underlying observational dataset**, to check **how the dataset alters with FAME** the input data (temperature and δ18Oeq) underwent statistical test**ing** (Figure 4 and Figure S3). Instead of a variable depth, we opted for fixed depths at 5, 149 and 235 m, giving a Eulerian view (Zhu et al., 2017a) in which to observe the implications of a dynamic depth habitat. By using a fixed depth, these results show that the shallowest depths produce populations that are significantly different both in terms of their mean values and their PDF. In the upper panel of Figure 4, the canonical El Niño 3.4 region is clearly visible at 5 m depth. Whilst differences exist between the temperature (Figure 4) and the FAME Anderson Darling results (Figure 3), for instance close to the Panama isthmus, there are significant similarities between the plots. These plots also show that our FAME data (Figure 3), in which we allow foraminiferal growth down deeper than the depths in Roche et al. (2018), are a conservative estimate and thus are on the low-end (Figure 4), to account for potential discrepancies with depth habitats.”

We do not fully understand the following comment of the reviewer:

> *“Page 8 line 1 – This paragraph is rather confusing to understand. It sounds like the authors are comparing a depth-weighted reconstruction and non-depth weighted reconstructions at fixed depths (Fig. 3 vs. 4)?”.*

In the way FAME is set up, the weighting for depth is based upon growth, without FAME we would be unable to integrate the required weighting function. Hence why it necessitated fixing the depth for the analysis of the input temperature and $\delta^{18}O_{eq}$ values.

Validation of $\delta^{18}O_C$

> *[“Validate the $\delta^{18}O_c$ predictions from the growth rate and geochemistry model. This was done in Roche et al., 2017 but is also somewhat circular because the sedimentary $\delta^{18}O$ values were used to determine the depth of production. I admit I am not sure how to actually validate the approach except from an additional validation dataset not used for determining production depth.*
>
> *Page 5 line 10 – Why was growth rate arbitrarily constrained to these different depths? First, foraminifera with algal symbionts should be in the photic zone. Second, didn’t the Roche et al., 2018 paper try to identify the depth-production relationship for the different species from the predicted $\delta^{18}O_c$ and measured MARGO $\delta^{18}O_c$? Why not use those depths?”]*

Thank you for your comment. It is important to stress that there is no geochemistry model in our approach. The optimisation procedure of Roche et al. (2018), gives the maximum allowed growth depth of each species. However, as the reviewer discussed previously, we should test how this influences the resultant distributions we generate, therefore we constrained the model to four depths including to a depth known to be below the photic zone. This is what is stated in page 4 line 7-19; page 5 line 10. We will rephrase this with clarity in mind.

Dismissive of Mg/Ca-Temperature?

[Include the analysis of Tc for Mg/Ca reconstructions. Inexplicably the authors refuse to analyze the temperature distributions even though those are the data from the common Mg/Ca method of individual foraminifera analysis (Sadekov, Ford, White). The author's stated reason is due to "...the complexity in reconstructions of trace metal geochemistry...and the potential error associated with determining which carbonate phase is first used when foraminifera biomineralise...". While there are ongoing methodological and calibration efforts for this and other proxies (including d18Oc), to ignore such a widespread type of analysis seems very shortsighted. If the authors do not want to forward model the Mg/Ca proxy itself they can simply analyse the temperatures in their dataset. Either way this is something that should be included in the manuscript.

Page 9 line 20 – Why discount trace metal temperatures (Mg/Ca)? Include the analysis of Tc for Mg/Ca reconstructions.

Page 11 line 14 – Why are the authors so dismissive of Mg/Ca analyses? The list of possible complications is important but it remains a fundamental observation that the Mg/Ca of foraminiferal calcite changes with growth temperature and has been validated in many different ways.]

