Response to anonymous referee #1
(black : referee's comments ; blue : our response)

The work provides a well-constructed template for local paleoceanographic pseudoproxy sensitivity experiments, and so provides a starting point for more quantitative analysis of uncertainty in paleoceanographic studies. Its conceptual simplicity should make it easily accessible to a wide audience. However, the applicability of the algorithm in estimating uncertainties in real-world reconstructions from oceanographic proxies is somewhat limited. In particular, the authors overstate the importance of their algorithm's ability to "[identify and estimate] systematic bias that would not otherwise be detected." MoCo is only able to identify systematic biases in a pseudo-proxy context, where the specific form of climatic nonstationarity over the reconstruction interval is already known.

We thank the referee for his time and effort in reviewing this paper. We stated in the paper and underlined more explicitly in the revised version that the applicability of the algorithm is limited to the reconstruction of paleoclimate statistics (annual mean value, mean seasonal amplitude, and interannual variability) from mollusk, corals and sclerosponge high resolution geochemistry. To fully understand the contribution of this work, it is necessary to be familiar with the constraints that characterize this type of archives. Modern corals live underwater in restricted areas, are often protected, and the analytical cost of a 20 year long $\delta^{18}O$ time series is about 3600 USD. It is therefore impossible to produce a modern dataset of hundreds of series over a wide geographical range for a robust statistical assessment of "*real-world reconstructions*" with sophisticated statistical analysis similar to those used for example in dendroclimatology.
Our method is indeed limited by climate nonstationarity and does rely on the assumption that the proxy-climate relationship is conserved under different environmental conditions, but is it not the universal condition in paleoclimatology?
We do not agree that MoCo is of limited interest in real-world context. The usefulness of simulations to the real-world are measured by the degree of realism and by the efficiency compared to alternative methods. Models are used all the time in real-world contexts for all types of applications. Here, the realism of error estimated by MoCo is necessarily higher than the error estimated by the mere proxy calibration since it involves several additional sources of uncertainties and implies a more precise definition of the reconstruction uncertainty through three categories of errors: the systematic error, the standard error and the potential systematic error.
Therefore, with all its limitations that we discuss with more detail, the method proposed here is a significant improvement for quantitative paleoceanography from corals and mollusks since, actually, uncertainties for paleoclimate statistics are generally not quantified. In the revised manuscript, we mentioned these points in the introduction, in the conclusion, and in the section 6.3. that was renamed "quantifying errors: contributions and limitations"
The issue of climate non-statonarity was explored in a new experiment.

The work also currently does not analyze the effects of proxy uncertainty on any measure of the reconstruction's ability to capture interannual variability. As interannual variability estimates are often the desired product of paleoclimatic reconstructions, the MoCo algorithm seems somewhat incomplete without this analysis.

Interannual variability is indeed one of the most valuable information provided by corals and mollusks, but there is a big misunderstanding here since this is precisely the purpose of this work. Throughout the manuscript we studied the reconstruction's ability to capture the

variance of the annual mean temperature and the variance of the annual cycle amplitude, which are both measures of the interannual variability.
In the revised manuscript, we changed the names of the reconstructed statistics for T_m, V_T, Δ, and V_Δ, and gave explicit mathematical definitions in section 2 to avoid this confusion.
We suspect that the referee referred in this comment to calendar reconstructions (where a calendar year is affected to every year of the record, as in dendroclimatology), which is not the context of this work. We made this point clear in the introduction and discussed shortly in section 6.4 the work that would be needed to extend this method to this type of reconstructions.

1 General Comments

1. The authors mention three types of error in the paper (systematic error, potential systematic error, and standard errors), which seem critical to many of their discussions. However, nowhere in the paper are these three types of error clearly and distinctly defined. Clear definitions for these central concepts are critical to the paper, especially since at least one of these labels (standard error) has a different meaning than used here in other contexts (eg. the concept of a “standard error" in statistics).

In statistics, the standard error of the mean is the standard deviation of the error in the sample mean relative to the true mean, which is exactly how the standard error is defined in our article. We merged the previous section 3.4 with section 2.1 and added the mathematical definitions of the 3 types of errors in section 2.1 to clarify this.

2. The above comment references the most important example of generally imprecise writing throughout the paper. Explanations of many of the methods need much more careful detailing, and are referenced in the following section listing specific comments.

