B. Metcalfe on behalf of the co-authors, Response to Reviewer 2

[reviewer comments as red text in blocks]

We thank Reviewer 2 for their time in submitting this review of our work, before we address the main and minor comments we like to (re)address a point raised by Reviewer 1 and expand upon application of our study.

> In their manuscript Metcalfe et al. present a forward modeling approach through FAME to investigate the use of individual foraminifera analysis (IFA) for ENSO reconstruction. Based on the modeling results, they conclude that this proxy is only valid in part of the Pacific Ocean. However, these regions are often characterized by low sedimentation rate, therefore limiting the use of this proxy.

The reviewer correctly identifies that we have carried out a forward model of foraminifera $\delta^{18}O$. Specifically, we used the FAME model, driven by observed climate data, to predict foraminifera population $\delta^{18}O$ accounting for their habitat water depths and growth season, to test whether foraminifera populations in the water can record the ENSO signal. We are not modelling IFA in sediment cores. We carry out a minor discussion of sediment core IFA, because, clearly, whether or not ENSO dynamics can be recorded by foraminifera populations in the water itself has consequences for sediment core records (which obviously source their foraminfiera tests from the water column).

> While the effort to incorporate forward models into paleoceanographic studies is commendable, I fail to see the practical application of this study.

The reviewer might prefer if our study was more similar to an inverse model approach, however, the manuscript that we submitted is a forward model. We regret if we have not elucidated enough in our paper the practical applications of our forward model. We see several applications:

Research question / proxy validation – Pg. 12 ln 30-32 "The use of ecophysiological models (Kageyama et al., 2013; Lombard et al., 2009, 2011) 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." One application can be validating the research question, i.e., whether foraminifera in the water will in the first place (even before they become fossilised, disturbed, winnowed away) record the signal we are looking (e.g., similar to Leduc et al., 2009; Ford et al., 2015). A pre-screening using the available information (temperature, salinity, etc.) and our understanding of certain processes can be used to focus our research, determine species selection. Which leads on to…

Site selection – In an ideal world the location of (palaeo-) data would be at least one datapoint every 1x1 degree latitude longitude (or a higher resolution) so that a direct comparison between model and data could be obtained. However, this is neither practical nor feasible in the real world where access to time and funding (etc.) is limited. The time and effort (from personal experience) that goes into sampling, washing, picking, measuring (and dependent upon pooled or individual, replicates or no replicates, etc) foraminifera is a lot. Forward models can therefore be used to select sites not on geographic proximity to existing published records but based on their strategic location, which can provide critical information about certain climatological processes for data-model comparison. Naturally, a bioturbation module would further elucidate such an approach. There are three models/analytical tools that exist that deal with picking and signal modulation: Sedproxy; FIRM or INFAUNAL. Therefore, we focussed upon the construction of a signal, which leads on to….

Sandbox modelling - Pg. 12 ln 26-27: "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. FAME is intended for climate models where boundary conditions can be varied.

> Inverse modeling would be impossible and the lack of comparison between the pseudo-proxy distributions and actual distributions of foraminifera prevents validation of the method.
>
> **Major comments**
>
> **Inverse problem**
>
> The manuscript focuses on forward modeling of IFA analysis. Although definitely a valuable exercise for data-model comparison (assuming that the climate model can make use of the forward model), it doesn't solve the inverse problem. It would be almost impossible to evaluate the growth factor in the d18O record.

Our commentary regarding the inverse problem is available with respect to Reviewer 1, we will surmise our key points here: (i) It is not suitable/applicable to our research question; (ii) a lack of a large scale dataset to perform a basin-wide analysis; and (iii) lack of a bijective relationship between $\delta^{18}O_c$ and the oceanic variables (T, $\delta^{18}O_{sw}$).