The reviewer is alluding to Figure 6, the Tc or calcite/recorded temperature, which is essentially our pseudo-Mg/Ca* produced with FAME (* = It is a weighting of temperature rather than $\delta^{18}O_{eq}$). This is discussed in the dataset, we also ran the temperature (Figure 4) of the dataset by itself (without the foraminiferal growth rates). It is true that we don't go into too much detail and we will expand our section discussing the FAME produced temperature (Tc).

However, it is not as the reviewer states as us being 'so dismissive of Mg/Ca analyses' (we are sorry if we created such an impression). A great many researchers are dedicating their time to this valuable geochemical analysis. However, given that the species-specific conversion from temperature to Mg/Ca is not as straight-forward as $\delta^{18}O_c$ and $\delta^{18}O_{eq}$, it would therefore require more parameters, which we are not confident in modelling at this stage. A pseudo Mg/Ca would also need to be validated (yet techniques are not standardised nor cross calibrated to a sufficient degree, with users using laser ablation; pooled specimens and/or whole shell) and the problems associated with dissolution, cleaning for analysis, are not easily parameterised in a model. We do welcome discussions on the computation of pseudo Mg/Ca and consider it something that could be included in the future, possibly in a second generation of the FAME model.

Removal of maps

[Remove maps of carbonate preservation/depth. It is fine for the authors to state the general problem in the text, but there are regions of shallow depth were carbonate is preserved that are not captured in the coarse DEM used; Remove map of sedimentation rate. Either quantitatively discuss the role of sedimentation rate and bioturbation or remove this map. The sedimentation rate threshold is intimately tied to the secular and nonENSO variability and thus is a much more complicated analysis than the general discussion in the text. I think the discussion is a starting point but the author's miss that the important factors are really the magnitude of other, non-ENSO sources of variability at the timescale of a sediment sample (plus bioturbation) compared to the magnitude of the ENSO change signal and the non-ENSO variability.]

As stated in the paper: - "The resolution of the ocean reanalysis data for the time period 1958-2015 would essentially be analogous with a sediment core representing 50 yr-1 cm-1 (or 20 cm-1 kyr-1). Based on our analysis, such a hypothetical core with a rapid sediment accumulation rate (SAR) could allow for the possible disentanglement of El Niño related signals from the climatic signal using IFA, but only in a best-case scenario involving minimal/no bioturbation, which is unlikely in the case of oxygenated sediments". This is an important caveat to communicate to the reader. We believe that removing the maps would remove this valuable piece of information.

Definition of ENSO components

["Definition of El Niño, neutral, and La Niña months there is a large body of literature and accepted methods for defining El Niño, neutral, and La Niña periods. In the text the authors take a simplistic approach, but there is no reason for this. Why not actually use the societal

> *and dynamically important definitions of these events including the requirement of a minimum consecutive number of months of anomalies and changing baseline for anomalies (to account for secular warming of the ocean)? This definition has a basis in theory as an El Nino (La Nina) event unfolds over a length of time and thus a single month anomaly may not be associated with the dynamics that are part of the coupled ENSO system."]*

What we said: "The tropical Pacific Ocean is divided into four Niño regions based on historical ship tracks, from east to west: Niño 1 and 2 (0° to -10°S, 90°W to 80°W), Niño 3 (5°N to -5°S, 150°W to 90°W), Niño 3.4 (5°N to -5°S, 170°W to 120°W) and Niño 4 (5°N to -5°S, 160°E to 150°W). One index for ENSO, the Oceanic Niño Index (ONI), based upon the Niño 3.4 region (because of the region's importance for interactions between ocean and atmosphere) is a 3-month running mean of SST anomalies in ERSST.v5 (Huang et al., 2017). However, Pan-Pacific meteorological agencies differ in their definition (An and Bong, 2016, 2018) of an El Niño, with each country's definition reflecting socio-economic factors, therefore, for simplicity we utilise a threshold of $\chi \geq +0.5°C$ as a proxy for El Niño, $-0.5°C \leq \chi \geq +0.5°C$ for neutral climate conditions and $-0.5°C \leq \chi$ for a La Niña in the Oceanic Niño Index. Many meteorological agencies consider that five consecutive months of $\chi \geq +0.5°C$ must occur for the classification of an El Niño event. However, here it is considered that any single month falling within our threshold values as representative of El Niño, neutral or La Niña conditions (grey bars in Figure 1). By using this threshold, three weighted histograms for each δ18Oc and Tc and their resultant distributions (El Niño; Neutral; and La Niña) were computed for every month and for every latitude and longitude grid-point for the 1958-2015 period."

