Response to review comment 1

This author response is organised for each comment as follows:
(1) Comment from referee/public
(2) Author's response
(3) Author's change in the manuscript and code

Summary: This is a revised version of a previous manuscript. The authors have addressed some, but not all, of the point I had raised, and some serious methodological points remain. The suggested revisions are minor in comparison to the first round, but I strongly urge the authors to either implement them or substantiate their position with a convincing rebuttal.

1 Scientific Comments

1.1 Paleo Mumbo Jumbo

A common feature of paleoclimate statistics is the introduction of methodological twists that have little or no theoretical justification. Witness the tortured Principal Component regression method of Mann et al. (1998), which has created endless backlash from statisticians for little gain. Other examples abound. The lesson is that, unless there is a clear theoretical or heuristic justification for modifying a tried and true method, one had best stick to the tried and true method. While the paper carefully describes the classic regression flavors or learning algorithms used here (PCR, elastic net, Random Forest, PLS), it also wraps them into an extremely unconventional form of bootstrapping (subselecting parts of the training period in an unspecified way) and averaging over this sample to obtain their "best" reconstructions (Section 2.2, point #6). I know of no justification for doing this, and it seems highly redundant with the cross-validation approach. I thought I had pointed this out in my original review, but I cannot find it there, as it got overshadowed by other considerations. It is now time to address this serious issue.

We thank the reviewer for highlighting this problem as we wish to develop a useful tool for paleoclimatologists while being verified by the statistical community. We decided to use the methodology proposed by Ortega et al (2015) and extend it to other statistical methods by varying different methodological parameters such as the learning period, proxy selection or the way train/test sampling is carried out (Section 4 of the submitted paper). The blind trust in the methodology of Ortega et al (2015) led us to this methodological approximation. Indeed, following this commentary of the reviewer, we have searched the bibliography for an explanation of this unconventional bootstrap approach and have found nothing convincing. Nevertheless, the methodology used to calculate scores and the cross-validation approach only applied to sample training for hyperparameter optimization is well known and frequently used in statistical learning.

Following the reviewer's suggestions we decided to keep an almost identical algorithm regardless of the method used in order to calculate the scores of a reconstruction in the same way as in the previous version. For the final reconstruction, rather than aggregating the individual reconstructions as before, the method is applied to the entire learning dataset with here also a cross-validation to optimize hyperparameters.

My recommendations are:

• Provide a theoretical justification

We have not found any theoretical justification and therefore decide to stick to tried and true methods.

• Demonstrate using simulated data that this is a sensible (I suspect this won't work, but I'm open to surprises).

• Clearly explain the rationale in the text.

• Make sure users can easily turn off this feature, in case they want to stick to tried and true methods.

We have now changed the way the final reconstruction given by the toolbox in the code is calculated and this is specified in the main text of the manuscript. There will therefore be no option for the users, who will directly obtain the result obtained by the tried and true methods (once their hyperparameters have been optimised using cross-validation and scores have been calculated using training/testing sampling). In addition of the scores calculated over testing samples, it now provides statistics (correlation, RMSE, CE) calculated for the final model that uses the whole years of observations for X and Y.

I cannot recommend the publication of this toolbox unless these conditions are met.

1.2 Statistical Models are Models too

I must reiterate the point that GMD is a journal about models, so it would be desirable to discuss the advantages of the methodological choices on modeling grounds: each of the regression methods models the data and uncertainties in various ways, and it would seem natural for such modeling assumptions and choices to be discussed here. One implicit modeling assumption they make is that the NAO is a linear combination of the proxy data, whereas the correct etiological relationship is the other way around (proxies react to climate, not climate to proxies). This inevitably leads to important biases (Frost and Thompson, 2000), as pointed out by Eduardo Zorita in his comment. Since the paper describes a toolbox, it is important that users be made aware of these caveats.

We understood from Eduardo Zorita's comment and Frost and Thompson's (2000) work that the bias is rather due to the uncertainties associated with predictors, in this case proxies. The uncertainties associated to biases due to biological/geological signals, other climatic influences, seasonal effects, etc... It means that the climate variable associated with the proxy is biased and thus the statistical link between the proxy and the climate index is underestimated, leading to biases in the reconstruction. Following the reviewer's comment, we decided to add a paragraph of discussions in section 2 dedicated to the limits of our approach and therefore of the tool we propose.