"(assuming that the climate model can make use of the forward model)" - FAME is built with climate models in mind. Data-model comparison studies suffer from an 'apples and oranges' problem, of which there are two key problems: (i) the conversion of units i.e., most proxies reconstructing temperature do not give values of temperature in degrees C or K but in their own proxy units (per mil, mmol/mol, etc) requiring a conversion, and (ii) a reduction in scales, i.e., models give a wealth of information (multiple layers) in the time-depth domain. FAME was produced (Roche et al., 2018) to (i) to generate pseudo-proxy time-series from model runs that can be compared with age-depth values down cores. Naturally, including a bioturbation, or mixing, module into down-core work is prudent for core datasets (please see comment below; and to (ii) reduce the information for a given time-slice into a manageable value using an integration that would make sense on a biological point of view (integrating various depths with equal weightings might seem logical, but foraminifera for instance grow at different rates depending on their temperature) and).

> It's also not visually obvious what the difference between the output of a non-weighted model is vs FAME in Figure S1. Some statistics would help, or plotting the resulting kernel distributions on a separate panel,

We will elaborate further on these plots to make them more understandable, Figure S1 are produced solely by FAME, the difference is one is weighted for a larger proportion of growth per month and the other bins the total number of months, these plots were picked at random from the dataset.

> Further, bioturbation is also likely to have a large impact on IFA, especially in areas of low sediment accumulation. Why not connect FAME to a bioturbation model and disentangle the influence of these factors?

There are three points we would like to address with this comment, the first is the reviewers admission that IFA (and by association pooled) distributions can be impacted by bioturbation. The second point is that it asks the follow-up question to our own research question, i.e. "Are the populations between different climatological end members significantly different, **and can this difference be resolved in the sediment record**", this second question would need different input parameters, as discussed next. The third point we would like to address is that our input dataset is ~60 years, in the grand scheme of things a relatively small contribution to the sediment. We choose not to link the dataset to a bioturbation module because the time series is not long enough, a 1cm kyr^{-1} SAR with a 10 cm mixed layer depth would need several thousand years of input data (please see comment further below). Furthermore, the application of bioturbation to a monthly time-series can be done in several ways, for instance should one could bioturbate the 'settling flux', i.e. the monthly signal, as soon as it is encoded or at the end of the time-series.

Statistical analysis

> Page 6, Line 25: Multiplying the bin counts will effectively skewed the results of a significance test. In practice, it would be impractical if not impossible to obtain 1000 samples in each bin.

We will reword this section as whilst we state 'total bin counts' we add a clarification: - "As the weighted distributions are effectively probability distributions, in order to fit a distribution, we multiplied the total bin counts by 1000 **(i.e. so that the total sum is 1000).**

> Similarly, page 7, line 4, how many foraminifera were artificial picked to produce these maps?

We did not artificially pick foraminifera (here we are referring to population rather than a simulated sample). We are testing the water populations that would be El Nino or La Nina or Neutral. As we are not modelling IFA in sediment cores, we therefore also do not include a picking routine.

IFA model - data comparison

> There are a number of recent studies with IFA results from the past 1000 years (some of them cited in the current manuscript). How do these distributions compare to the statistical ones?

We will address this in a revised form of the MS.

Effect of SAR

> Since a model of bioturbation was not implemented here, it's hard to examine the effect of bioturbation on the IFA.

We have indeed not specifically modelled the effect of bioturbation upon our data, because it is simply not possible to carry out a transient bioturbation model run upon 60 years data. Therefore, to avoid appearing biased against sediment archives in our SAR map, we intentionally used a very generously low SAR cutoff of 5 cm/ka. However, it is not necessary to run a transient bioturbation model to demonstrate the limitations of a SAR that is less than 5 cm/ka. We can carry out simple calculation following established understanding of the influence of bioturbation upon the age dispersal of single foraminifera (Berger and Heath, 1968; Berger and Johnson, 1978; Berger and Killingley, 1982), the same understanding that is included in transient bioturbation models themselves (e.g. Trauth, 2013; Dolman and Laepple, 2018; Lougheed et al., 2018). In such a case, assuming a bioturbation depth of 10 cm (Peng et al., 1997; Trauth et al., 1997; Boudreau, 1998), we can calculate that the 1 sigma age value of foraminifera contained within a single cm of a 5 cm/ka core is 10 / (5/1000) = 2000 years (from which follows that 2 sigma = 4000 years). – We will include a version of this paragraph in a revised manuscript

> Furthermore, rapid accumulation rates should be possible around islands. The coarse map overlaid here fails to account for these. I would suggest adding to the text that in strategic locations (in the blue areas), sedimentation rates may still be high enough.