The whole manuscript was deeply revised for more precision and clarity, especially in the description of the methods. We provided mathematical definitions of all our variables. We focused especially on the aspects that were not clear for the referees.

3. Although the two words are sometimes used interchangeably, the authors may want to reconsider their heavy use of the word “error" throughout the paper and in its title, and replace it with the word “uncertainty." The former word implies a mistake in the analysis, while the latter invokes the effects of an inescapably stochastic and/or nonlinear nature of the proxy formation process on the resulting record of climate. The latter will also tie the work in to a greater body of literature on uncertainty quantification in climate studies.

We agree with the referee’s definition of the uncertainty and we add that its meaning is quite general. However, we do not agree with his definition of the error. The term error does not imply a mistake in the analysis. In statistics, the error is defined as the difference between the value obtained by the estimator of a statistics and the true value of this statistics. Since this definition is more specific and correspond exactly to what is calculated by the algorithm, we kept using this term.

4. The authors statement of the work’s contribution in the abstract and conclusion is falsely inflated. Specifically, the authors need to make it clear in the abstract and conclusion that MoCo is a tool designed specifically for synthetic, pseudo-proxy studies, and that the results mentioned apply only to these contexts.

We made clear in the abstract and conclusions that the errors calculated by MoCo are the output of a simulation that implies approximations and assumptions. However, we insist that this tool, with all its caveats, does represent a significant improvement for quantified paleoclimatology using mollusks and corals. Today, calibration studies of coral and mollusk geochemistry against an environmental variable yield an estimate of the uncertainty for data points. When it comes to climate statistics (annual mean, seasonality, interannual variability), values are generally provided without error bars, and systematic biases cannot be detected because modern datasets are not large enough to get reliable estimates of these uncertainties. The sources of standard error have also never been explored. MoCo does provide a statistical estimate of the standard and systematic errors for 4 statistics and represents a methodological framework for more complete uncertainty studies. It is based on a realistic simulation of the proxy climate reconstruction process that involves all the main error sources in this context. It is thus indeed designed to be used with real-world reconstructions and not only for pseudo-proxy experiments.
We revised the introduction, the discussion and the conclusion for a better presentation of the actual context of quantitative paleoclimatology using mollusks and corals, so that the contribution and the novelty of this method appears more clearly.

5. There should also be a section in the body of the paper that clearly and thoroughly addresses the limitations of the algorithm in application to real-world problems.

Shortcomings of the method were already exposed in the section 6.3. For more clarity, its title was changed to “quantifying errors: contributions and limitations”. We also insisted on the strong influence of climate non stationarity, and explored further and quantitatively its impact in a new experiment. Experiment 6 (figure 7) showed that in a location, the standard error is linearly correlated to the interannual variability. Based on this result, we propose to estimate and use this relationship to minimize the bias due to climate nonstationnarity for standard errors.

6. Because an estimate of interannual variability is often the desired product of many paleoclimatic reconstructions, the analysis should look at some measures of the reconstruction’s ability to capture interannual variability in addition to the measures already examined (Tm, var(Tm), _T, and var(_T)).

As stated earlier, var(Tm) and var(ΔT) (V_T and V_Δ in the revised version) are measures of the interannual climatic variability. Hopefully the explicit mathematical definition will clarify this.

Suggested statistics are the correlation of the estimate with the target, the significance of that correlation, and the coefficient of efficiency and/or reduction of error statistic commonly used in dendrochronology (see Cook and Kairiukstis (1992), eg.). It would also be useful to look at the effects of multiple sources of uncertainty on some measure of variance loss or amplitude attenuation. Analysis of these statistics in addition to the ones already examined will also help place the work in the context of other pseudo-proxy studies in the paleoclimate literature (eg. Smerdon et al (2010), Mann and Rutherford (2002), von Storch et al (2009)).

This would be indeed a very interesting study, but it is beyond the context of this work. This would apply to reconstructions of calendar series. We added this statement in introduction: “*We only consider here the case where the climate statistics of long time periods are estimated by a sample of short windows (similarly to Tudhope et al., 2001). We do not*

address the case of calendar reconstructions (similarly to Cole et al., 1993 or Cobb et al., 2003).”. We suggest in section 6.4 that the method should be adapted to calendar reconstructions. In this case the record has the same length as the target time series. The Monte Carlo analysis would be performed by repeated reconstructions of varying synthetic time series. Every reconstruction would yield a correlation coefficient, a level of significance or other statistics. The standard deviation of these coefficients for the whole population of reconstructions would yield an estimate of the skill of this type of reconstruction. Estimating the influence of stochasticity, proxy limitations and age uncertainty on this type of reconstruction would require a specific study in another article.