Why did we do this? Because if a foraminifer lives for 30 days then how appropriate would *"including the requirement of a minimum consecutive number of months of anomalies and changing baseline for anomalies (to account for secular warming of the ocean)"* be? Because sediments can't resolve annual/sub-annual resolution (like corals or molluscs), therefore the periods where the threshold passes 0.5 would in the sediment be mixed with with El Nino or La Nina (as in they would have potentially similar values as an El Nino, and as time cannot be resolved they would be considered as El Nino). In the sediment the minimum consecutive months is not used to define an El Nino, as it is impossible, an arbitrary value or any kind of quantile- or sigma distribution is.

Next paper.

> *["Examine how ENSO amplitude vs. frequency change IF distributions. The authors raise an interesting point in their conclusion that has not been well addressed, namely how do changes in the statistics of ENSO (frequency, amplitude) affect IF distributions and reconstructions of ENSO variability. Evaluating these two different questions would be an important contribution to IF analysis of ENSO change. But, introducing the idea in the conclusions without a previous discussion in the manuscript is not a good idea in my opinion."]*

We are not attempting, at this stage, to specifically reproduce single foraminifera analysis. How ENSO amplitude and frequency impact foraminiferal distributions is a separate paper we are working upon, because it actually cannot be dealt within as a simple discussion topic (and would require a specific and different dataset from the current paper's dataset), it is something that we thought about as we worked on this manuscript. Therefore, we suggested this approach in our conclusions / perspectives as something that could be worked on in the future, which is something that we believe is normal in scientific manuscripts. We will attempt to make clearer that we are referring to possible future work.

Specific comments

> *Page 1 line 17 – "Furthermore, a large proportion of these areas coincide with sea-floor regions exhibiting a low sedimentation rate and/or water depth below the carbonate compensation depth, thus precluding the extraction of a temporally valid palaeoclimate signal using long-standing palaeoceanographic methods." The role of sedimentation rate in IF analysis is important but there is not any investigation of this effect in the present manuscript so it is not really a conclusion or finding. This statement should not be included*

> *in the paper in its present form; Page 1 line 17 – "Furthermore, a large proportion of these areas coincide with sea-floor regions exhibiting a low sedimentation rate and/or water depth below the carbonate compensation depth, thus precluding the extraction of a temporally valid palaeoclimate signal using long-standing palaeoceanographic methods." The role of water depth and carbonate preservation is also important. But, there is not any investigation of the sedimentation rate effect in the present manuscript so it is not really a conclusion or finding. Furthermore, there are seamounts and other shallow sites not captured in the gridded dataset that can contain records for palaeoceanographic investigations. This statement should not be included in the paper in its present form*

We disagree, we believe that the inclusion of SAR and water depth (CCD) adds important context to our paper. Those are the two main factors that allow for the carbonate signal to be preserved in sufficient temporal resolution. We can consider adding the location of sea mounts to the map, thank you for this idea.

.

> *Page 4 line 1 – The new model for foraminifera growth only uses the temperature component of the previous model. Why? How different are the results?*

The 'why' have been dealt within in Roche et al. (2018). The 'how different' is comparing apples and oranges, FAME requires temperature as an input whereas FORAMCLIM needs temperature, light, and organic carbon (food). Light and food are not included in many datasets, nor are they parameterised or have proxies. A validation step of the two different models using the same observational input data is thus not simply attainable. The input data here does not have either of these additional variables (we considered for example using a long term chlorophyll record from satellite data, but such datasets ignore the deep chlorophyll maximum).