We therefore address this fundamental problem by specifying in the main text that since climate variations affect variations in proxies, we can then attempt to estimate past climate variations using the statistical methods proposed. We also discuss the problem that proxies uncertainties due to measurement and transfer methods lead to an underestimation of the link between climate variables translated by proxies and climate variations.

I also second Eduardo Zorita's suggestion of including metrics of variance in the validation, as this is easier to do for indices than fields. This and other diagnostics (e.g. R, CE) should be included on Fig 10, for instance, or in a Table.

In the second version of the manuscript we already added Tab. 4 in supplementaries. This table addresses Eduardo Zorita's comment as it presents the variance of reconstructions for different periods

or groups of periods: Training periods, testing periods, reconstruction period and learning period. To address this comment.

These statistics R, CE and RMSE are now included in Fig 10 in the next version of the manuscript.

1.3 Perfunctory Validation

Validation has improved in this version with the addition of more metrics, and an analysis of residuals, which will be very reassuring to informed readers. A more fundamental issue is the sampling: the authors currently do not specify how they perform cross-validation, with substantial implications for the estimation of generalization error. Put simply, cross-validation is a way to estimate generalization error, that is, the error that one would make by estimating values of the target (here, the NAO index) that lie outside the range of the training interval (Arlot and Celisse, 2010). That is ostensibly the goal of reconstructing pre-instrumental values of a climate index, so one wants a way to estimate this generalization error using instrumental values. Cross-validation does this by selectively removing a subset of the training interval, and using it to compute validation metrics. If one does this in a sensible manner over suitable permutations, one can show that CV provides a good estimate of the actual generalization error. Here is the rub: a lot depends on the sampling mechanism. There are basically two choices: removing points at random, or removing blocks of consecutive points. The first looks like Venetian blinds in the data matrix, so it is sometimes called "blinds-style cross-validation"; the other is called "block-style cross-validation", for obvious reasons. This makes little difference when the data are independent; but in climate timeseries, autocorrelation is almost always incredibly large, so that one gets skill from persistence alone. Thus, if you remove the year 1911 from your training sample but have both 1910 and 1912, you can produce a skillful estimate of the NAO index in 1911 without having any proxies at all! You will get enough information from past and future values of the index to produce a reasonable estimate of the index at that withheld point. In climate timeseries it is essential that cross-validation be done in the block style. Is this what was done here?

We currently use purely random sampling (i.e. the blinds-style CV) cross-validation as well as train/test sampling. We would like to specify that it is by using the "hold out" approach (e.g. train/test sampling) that we can calculate the generalization error while cross-validation is used to optimize the hyperparameters of the associated regression method. The hold-out approach differs slightly from cross-validation in that no different blocks are built, each of which is set aside at each iteration. This involves setting aside part of the sample (test sample) to finally estimate the quality of the statistical model.

However, we understand the reviewer's comment and have decided to now use the block-style approach rather than blinds-style approach for both the hold-out sampling and the K-fold cross-validation sampling and we now use the reviewer’s argument in the main text..

We emphasize that the block-style approach results in a finite number of samples regardless of the size of the train sample chosen. This therefore leads to an estimate of the generalization error or optimal hyperparameters no longer dependent on sampling and are therefore unique for a given K. Thus, if, for example, one year of instrumental measurement is incorrect (say twenty years), a block-style approach would suffer much more than a blind-style approach to bias since these data will pollute the calculation of scores and the calibration of data on a permanent basis. The blind-style approach would not completely eliminate bias, of course, but by randomly distributing potentially poorly measured data across the different samples, it can reduce bias.

Block-style vs. Blind-style approaches is now discussed in the paper but only the block-style one is used for the study. It should be stressed that if a block-style splits is performed for hold-out, the number of training splits, R, is now determined by the size of the testing (or training) samples relative to the size of the whole learning sample. For ClimIndRec users, if a block style hold-out is performed, R input is ignored and the real value of R is determined. Following the block-style splitting and in order to produce the maximum splits as possible, the first testing period encompasses the first n_{test} time steps. The second testing period is then the shifted by one time step version of the first testing period. And so on until each data of the learning period has been used at least once. These informations are specified and explained in the new version of the manuscript.