2 Specific Comments

2.0.3 Abstract

This is well-written and motivated, with the important exception of having swept the pseudo-proxy context under the rug (as noted in the general comments).

As discussed earlier, we do not think that the pseudo-proxy context prevents our method from being used for real reconstructions so that nothing needs to be “*swept under the rug*”. It is clear in the text that errors are calculated from numerical simulations. We mentioned that paleoclimate statistics were so far reconstructed from mollusks and corals without any error bar whatsoever, to make clear that error estimates from MoCo, though obviously imperfect, are a significant improvement.

2.0.4 1. Introduction

• pp. 2479, lines 16-19: Although I agree broadly with the authors’ statement about the assessment of uncertainty in most paleo-oceanographic reconstructions, the authors should be aware of work by Evans et al (1998) on the sensitivity of reconstructions to network choice, and work by Brown et al (2008) and Thompson et al (2011) on the forward-modeling of isotopic signals found in corals.

We should have indeed mentioned these studies related to uncertainties in coral-based reconstructions. We added a paragraph in the introduction to summarize their contribution in the context of this work.

• pp. 2481, line 2-5: I disagree that the technique presented here is conceptually similar to Haslett et al (2006). The latter study is a climate reconstruction from observed data, and presents modeling that describes the relationship between the proxy signal and the climate, as well as the relationship of the proxy and climate fields to themselves across space and multiple sources of uncertainty inherent in each one. In such Bayesian hierarchical modeling, the data are used to constrain the uncertainties arising at every other level of the model. The work in the present paper is a pseudo-proxy experiment, rather than a reconstruction, and the modeling aims only to represent the effect of various sources of uncertainty in the proxy data.

We removed the reference to Haslett et al. (2006), although this comment sounds strange to me since we did not write that our technique is conceptually similar to Haslett et al. (2006) but that it is conceptually simple compared to Haslett et al. (2006), whose aim is also to constrain reconstruction uncertainties.

• In the last paragraph of introduction, the known target and pseudo-proxy context necessary for the operation of the algorithm should be made clear.

We added this information in the last sentence of the introduction. The context of the simulated reconstructions was also made more explicit in paragraph 4.

2.0.5 2. MoCo Algorithm
2.0.6 2.1
• pp 2482, first paragraph. Looking at effects on metrics of reconstruction of interannual variability would also be useful (for example, correlation and p-value of the reconstruction with target in low and high frequency bands; coefficient of efficiency or reduction of error statistic; reconstruction bias).
We answered earlier to similar comments

• pp. 2482, line 22: The authors describe the proxy formation processes as “stochastic," which may not be strictly accurate (even if inherent nonlinearities in the processes make stochastic models for them appropriate).

This is a simplified interpretation of our sentence: “*The formation of the proxy record involves a complex chain of physical and biological processes (for instance mechanisms of Strontium incorporation into coral aragonite) that introduce non-climate-related stochasticity*”. Nonlinearities are likely involved, but there are also a large range of factors related to ecology, biology, biochemistry, microstructure, microenvironments, that influence the proxy value in a way that is stochastic at our space and time scales. For example, an idealized proxy model would consider that a contemporaneous aragonite layer in a mollusk has a homogeneous Sr/Ca ratio. In the reality, there is a microscale heterogeneity due to growth rate variations, or crystal-scale localized diagenesis. The spatial repartition of these anomalies is stochastic and the location of the sampling within a layer is also random. It is beyond our scope to review all the sources of uncertainties and their mechanisms. We added two references related to strontium incorporation. We only need here to have a quantified estimate of their combined effects.

• The rest of the section needs to be re-written to clearly define the author’s working definitions of “standard error," “systematic error," and “potential systematic error." In Figure 1, definitions for two of the three types of errors are given in equation format; using these equations and elaborating on them in the body of the paper would be useful in providing clear definitions. Note that giving examples of the different error types does not constitute a precise definition.