> *Page 4 line 15 – Allowing symbiont-bearing foraminifera to possibly grow to 400 m simply based upon optimal temperatures seems not correct. They need to be in the photic zone.*

Four different depths (60; 100; 200 and 400 m) have been used in the model, the use of the shallow and deeper depths likely don't capture one or more of the species actual ecologies, however that is why we ran it with different depths to understand how chosen depth alters (or doesn't alter) the results.

What we said: - "Consequently, we allow all the species of foraminifera to grow down to ~ 400 m (depending if optimal temperature conditions are met) to capture the total theoretical niche width. As the optimised depths of Roche et al. (2018) are shallower, and upper ocean water is more prone to temperature variability, our approach likely dampens both the modelled δ18Oc and Tc. Therefore, the sensitivity of the model was tested by applying the same procedure but with the limitation of the depth set to 60; 100 and 200 m."

–Methods–

> *Page 5 line 5 – The conversion of VSMOW to VPDB looks to be in error. The correct formula for this conversion is [d18O_VSMOW+1]/[d18O_VPDB+1] = 1.03091 where d18O does not include the 10ˆ3 term. Thus d18O_VPDB = d18O_VSMOW/1.03091 +(1/1.03091)-1 or d18O_VPDB = 0.97002*d18O_VSMOW 0.02998. In d18O expressed with the 10ˆ3 term, the equation would read: d18O_VPDB = 0.97002*d18O_VSMOW - 29.98.*

The reviewer is referring to the fractionation difference between water and carbonate for which the given expressions are indeed correct. However, what is referred to in the manuscript is the conversion between two scales with measured $\delta^{18}O_{sw}$; those can be converted from V-SMOW to V-PDB by – 0.27 ‰ (please see Figure 1 in Hut, 1987 for the original reference).

> *Page 6 line 1 – "...these for now can be ignored." Why can the other factors determining foraminifera growth be ignored? This cannot be a statement unless it is backed up. Or, the authors use only temperature but then go through an appropriate validation process (more than what is shown in Roche et al., 2018) as suggested above.*

We agree with the reviewer that our initial sentence was somewhat ill-formulated. What we meant here is that the major driver of foraminiferal growth is temperature and hence taking it (temperature) into account will provide the first order signal, as was discussed already in Roche et al. (2018). A revised version of the manuscript will include a modified statement as follow: "these variables for now can be set aside as temperature provides the dominant signal, it is worth noting that in all probability some variance will arise from these processes and deviation between observed and expected values should consider this." Regarding the validation process we refer the reviewer to the discussion already given above.

> *Page 6 line 11 – Starting here, it is very unclear how and why the particular set of conditions for El Niño, La Niña, and neutral periods were chosen. What time series of sea surface temperatures were chosen for computing anomalies (in each grid square, Nino 3.4, Nino 3, Nino 4, etc.)? Were the anomalies based upon a 3-month running mean? Were the anomalies computed relative to a fixed period or, as is now the accepted approach, relative to 5-year interval means? Why not use the definition of El Nino etc. events that include the requirement for consecutive months of anomalies? This definition has a basis in theory as an El Nino (La Nina) event unfolds over a length of time and thus a single month anomaly may not be associated with the dynamics that are part of the coupled ENSO system.*

Pg. 6, Line's 8 to 10. The ONI dataset was used, this is based upon the 3 month smoothed anomaly in the El Nino 3.4 region, we decided to set thresholds including anything above 0.5 and below -0.5 in their respective EL Nino and La Nina bins because unlike, e.g. corals, a palaeoceanographer using sediment core foraminifera cannot discern a specific year. An 'almost El Nino' anomaly won't be discernible from a full El Nino period in the fossil record, because unlike coral records we cannot determine what the previous 3 months were like.

