We would like to thank the anonymous reviewer (Reviewer 3) for their constructive suggestions and comments, which have helped to improve our manuscript entitled “SCUBIDO: a Bayesian modelling approach to reconstruct palaeoclimate from multivariate lake sediment data.” Below, the reviewer’s comments are shown in red, and our responses in black.

The authors frequently mention quantitative proxy values (say line 78) but are the proxy records they collect not quantitative data, in terms of intensities, as is much of the proxy data collected in terms of tree rings, battles, diatoms, pollen etc. It would be good to clarify this as qualitative proxy records are mentioned throughout but it is not clear to me what this refers to.

We thank the reviewer for this comment, and we agree that this might be confusing. We meant that the climate information derived from proxy data is qualitative but agree with the reviewer that proxy data are quantitative. We will be more specific in the new revised version of the manuscript and will change “qualitative proxy value” (e.g. line 78 in the previous version of the manuscript) for “qualitative climate information derived from proxy values”.

(Line 225 and Figure 1). From a modelling perspective there is no clear rationale made for why a quadratic relationship is appropriate, perhaps some form of smoother would work equally well without the tail assumptions of the polynomial model. In Figure 1 visualisations of the fits for each chemical is made however there is no clear relationship and the figures reflect a lot of noise., This may be due to only one climate proxy being presented (temperature?) whereas it is possibly the case that the proxy could depend on several - the authors should comment in this regard.

The reviewer makes some great points here which mirror the responses to the other reviewers. Please see the response to Reviewer 2 about the quadratic relationship. We have tried a P-spline model but found that this significantly underestimated the variability and is very computationally expensive.

We have responded to Reviewer 1 about the lack of a clear relationship between the individual elements and temperature in Figure 1, but as a summary this figure should not view each of the element relationships as independent as it is the joint relationship between all these elements which provides us with the relationship between climate and XRF data used to calibrate.

We completely agree that some of the noise is going to be coming from other meteorological processes and variables. However, this is the reason why we wanted to use a probabilistic approach as there are likely many processes which are involved unrelated to temperature (or another meteorological variable we are reconstructing), and thus we want to account for these within the uncertainties. If we were to have other bits of information that we can include, such as precipitation etc, then it is likely that the uncertainties would be reduced. However, it is challenging to know what those other drivers are, and it is also challenging to get data for this which is of good enough quality to put into the model and have enough computational power to fit within a reasonable timeframe. We thank the reviewer for this for this suggestion though as these are some very interesting points, but at present this is something that we are not able to here. We will add this as a potential avenue for future model updates.

The authors mention that uncertainty quantification is a strong basis of their approach and note (Line 417) "Nevertheless, gaining an 80% coverage percentage is acceptable for this modelling approach ". It appears that the constructed 95% HPD regions only contain 80% of observations. Why is the 80% acceptable, or a stronger argument needs to be made in this regard. Perhaps the reduction in uncertainty coverage is due to the log-ratio transformations of the XRF data and modelling inter-element relationships with a multivariate random effect which is acting as a poor equivalent to accounting for the compositional nature of the data? While the authors claim reasonable performance from an uncertainty quantification perspective, insufficient discussion is made of mean/median (unclear which) predictions in Figure 2. I note that authors later calculate the correlation between Diss Mere and the LMR but provide no similar calculation here? Figure 2 seems to suggest that the R^2 in this case would be very poor, is this why it is not presented?

We thank the reviewer for this comment. We have since re-run the model for Diss Mere and started the calibration period at 1700 CE rather than 1659 CE due to higher uncertainties in the instrumental data and missing years in the XRF data. In addition, after carefully reviewing the validation code we identified a small bug which was rounding up the age variable to the nearest integer. This shifted the predicted temperature by one year and therefore was not directly comparing to the same year in the validation dataset. However, since accounting for this misalignment, we now have a much stronger relationship between true and predicted temperature, and a very good coverage percentage (97.4% of the reconstructed values fit within the 95% confidence intervals). We thank the reviewer for the suggestion about adding in the r^2 values, we think this is a great idea and we will add these into the figure. Based on this comment we will also add in a few sentences talking about the median value and how they do not perfectly align and potential reasons as to why.