Section 2.1 was completely revised to clarify this. We added mathematical definitions of these three types of errors. We added also precisions in the result and the discussion sections to clarify how these categories of error are estimated, how different they are, and how they should be treated.

2.0.7 2.2 and 2.3
• By the end of section 2, the reader should have a clear idea of the general MoCo “workflow." However, I found myself later coming back to this section and comparing it with the details presented in section 4, to try to understand the workflow by example. A key point of confusion for me is whether MoCo served only to perturb the climatic target, and requires being coupled to a forward model or proxy formation in order to be used for pseudo-proxy

experiments (which seems to be how the case study works). If this is the case, this should be clearly stated in section 2, and the coupling to a separate forward model should be included in the diagram in Figure 1. On the other hand I can also envision the ensemble of perturbed target climates being interpreted as the “pseudo proxy" signal, without the use of a separate forward model, as is currently diagrammed in Figure 1. Perhaps the authors intend for MoCo to be used in either way; whatever the case may be, it should be clearly described here. If I understand correctly, the intended output of MoCo is a "pseudo proxy", and so should be labeled with a different letter than climate (perhaps use Pi) in Figure 1. These pseudo proxy series must then undergo a reconstruction method before the Pi are translated back into estimates ^ Ci of the climate to be compared with the target C0 (as described later in section 3.4). This reconstruction step should also be described in the section and diagrammed in the figure.

Things are actually as the referee “*envisioned*” them from Figure 1. We do not use a forward model *stricto sensu*. The outputs of MoCo are estimates of (1) the standard error, (2) the systematic error and (3) the potential systematic error due to the proxy calibration. The calculation of the first 2 errors involves repeated production of surrogate proxies, and the 3rd one is simply calculated from the calibration dataset. Surrogate proxies are simply produced by perturbing SST series.
The title of section 2 was simplified to: “The MoCo program”. We started this section by a simple description of the analysis performed by MoCo, the input and output of the program, trying to clarify the questions pointed by the referee.
There was probably a confusion between forward models and what we called the proxy model, which is simply an empirical linear equation between the proxy and the climate variable. We clarified the definition of the proxy model in a new section 3.1.
Section 3.4. was probably confusing the referee. It was deleted and merged into section 2.1. with the description of the other errors.
The step from Pi to Ci described by the referee does not exist so there is no need to add it to figure 1.
We also made clear in section 4 that in the sensitivity experiments, MoCo was used repeatedly with varying values of a parameter to study the sensitivity of the errors to this parameter.

2.0.8 3. Inputs to the algorithm
2.0.9 3.1
• line 13-15: The requirement on the length of the target climate series should be both quantified (longer than the proxy record by what factor? Does that factor depend on any other characteristics of the typical proxy series?) and justified statistically.

We corrected section 3.1 (now section 3.2) to make this point clear, and added: “*The number of years N_0 of the time series should be larger than the total number of years N_S of the proxy records sample to allow adequate random sampling. A N_S/N_0 ratio lower than 0.2 would keep the average overlap rate in the samples under 10%. The length of the target time series should also be chosen according to the period the sample is expected to be representative of*.”

• line 17: Such a long time series could also be statistically generated.

Right. We added this comment in the text.

2.0.10 3.3.1

• Implicit in the statement that several specimens should be analyzed to average out the effects of spatial heterogeneity is the assumption of some larger-than- local-scale that the target climate represents. This may not always be the goal of a reconstruction (perhaps one would like to reconstruct climate local to a given proxy), in which case this noise can be set to zero.

This is true, but not implicit: we mentioned this in the second paragraph of section 2.1. And even for local reconstructions this noise may be useful to represent microenvironment heterogeneity. We added: "*σ_S should be estimated according to the studied geographic scale*."

2.0.11 3.3.2
• It is not precisely clear what is meant by "proxy analytic error."

It is stated in the introduction that the study is focused on geochemical proxies in accretionary calcium carbonate skeletons. The geochemical proxy ($\delta^{18}O_{carbonate}$ in the case study) is generally measured by mass spectrometry with an associated analytical standard error estimated from repeated measurements on known standard material. We corrected this part to be more specific: "*geochemical analysis standard error*".

• Perhaps explain why these uncertainties add this way to less statistically-oriented readers. For clarity in the equation for σm, either the letters m; a;w; and c should be put in subscript.