Important note for the reviewer: We understand what is embarrassing for the reviewer as our approach seems to him to apply a double validation. We insist that what we call “hold-out” (Sammut et al. 2009, *Encyclopedia of Machine Learning, p.507*) is the validation while KFCV is the method we use to tune parameters. We did not find a way to combine both and we were seriously working on for this round of review. If this can help for understanding, in Ortega et al. (2015), Nat. Geosci., they use the Preisendörfer’s rule N (Presiendörfer, 1988) to tune the number of Principal Components used over their calibration (here training) samples. As this method is only applicable for PCR, we chose the KFCV method to tune parameters as it can be used for any regression method. Hence it is very important to see KFCV not as a validation procedure but as a tuning parameter method such as the Preisendörfer’s rule N is for PCR. As we use ClimIndRec for other studies, we are actually searching for an approach to overcome the double sampling we actually do. Hence, if the approach we use in this version of the manuscript appears still irritating for the reviewer we are fully open to discuss and find a compromise with the reviewer about this, around another round of review if needed for instance. We emphasize that we do not think we are doing wrong but we know that this might not be an optimal approach.

Another issue is the value of K: large ones lead to more variable estimates, whereas small ones lead to more bias. In the case of the block-style cross-validation that needs to be applied here, K = 5 is usually found a good compromise, but obviously the code can be made flexible enough to adjust this. The other limit is leave-one-out cross-validation, where K = n. Evaluating the sensitivity of parameter tuning to this choice would be important. Note that this would fall under section 4.1.2, as under block-style validation, the length of the validation period depends on K.

This interesting comment highlighted to us that we chose K in a completely arbitrary way. In the first version of the code and paper implemented, we had a cross-validation leave-one-out. The problem we had encountered was the execution time since with R=100 and K=n doing so as 100^{n} models were built for each reconstruction studied in the paper when we had to respect the deadline of the journal. This is why in the current version we have switched to a 10-fold CV method that is less expensive in terms of computing time. We performed for each reconstruction method 3 reconstructions over the reconstruction period 1000-1970 with different values for K=5,10 and n. NSCE scores shown Fig S1 indicates that the choice of K is not affecting scores. In addition reconstructions for a same regression method but for different choices of K never have a correlation lesser than 0.