We have included the new updated true vs reconstructed figures to the response to Reviewer 1.

A weakness (in my opinion) of the results presented in the two case studies later on stems from an insufficient evaluation of predictions of the mean temperature for the calibration dataset at Diss Mere - predictive performance in terms of the uncertainty may be 80% coverage, but it is not clear that the center of the prediction intervals are accurate - this perhaps explains the commentary on the performance of Diss Mere later on. Was an analysis of the calibration approach carried out at the Finnish site? If so, does it explain why the predictive performance is potentially better there than Diss Mere? Alternatively, do the weaknesses in predictive performance in Figure 2 also manifest at the Finnish site? Similarly annual mean temperature is used - is precipitation potentially useful to incorporate here or is predictive performance poor when it is incorporated as observed at Diss Mere? Additional evaluations in this regard could be included in the Supplementary materials.

We thank the reviewer for this, and we hope that the response to your previous comment has addressed the concerns about the predictive performance. We think it is a very good idea to add in the validation information for Nautajärvi, and we will include this as part of the supplementary information as a response to this comment. Please see Figure A1-4 in the response to Reviewer 1 for the Nautajärvi figure.

In terms of model assessment, the sensitivity of results to specified priors is not provided - the priors as presented are very vague but could some information be incorporated to make these more informative? Since the XRF is rescaled using a centred log ratio - is it plausible that intercept values of +-200 are possible, which is what is suggested by the vague prior. Similar arguments apply to the priors for the other components - are some of the values suggested by the prior impossible?

We are grateful to the reviewer for pointing this out. However, we think there might be a misunderstanding of our probability distributions. We are using this standard nomenclature of a normal distribution being represented by its mean and variance. Thus, when we write N (0, 100) we are referring to a normal distribution with mean zero and standard deviation 10 (variance 100). Thus, the intercept has an a priori 95% range of -20 to 20 which seems plausible for CLR transformed data. This is still a vague prior, but we would argue that it is weakly informative for the scales of the data that we are modelling.

Please refer to the response to Reviewer 2 about or choice behind the different types of priors used within this model.

The manuscript also requires substantial editing as there are a number of grammatical errors, typos and excessive use of language which makes the manuscript difficult to follow at times. For eg, SCUBIDO is spelt incorrectly twice and some of the text used either side of equations causes confusion. I have noted several of these below.

We thank the reviewer for identifying these and we will change all of these in the new version of the manuscript.

Line 309 - "found a qualitative link" - what is a qualitative link?

We suggested that there was a qualitative link as there was a good visual relationship between Holocene temperature evolution and the Ca record. We will explain this in more detail.

Line 334 - "and thus the model did not find a good enough relationship. Annual mean temperature on the other hand worked well, which support the temperature signal recorded in the qualitative XRF-CS data during the Holocene " - what is meant by a "good enough relationship? Why was the temperature signal and XRF-CS relationship deemed good enough?

We thank the reviewer for identifying that more detail is needed here. We will explain in more detail why precipitation was not used as the reconstruction was flat, there was no predictive power between the elements and precipitation and the validation showed no relationship between true and reconstructed precipitation.

Line 337: "Another point to highlight at this stage is that we run the Bayesian model using a multivariate dataset made of the elements measured by the XRF scanner, which differentiates SCUBIDO from other recent reconstructions based on varved sediments " - How does it differ?

We were referring to other approaches using only single elements to infer climate, or a pair of elements in the form of a ratio. However, we will now remove this sentence following the suggestion from Reviewer 1 as we mention this later in the manuscript.

Line 339: “We therefore rely on the Hadley Central England Temperature (HadCET, Met Office) data” - is this proximate to the site?” As such, does it capture temperature change at the site reasonably?

Unfortunately, we could not find another station which is closer to the site and is also long enough to calibrate the data as the top of the XRF record is at 1932. We will add some additional information in the text about this in response to this comment.

Line 350: “$XRFm$ was resampled to annual means “ - How was it sampled or adapted? Was the XRF data not at annual level in any case?

We used linear interpolation to downscale our resolution form many data points per year to one. We will clarify this in the text based on this comment and the comment from Reviewer 1.