We slightly corrected the sentence to make it clearer: "*These three sources of uncertainty add in quadration because they are independent*".
All the variable indices were put in subscript throughout the manuscript.

• The choice of temporally uncorrelated errors should be justified for each of these three errors, or perhaps temporally correlated errors should be considered by the authors.

I do not think it is really necessary to justify that mass spectrometry uncertainty, weather during the shell life, and shell post-mortem diagenesis are not correlated.

2.0.12 3.3.3
• It should be stated clearly that (Tls; Tli) represent thresholds below and above which precipitation of skeletal material stops. References to the literature to support the existence of these kinds of thresholds should also be provided.

We added this statement in section 3.3.3. However, since there are obviously upper and lower lethal temperature limits for every living species, it did not seem necessary to add references here.

• line 16-18: If these breaks correspond to the input variable "gap" in Table 3, this should be clearly stated.
corrected.

2.0.13 4: Sensitivity Experiments
• Stochastic noise could also be added to the parameters in model (2) to account for uncertainty in the proxy formation process. Why do the authors choose not to add noise here? (In the language of Bayesian hierarchical modeling, the authors put all the uncertainty at the

emphprocess level, and none at the data level- this is another reason this study is significantly different than that of Haslett et al (2006)).

This comment is hard to understand and seems contradictory: the referee asks why we did not add noise in the proxy formation process and then says we put all the noise at the process level. I feel that the referee does not clearly understand what we do. Hopefully our revisions will clarify our experiments. As for the stochastic noise, σ_m involves the analytical uncertainty σ_A (which represents uncertainty at the data level) AND σ_C, a noise related to the carbonate heterogeneity (which is an uncertainty at the process level). σ_W and σ_S represent uncertainties related to spatial and temporal variability that can be assimilated to the process level.
As for the uncertainty related to the proxy calibration, it is explained that this is measured separately by the "potential systematic error".

• Define V-PDB and V-SMOW before using these acronyms.

V-PDB (Vienna Pee Dee Belemnite) and V-SMOW Vienna Standard Mean ocean Water) are the names of international standards for isotopic analysis. Defining these acronyms would make the text heavier without bringing useful information. We added the reference Coplen (1996).

• A citation is needed for IMARPE instrumental record mentioned in line 3.
Unfortunately, this record is not associated to any publication. We defined the IMARPE acronym, cited the web page where it is available, and we added the IMARPE web page.

• To clarify the organization of the following experiments, the authors should consider expressing them in terms of statistical factors and treatments. This language could also be incorporated into Table 3.
It was not clear to us what corrections were expected here by the referee. The statistical treatments are the same in all the experiments. The aim of every experiment is to explore the sensitivity to one or several parameters, which is expressed in the corresponding section title.

• The authors may want to consider combining and reorganizing this and the next part of the paper, so that the explanation of each experiment is immediately followed by its results. This would make it much easier to follow.

We considered this possibility when writing the manuscript. However we finally chose the actual organization because the influence of some parameters is explored in different ways in several experiments. For example the influence of σm is studied in experiment 1 in a on/off mode and compared to other influences whereas it is studied over a larger continuous range of values in experiment 3. The effect of the "target" time series is also studied under different conditions in experiments 1, 3, 4, and 5 (table 3). It seemed therefore better to have all the experiments described before presenting the results focusing on one parameter and putting together the information from different experiments. This is why the results sections include in parenthesis the numbers of the corresponding experiments.

2.0.14 4.1
• I believe the authors mean to say experiment 1 tests the effect the *number of replicates* or *proxy sample size* had on the standard and systematic errors, rather than the "effect of

sampling" or “effect of random sampling". I had to look at the table to clarify what they meant here. This language should be changed for clarity both here and in the results section.

The referee seems to misunderstand the way MoCo works. MoCo draws a random sample of N short time windows and estimates the climate statistics from it and repeat the operation 5000 times. At every iteration, different years are sampled so that the estimate is different and thus the error. This is clearly the “effect of random sampling”, which includes the sample size. We cannot speak of replicates here because the years drawn in the sampling are different. Hopefully this will be clarified by our revisions of section 2 and section 3.