> *Page 6 line 18 – Why and how was the pdf/cdf from the actual data fitted and smoothed with an Epanechnikov kernel? What impact did this fitting and smoothing (particularly the choice of bandwidth) have on the Anderson-Darling test and the results overall?*

The data was fitted using a fit distribution procedure in MatLab because the statistical function requires a distribution to test. We chose to use the kernel distribution because it mimics the underlying dataset well and we were testing a large number of grid points, therefore we decided to keep numerous parameters constant (for instance we could have decided to change the distribution using a find the best fit distribution but this would have made intercomparison problematic), however to allow our fitted distribution to better mimic the underlying distribution we allowed the programme to vary the bandwidth between grid points for an optimal kernel distribution.

> *Page 6 line 24 – This paragraph is very unclear and the errors associated with binning prior to analysis of the pdf seem avoidable. For example, why not take the growth rate in each of the 696 months in each grid at each depth, and scale the growth rate to calculate an effective # of individuals such that they sum to 1000 across all months? Round those numbers to integers and then use the integer # of individuals for each month to replicate that actual months Tc or d18Oc value. The resulting ordered list of values can then be binned/smoothed etc. and represents a pseudo-distribution that one might find in a sediment sample?*

The reviewer is correct that it would solve the minor binning error – but it wouldn't solve the rounding error, if you round these numbers….

What we wrote:- "As the weighted distributions are effectively probability distributions, in order to fit a distribution, we multiplied the bin counts by 1000, effectively converting probability into a hypothesised distribution. Using the repeat matrix function (MatLab function: repmat), a matrix of δ18Oc was produced using each bin's mid-point (δ18Omid-point) there is a threefold error combined with this methodology which may account for minor variation between discrete runs of the model: first the counts values were rounded to whole integers so an exact number of cells could be added to a matrix; secondly the δ18Omid-point was used which gives an error associated with the bin size (±0.05 ‰) that is symmetrical close to the distributions measures of central tendency but asymmetrical at the sides; and finally, the associated rounding error at the bin edges within a histogram (±0.005 ‰)."

–Results–

Page 7 line 3 – It says that the mean d18Oc for El Nino and neutral months are compared. How? Earlier and later it is stated that the A-D test is applied to compare distributions. What is meant by these lines?

We will reword this sentence for clarity.

Page 7 line 5 – "...ENSO events can potentially be detected by paleoceanographers and unmixed using, for example, a simple mixing algorithm with individual foraminiferal analysis..." This is not really practicable because it assumes complete stationarity in the El Nino, La Nina, and neutral distribution. This is unlikely as all are expected to change, and do in models and data (e.g. coral time series from middle Holocene show changed seasonal amplitude and ENSO cycles).

Here we are discussing an unmixing analysis for a single time slice, if enough foraminifera are measured then it can be possible to disentangle mathematically the various components that go into a single distribution. However, this is only possible if the values of El Nino, La Nina etc. have a different absolute δ^{18}O value, our point here. This is true regardless of an unmixing analysis or and holds true for any proxy.

Page 7 line 7 –"In cases where FPEN and FPNEU do not exhibit significantly different means, then the chosen species and/or location represent a poor choice to study ENSO dynamics." This may not always be the case because the mean values could be similar but the distributions wildly different (such as long tails with different signs). Changing numbers of El Nino and neutral and La Nina events could that quite dramatically change the shape of the combined distribution that is ultimately preserved in sediments. And, it may be possible to find regions of such a distribution that can be used to diagnose changing ENSO.

Page 7 line 20 – Why is Anderson-Darling test done here but the mean values are discussed above? If the A-D test shows that the El Nino and Neutral distributions are different (at some statistical level) then that means alteration of those distributions (more/fewer, stronger/weaker events) would alter the summed distributions that one gets from a sediment sample. But, how would this actually be detected in the sediment sample? That the AD test demonstrates the El Nino, Neutral, and/or La Nina distributions are different is helpful but it does not get at whether ENSO change could actually be detected in a sediment sample.