We agree seamounts and islands are places where the depth is shallower than our prescribed carbonate compensation cut-off depth and could also potentially have higher SAR, we will add a clarifying statement to the figures and the main-text.

Improper referencing

> This is not the first study to use pseudo-proxy to examine whether IFA can be used for ENSO reconstruction. Thirumalai et al. present a model that can be more easily applied to a real application. First, reference this study (and others) at the beginning of the manuscript and second, why not extend their "picking" model to also evaluate the contributions of sample size?

We reiterate that our forward model seeks to model the d18O of the full foraminifera populations in the water. We are not modelling sediment core IFA.

Minor comments

> Abstract: Should state that this is an IFA technique.

See answer to your previous point.

> Page 1, Line 23: specify that the interaction on interannual timescale is known as ENSO. On decadal, it's known as the PDO.

We will rephrase to:
"Predictions of short-term, abrupt changes in regional climate are imperative for improving the spatiotemporal precision and accuracy when forecasting future climate. Coupled ocean-atmosphere interactions (wind circulation and sea surface temperature) in the tropical Pacific, collectively known as the El Niño-Southern Oscillation (ENSO) **on interannual timescales and the Pacific Decadal Oscillation on decadal timescales**, represent global climate's largest source (Wang et al., 2017) of inter-annual climate variability (Figure 1)."

> Page 1, Line 27: SO is part of ENSO. Should rephrase as we have long instru- mental records of the atmospheric variability but not the ocean.

We will rephrase to:
"The instrumental record of the past century provides important information (*i.e.* the Southern Oscillation Index; SOI), however, detailed **oceanographic** observations **of the components** of ENSO **(both the El Nino and Southern Oscillation)**, such as the Tropical Oceans Global Atmosphere (TOGA; 1985-1994) experiment only provide information from the latter half of the twentieth century (Wang et al., 2017)."

> Page 3, line 3: Stott et al is not the only reconstruction in the Western Pacific, either use e.g., or as done previously cite multiple sources.

We will include multiple sources and an e.g., so it will read as follows: "and (3) those associated with the trace metal geochemistry (**e.g., Ford et al., 2015; Sadekov et al., 2013;** Stott et al., 2002, 2004; **White et al., 2018**), more specifically the natural logarithm of the relative concentration of Mg and Ca (ln(Mg/Ca), of the shell, based upon the temperature dependent (Elderfield and Ganssen, 2000; Nürnberg et al., 1996)incorporation and substitution of a Mg cation into the calcite lattice (Branson et al., 2013, 2016).".

However, we would like to clarify that the rationale for the citation here is not referring exclusively to the Western Pacific pool, but rather citing papers (including earlier papers) around the topic of 3 proxy types as noted on Pg. 2 ln23:

"Proxies of past ENSO and Pacific SST (Ford et al., 2015; Koutavas et al., 2006; Koutavas and Joanides, 2012; Koutavas and Lynch-Stieglitz, 2003; Leduc et al., 2009; Sadekov et al., 2013; White et al., 2018) are based upon the biomineralisation of the calcite, or a polymorph such as verite (Jacob et al., 2017), and shells of foraminifera (Emiliani, 1955; Evans et al., 2018; Zeebe and Wolf-Gladrow, 2001). There are three major types of foraminifera-based palaeoceanographic proxies: (1)…. (2)… and (3)…"

Page 3, Line 30: Mg/Ca is not a simple function of temperature. There is a growing body of evidence that suggests that Mg/Ca is also sensitive to salinity and pH. In addition, the calcite saturation of the bottom waters on post-depositional preservation of the signal.