2.0.15 4.5
• How were the two temperature thresholds sampled from the intervals described here? From a uniform distribution on the interval? From a truncated normal distribution? There are many many ways to imagining sampling from this interval, and the authors should be specific about precisely how they did it.

It is difficult to understand the referee’s point, and especially what he means by “sampling a temperature threshold”. In this experiment, a threshold is imposed, which means that all temperatures beyond this value are not recorded (excluded from the sample). Systematic and standard error values are obtained by MoCo under these conditions. Then, the value of the upper threshold decreases from Tmax to Tmax-10°C and the corresponding standard error values are plotted against it. We clarified the description of this experiment in section 4.5.

2.0.16 5: Results
2.0.17 5.1
• line 5-7: It is true that systematic reconstruction error increases with the difference between climatic conditions in the reconstructed and calibration intervals. However, it is important to note that this is just one specific manifestation of the general problem that nonstationarity poses to climate reconstructions. In realworld reconstructions, there is no calibration test or screening method to detect this kind of error without independent a priori knowledge of the climate outside the calibration interval. MoCo can only detect systematic error resulting from climatic nonstationarity in “pseudo-proxy" experiments.

There is again a misunderstanding here. The result presented in section 5.1 is about “potential systematic error” produced by uncertainties in the proxy model (the empirical calibration, not a forward model). It is not produced by surrogate proxies experiments but simply calculated by error propagation equations. These equations were added in section 2.1. So, we are not here in a pseudo-proxy context. Furthermore, the error mentioned in this section is not a systematic error. As the referee says, there is no screening method to detect this kind of error. This is why we gave it a different status: a potential systematic error, that should be represented as an independent error bar. We tried to clarify section 2 and section 5.1 to avoid these confusions as far as possible.

2.0.18 5.2
• Again, the authors use “random sampling" when I believe they mean to make statements about the effect of sample size on the reconstruction results.
The source of error is the randomness of the sampling. “Random sampling” involves more than just the sample size. In experiment 1, yes, it is equivalent, so we corrected the text accordingly. However, in experiment 2, we keep the same sample size (or total number of

years of the sample), but we change the way it is sampled. The effect of random sampling is different for one 200 year long record or 200 one-year long record.

• line 2- 4: The statement comparing errors from short and long records is difficult for the reader to see immediately by comparing Figures 1 and 2. The figures should be made more immediately comparable by plotting N = 200/Ny on the horizontal axes in Figure 2.

We do not think this would help since it would draw the reader's attention on a parameter (N) that is not the aim of the experiment, as the referee thinks.

2.0.19 5.3
• pp. 2491, line 20: Surely the interpretation should be that the effect of spatial variability on standard error decreases with the number of records (Central Limit Theorem!), rather than the effect increasing with record length as stated. The authors see an increase with record length only because they keep N _ Ny fixed, so N and Ny are inversely proportional. If the authors really want to make statements about the effects of record length, they need to keep the number of records fixed as record length increases so as not to confound influences.

We acknowledge the aim of the experiment was not well stated. We changed the title of section 4.2 for "*Experiment 2: should long or short records be preferred?*" and corrected the experiment description to avoid confusion.
We agree with the referee that N is not constant and thus that influences are mixed. However, if N was constant, we would just be increasing the record length and thus the sample size, which has already been explored in experiment 1. The longest record will always give better results than the shorter, there is not much to be learned in this.
On the other hand a legitimate question is: what are the advantages and drawbacks of using short or long records for a reconstruction considering a constant sample size. It is generally assumed that longer records are to be preferred. We show here that in some aspects, 30 one-year long shells give better results than one 30-year long shell. Again, what changes is the way the series is randomly sampled.
We also clarified the result description in section 5.2.

• pp. 2492, line 1- 13: This is very unclear. First of all, if the variability as measured by σm can already be translated to temperature variability (as the authors state that σm of 0:5‰translates to 2 degrees C), then why bother doing the Monte Carlo simulations? Is the 2 °C before or after aggregating other sources of uncertainty? Clearly I am misunderstanding something about this experiment, and so it needs to be described more explicitly.