True, that is why we tested the distributions as well. Here we are discussing the fact that similar values would be impossible to unravel – we will make this clear. This *("Changing numbers of El Nino and neutral and La Nina events could that quite dramatically change the shape of the combined distribution that is ultimately preserved in sediments. And, it may be possible to find regions of such a distribution that can be used to diagnose changing ENSO")* is why we chose to use a statistical test that looks into the distribution.

Page 7 line 26 – Applying a 1-sigma value from modeled minus coretop comparisons to the AD test value does not seem appropriate. This value assesses the accuracy of the model in predicting the absolute value of the mean of a coretop sample. But it is not an appropriate estimate for the significance of the difference between two different IF values or the difference in the AD statistic.

We disagree, if the model has some measurable error, it is appropriate to advertise the fact to readers that at some locations the distributions whilst significantly different with one test fall within the model 'error'. Hence the use of hashing.

Page 8 line 9 – Unclear what "on the low-end" means.

We will clarify this.

> *Page 8 line 16 – "...a large percentage of the tropical Pacific remains accessible to palaeoclimate studies." This is very much not the message in the Abstract and from the title of the paper. Those sections should reflect this finding.*

First the title is "On the validity of foraminifera-based ENSO reconstructions" is ambiguous as to whether they are or are not valid. Second the abstract is referring to the entire discussion, the calculated distributions and the SAR/depth. We will clarify this statement.

> *Page 8 line 25 – "Indeed, one should view discrete sediment intervals, and the foraminifera contained within them, as representative of an integrated multi-decadal or even multi-centennial signal..." This is exactly how foraminifera paleo-IF studies have viewed them and should be stated up front (start of paragraph for instance).*

Yes, we will make this clearer. We note that we cite our individual foram paper (Lougheed et al, 2018) using single shell ^{14}C to show that a single cm can be not just multi-centennial, but multi-millennial.

> *Page 8 line 28 – "Therefore, in order to reliably extract short-term environmental information from foraminiferal-based proxies, the signal that one is testing or aiming to recover must exhibit a large enough amplitude in order to perturb the population by a significant degree from the background signal, otherwise it will be lost due to the smoothing effect of bioturbation..." This statement does not make sense to me. The background signal IS the signal, i.e. the seasonal cycle, ENSO etc. Changes in ENSO must be such that they alter that signal (the distribution of IF analyses), but bioturbation etc. should not erase the signal unless one is looking for short periods of change less than the time integrated into the sample.*

The reviewer has stated what we are stating in that sentence, we will rephrase for clarity. However, ENSO is not the background signal otherwise it would not be detectible through a temperature anomaly. It is a short term climatic event when one considers that a single cm in ocean sediments can reflect hundred to thousands of years.

> *Page 9 line 6 – "...a series of high magnitude, but low frequency El Niño events could be smoothed out of the downcore, discrete-depth record." They will not be smoothed out as the authors state. Those anomalous IF values may be rare, but will be present in the sediment sample and if measured can be used to examine changing ENSO.*

Thank you for pointing this sentence out. We will clarify this sentence to explain in better detail what we mean. Firstly, the absolute magnitude of events would obviously be smoothed out if one were to be apply discrete, multi-specimen sample downcore analysis (i.e. not single foram analysis), as the reviewer is obviously aware of. Were single foram analysis to be applied, the single foram values corresponding to high-magnitude ENSO events would indeed still be present in the sediment record, as the reviewer correctly points out. However, single foraminifera from multiple ENSO events, and non-ENSO climate, would all be mixed into the same discrete interval, meaning that a time-series of ENSO is essentially not possible to produce, and therefore: (1) the frequency of ENSO events becomes difficult to detect, and (2) that one is forced to make *a priori* assumptions regarding the behaviour of background climate and ENSO climate in past times in order to differentiate between ENSO and non-ENSO single foraminifera in the palaeo record.