We agree that is why on Pg. 11 Ln 16 – 23 we state : "several other parameters can alter this technique, this includes abiotic effects such as salinity (Allen et al., 2016; Gray et al., 2018; Groeneveld et al., 2008; Kısakürek et al., 2008) or carbonate ion concentration (Allen et al., 2016; Evans et al., 2018; Zeebe and Sanyal, 2002); biotic effects such as diurnal calcification (Eggins et al., 2003; Hori et al., 2018; Sadekov et al., 2008, 2009; Vetter et al., 2013); or additional factors such as sediment (Fallet et al., 2009; Feldmeijer et al., 2013) or specimen (Barker et al., 2003; Greaves et al., 2005) 'cleaning' techniques. Given the role of Mg in inhibiting calcium carbonate formation, the manipulation of seawater similar to the modification of the cell's *p*H (de Nooijer et al., 2008, 2009) may aid calcification and explain the formation of low-Mg by certain foraminifera (Zeebe and Sanyal, 2002)"

As such we are producing a pseudo proxy of what the proxy aims to test, i.e. temperature, we will add some text with regards to these later lines, where the reviewer has pointed it out, as follows:

"Here, we use the recently developed *Foraminifera as Modelled Entities* (FAME) model (Roche et al., 2018) to take into account potential modulation of $\delta_{18}O_c$ and the Temperature recorded in the calcite, herein T_c, by foraminifera growth. Tc can thus be considered as an estimate of the proxy Mg/Ca (**albeit one uninfluenced by secondary factors**)."

Page 5, Line 3: Why not used species- specific equations?

A first reason is that the use of equilibrium as opposed to species-specific equations places all foraminiferal species against an equal benchmark. More importantly, most species-specific equations have been produced in culture (though some others in tows and pump samples) so that they contain variability in both growth and environmental (the same data that produced the growth functions in FORAMCLIM/FAME) conditions.

As an equilibrium oxygen isotope matrix (time-depth) is used to produce the FAME weighted distributions, said equilibrium can be switched out for 'equilibrium species values', however, such a terminology does not clearly outline the influence of time (growth season) and depth on the different species' oxygen isotopes.

Page 5, Line 12: Not sure what is meant by "Which can compute eight foraminiferal species". Do you mean growth?

The sentence will be altered as follows:

"FAME is based upon FORAMCLIM which can compute **the growth of** eight foraminiferal species (Kageyama et al., 2013; Lombard et al., 2009, 2011; Roche et al., 2017), however comparison with a core top database has been limited to five foraminiferal species (Roche et al., 2017)."

Page 6, line 11: There is an abundant body of literature dealing with the definition of an ENSO event. Why not start there?

We are not attempting to challenge or redefine ENSO or discount ONI or the BJ-Index (etc.) derived event chronologies, instead our simplification is based upon two things: (1) the simplicity of the input variables and (2) whether foraminifera (~4 weeks life) would 'sense' the event (as stated in our previous comment to reviewer 1). To clarify we will alter the paragraph as follows:

Pg. 6 Ln 10 – 15: "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 $\chi \geq +0.5°C$ **(where χ is the value of ONI)** 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). **This simplification reflects the lifecycle of planktonic foraminifera (~4 weeks) as the population at time step t knows not what happened at t-1 or will happen at t+1. As we are producing the mean population growth weighted $\delta^{18}O$ values, 'almost' El Niño or 'almost' La Niña would be indistinguishable from the build-up and subsequent climb-down of actual El Niño and La Niña events. Therefore, these 'almost' El Niño or 'almost' La Niña are placed within their respective climatological pools as El Niño or La Niña.**"

Page 9, line 20-25: Most of these studies are based on pooled samples and were referencing to an ENSO-like signal rather than the interannual mode of variability that IFA is targeting.

We state that these datasets are from pooled and individual (Pg. 9 ln 18-20 “Several authors have focussed on individual foraminifera analysis (IFA) or pooled foraminiferal analysis in the Pacific region, either for trace metal or stable isotope geochemistry.”) and later we point out that these are the authors inference (Pg. 9 ln 22-23 “The resultant data of such studies have been used to infer a relatively”). We agree that depending upon the method authors can be referencing ‘ENSO-like’ and/or ‘interannual mode of variability’, we make reference to this point at Pg 10 ln 9 -26 (e.g., “Our own analysis using our FAME $\delta_{18}O_c$ and T_c output mimics foraminiferal sedimentary archives, pooling several decades worth of data in which the resolution is coarse enough to obscure and prevent individual El Niño events being visible but allowing for some kind of long-term mean state of ENSO activity to be reconstructed”).

As we already take into consideration the reviewers point, we can assume that this comment merely reflects the opinion we are performing IFA.