In my opinion the confusion does not come from the description of the experiment but comes from the beginning with a wrong vision of the referee about the way MoCo works. The referee seems to look for analogies with techniques and issues from the domain of dendroclimatology that do not apply here. Hopefully the revisions of the introduction, the description of MoCo, and the method section will clarify everything.
σm translates into temperature simply using the proxy linear calibration model and is only valid at the monthly scale. The referee forgot that we are not interested in errors of reconstruction of single monthly values but in the reconstructed climate statistics.
The Monte Carlo simulations are needed because other sources of uncertainty are involved at other time scales and that impact the reconstruction statistics differently. This is stated in the introduction.

• The authors may want to consider making the maximum value of σm a fixed factor times the standard deviation of the proxy signal, as in other pseudoproxy experiments (eg. Smerdon et al (2010), Mann et al (2005)) so that the noise is expressed on a scale of signal-to-noise ratio.

This is an excellent suggestion. The experiment is the same but the error values were plotted against alpha*σm/σ(time series) in figure 4 which represents a noise/signal ratio and is thus more meaningful. We corrected the text accordingly, but the interpretation remains unchanged since the curves were notably similar.

2.0.20 5.4
• Discussion of results of experiment 4 are refreshingly clearly interpreted.
Thanks!

• The non-monotonicity of the standard error in Var(Tm) and Var(ΔT) for P.Chicama is quite surprising! This should be explained/interpreted, otherwise this reader is left with suspicion of a bug in the code.

We double-checked the code and did not find any bug. This is also unlikely because this non-monotonicity was only found for the Puerto Chicama time series. We added 7 lines in section 5.4 to explain this result.

• The asymmetric response to the changing the upper versus lower limit likely has to do with how anomalous the points Tmax=Tmin were in the context of the usual climatology. It could be interesting to choose Tmin and Tmax based not on the min and max temperatures in the calibration interval series, but on some number of standard deviations above and below the mean temperature.

The asymmetric response is simply due to the asymmetric variance of SST in the annual cycle. In the Eastern Pacific, El Niño variability is mostly concentrated during the austral summer (Dec-March). The referee's suggestion is interesting but we do not think it would add much in this context.

2.0.21 5.5
• These are good points, but should be discussed in turn as experiments 3, 4 and 5 are discussed, rather than being a separate section.

As we previously discussed, we chose to organize the result section by source of error rather than by experiment, because this seemed more informative for the reader. Each source of error was explored in different ways in several experiments. We brought these observations together in separate subsections. We think that the effect of the target time series is important enough to be summarized in a specific section. It is exactly the effect of climate non-stationnarity the referee pointed out.

2.0.22 6: Discussion
2.0.23 6.1
• line 9-10: This is simply the well-known advantage of Monte Carlo technigues that from an ensemble of realizations, it is easy to look at distributions of any function of the reconstruction (eg. distributions of the amplitude, amplitude variance, etc).

Right. And it is still worth mentioning since this well-known advantage has not been used so far.

2.0.24 6.2
• line 21-22: Citations should be provided for this statement for readers unfamiliar with paleoceanographic studies. It is unclear what is meant by this precision value.

This sentence was not clear. We corrected it. Two references were added.

2.0.25 6.3
• In this section yet again, the authors claim that MoCo is an “improvement" over standard reconstruction techniques because it can detect systematic biases, when in fact the algorithm only does this in a pseudo-proxy context. In lines 8-9, the recommendation to seek high temporal variability in target time series seems impractical at best, and fundamentally flawed at worst. In fact the most realistic results will be yielded from a target series that is as realistic as possible, rather than as variable as possible.

Again, the results obtained from pseudo-proxy experiments can be used in real reconstructions if the simulation is realistic (parameterization, target series…). And again, it can only be an improvement since we start from nothing.
Of course the most realistic target series should be used and not the more variable. We meant the most variable within the range of realistic series.
We rephrased this part more precisely.

2.0.26 Figures
• Figure 6: Top panels are unnecessary and confusing; delete them.

3 Typographical Errors and style
• Even though subscripting isn’t possible within the body of a code, parameters of the MoCo code, and MoCo input variable names, should be referred to with subscripts in the paper, eg. _m and Tli instead of _m and Tli. If the inputs and parameters are described clearly enough in the paper, and if the code is commented well, then the correspondence between variable names in the code and in the paper should be clear.

Corrected.

• Figure 5: blue and red are not distinct enough; both look black when paper is viewed at up to 200% magnification.

We used lighter blue for a better distinction

• pp. 2483, line 19: the word “of" should be the word “one."
corrected