> *Page 9 line 7 – The sediment accumulation rate needed to observe/reconstruct changes in ENSO is not fixed. It depends upon the magnitude and duration of secular trends, and variability with respect to both the time integrated in a sediment sample and the magnitude of the ENSO signal and its change. This is a quite interesting but also complicated subject and arbitrarily cutting the sedimentation rate at 5 cm/ky is not justified.*

We agree with the reviewer it is not fixed, the '*sediment accumulation rate needed to observe/reconstruct changes*' ideally would reflect the percentage of foraminifera within the sediment growing during ENSO events and the magnitude of the events not just the number of events. We furthermore note that, to avoid using a SAR cut off that could be considered arbitrary, we intentionally used a very generous cut-off of 5 cm/ka. Were we to set the cut-off to be higher, following the more traditional lower cut-off of 10 cm/ka (Bard et al, Shackleton et al), then the areas of the Pacific basin that could be considered suitable would be even more limited.

> *Page 9 line 9 – The map of water depth is quite coarse and misses important locations that are above the CCD, accumulate carbonate (and foraminifera), and can be used for palaeoceanographic reconstructions. Thus, while the overall point is true, the map as shown is misleading.*

We used the latest GEBCO, but we would be more than happy to include higher resolution data.. However, we are obviously not saying a seamount would not be useable. We can consider adding sea mounts to the map.

> *Page 10 line 3 – The references to Cole and Tudhope, 2017; White et al., 2018 seem to be in error. These papers do not discuss lake core colour etc.*

We will rephrase this sentence. Here we referring to the interpretation, for instance figure 19.3 of Cole and Tudhope (2017)

> *Page 10 line 3 – "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." This statement is not justified by the author's analysis or a citation.*

We will add a citation(s).

> *Page 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." Also included is the magnitude of change in ENSO statistics and resulting foramifera Tc or d18Oc, sampling uncertainty on the IF distribution. See also note above on the role of sedimentation rate.*

At the reviewer's suggestion we will add in 'sampling-bias' into the sentence

> *Page 10 line 9 – "The results presented here imply that much of the Pacific Ocean is not suitable for reconstructing ENSO studies using palaeoceanography, yet several studies have exposed shifts within σ(d18Oc) of surface and thermocline dwelling foraminifera. One can, therefore, question what is being reconstructed in such studies." The results presented here don't really test whether individual foraminifera d18Oc (or Tc) studies can reconstruct ENSO. Furthermore, the water depth and sedimentation rate constraints are the reason for excluding much of the Pacific. This statement is therefore incorrect and the search for other explanations does not follow.*

This sentence may have led the reviewer to misinterpret our results as sediment-based individual foraminiferal analysis centric, we will rephrase this sentence for clarity and suitability.

> *Page 10 line 19 – This second part of the paragraph is interesting and has been commented on before. But, at no point do the authors actually evaluate any of these effects or approaches so they can't really assess the different factors they raise here.*

Here we are discussing other's findings, for instance, Zhu computed the variance and found that some of the signals detected could be a by-product of the annual cycle.

–Conclusion–

> *Page 12 line 17 – "Previous work..." The only citation here is to Zhu et al., 2017. There has been a lot of work comparing IFA different time slices (both d18Oc and Mc/Ca) that should be cited here (Koutavas et al., Leduc, Koutavas and Joanides, Sadekov et al,Ford et al, Rustic et al, White et al). Furthermore, they have not all used 1-sigma d18Oc as the metric for detecting change.*

The reviewer is right, "*they have not all used 1-sigma d18Oc as the metric for detecting change*", that is why they are not cited. The reviewer would be justified in suggesting these references here, had we not repeatedly cited them throughout our paper. Whilst we will attempt to make this clearer for the reader, it is worth noting that a few sentences later, we directly refer to the papers the reviewers cites (see comment below:) .

> *Page 12 line 21 – "Overall, our results suggest that foraminiferal δ18O for a large part of the Pacific Ocean can be used to reconstruct ENSO..." This contradicts what is said in the abstract and in some places in the text (but is similar to in other places in the text). Which is it?*

We will add clarity to this statement, however we would like to note that this quote neglects the second part, the first word overall here being 'generally speaking' eludes to the fact that it's a sentence that has a follow up: "Overall, our results suggest that foraminiferal δ 18O for a large part of the Pacific Ocean can be used to reconstruct ENSO, **especially if an individual foraminiferal analysis (Lougheed et al., 2018; Wit et al., 2013) approach is used (Ford et al., 2015; Koutavas et al., 2006; Koutavas and Joanides, 2012; Koutavas and Lynch-Stieglitz, 2003; Sadekov et al., 2013; White et al., 2018)**, contrary to previous analysis (Thirumalai et al., 2013). **However**, the sedimentation rate of ocean sediments in the region is notoriously slow (Olson et al., 2016) and much of the ocean floor is under the CCD. **These factors reduce the size of the area available for reconstructions considerably (Lougheed et al., 2018), thus precluding the extraction of a temporally valid palaeoclimate signal using long-standing methods."**

> *Page 12 line 24 – "However, the sedimentation rate of ocean sediments in the region is notoriously slow (Olson et al., 2016) and much of the ocean floor is under the CCD. These factors reduce the size of the area available for reconstructions considerably (Lougheed et al., 2018), thus precluding the extraction of a temporally valid palaeoclimate signal using long-standing methods." This is generally true, but there are seamounts and other regions that may actually preserve carbonate. Furthermore, the sedimentation rate constraint is also somewhat arbitrary and depends upon secular trends and non-ENSO variability encompassed in a particular sample.*

As the reviewer states our statement is 'generally true', therefore the difference is our interpretation vs. the reviewer's interpretation, we politely disagree.

> *Page 12, line 27 – "We further highlight that the conclusions drawn from foraminiferal reconstructions should consider both the frequency and magnitude of El Niño events during the corresponding sediment time interval (with full error) to fully understand whether or not a strengthening or dampening occurred." While this is true, nowhere in the manuscript is this issue addressed. Inclusion as a conclusion to the paper is therefore not warranted; Page 12 line 30 – "The use of ecophysiological models...are not limited to foraminifera and provide an important way to test whether proxies used for palaeoclimate reconstructions are suitable for the given research question." This is not really a conclusion of the study. And, given the uncertainties and lack of rigorous testing of the foraminifera model in this study, this is a questionable statement overall.*

Conclusions do not have to include the main focus of the study but can include information that present the findings in a different light (as in what it means to the readers) or what can be next done. We don't agree with the reviewer's suggestion that it is an untested model, but also doubting that Ecophysiological models are not limited to foraminifera – this is a factual statement and they can be of use.

–Figures–

Where we have used FAME they are growth rate weighted values – this is explained in the methodology and Roche et al. (2018). Figure 4 and supplementary figure 3 use the input data therefore they are non growth rate weighted.

> *Figure 3 – Why are there white and grey areas that mean the same thing?*

As the key shows they represent where the populations are statistically different, the hashing draws the eye too much so for those panels with hashes we make it grey. As one of our species does not have these hashes it remains white.

Figure 4 – Are the temperature data growth weighted? What species? If not, why not analyze the Tc data in parallel to the d18Oc data to evaluate what advantage/disadvantage the two different signals have (e.g. from S).

What the caption says: - “Figure 4. Results of an Anderson-Darling test between El Nino and Neutral climate conditions based upon the Temperature input data: Fixed depth.” This is the temperature input data

Figure 5 – Why are the white and grey areas grouped together? What do they mean? Are these panels based upon growth-rate weighted values?

As the key shows they represent where the populations are statistically different, the hashing draws the eye too much so for those panels with hashes we make it grey. As one of our species does not have these hashes it remains white.

Figure 6 – Are these panels based upon growth-rate weighted values?

Where we have used FAME they are growth rate weighted values