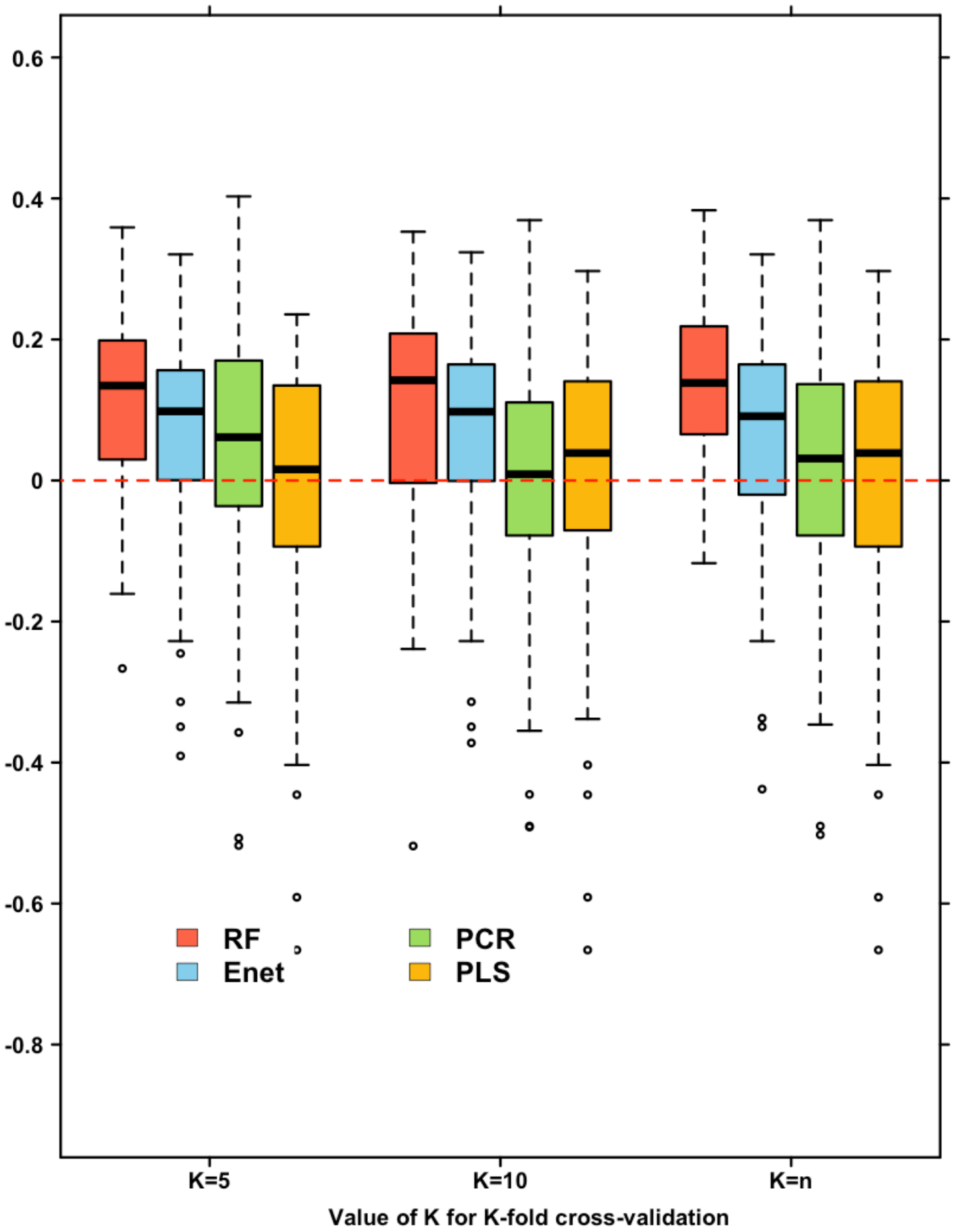

Figure S1: NSCE scores obtained by reconstructing the NAO over the period 1000-1970 with each regression method for different values for K.

Hence, in view of the reviewer's comment, we have decided to study by default the case K=5 which is now discussed in the paper. In the next version of ClimIndRec, the choice of K can be determined by the user. It should be noted that if the user chooses a block-style approach for train/test sampling, then R is ignored, and n_{train} is used to determine the samples.

1.4 Regression Methods vs Inferential Framework

I commend the authors for including a discussion of methodologies to deal with missing data (Section 5.1). One aspect that does not come out as clearly as it should is that the inference framework (e.g. the Expectation-Maximization algorithm, or a Bayesian hierarchical model) is distinct from the modeling choices, a point made eloquently by Tingley et al. (2012). Thus, all the methods used herein could be used in the framework of RegEM for instance (they would have to be embedded within), or in a Bayesian framework.

In view of the large number of changes we had to make in the last round of reviews, we failed to highlight this important problem with our toolbox. We do not exclude the possibility that in the future, this type of Bayesian approach could be implemented in the tool, which would lead to a major improvement in the exhaustiveness of the use of the proxies database. Unfortunately, we think, because we are currently limited in our theoretical and technical knowledge of this type of approach, we will not have time to look at this aspect for this round of review.

We will add a discussion largely based on these limitations of the toolbox in the paragraph that discuss the rationale of ClimIndRec, its limitations and its added-value to the classical R packages (see section 2 of response to reviewer 3) (i.e. section 2.1 in the new version of the manuscript).

1.5 Modes vs indices

I understand that the paper's motivation is the reconstruction of climate modes, particularly the NAO. However, as the authors seem fond of overwhelming readers with superfluous details, it is fair to provide a detail-oriented review. Strictly speaking, CliMoRec enables the reconstruction of indices, not "modes". The authors should explain that it would also work on ANY timeseries, including hemispheric averages (e.g. Northern Hemisphere Temperature). Accordingly, they might also consider rebranding it: ClimIndRec, perhaps?

Very good catch. We already highlighted that the toolbox can reconstruct any climate timeseries so that we indeed choose an inappropriate name for it.

The reviewer's suggestion being very relevant, we decided to change the name of the toolbox by ClimIndRec which is more appropriated.

1.6 Forcing attribution

The authors used the common method of Superposed Epoch Analysis to evaluate the response of the NAO to volcanic forcing, but do not quantify uncertainties. Obviously, not all of the wiggles are meaningful, and some methods exist to tell which ones are (e.g. Rao et al 2019, 10.1016/j.dendro.2019.05.001). As it stands, the authors mention significance, but I could not get details on how that was established. Without proper uncertainty quantification, one cannot rule out the possibility that none of the wiggles stand out of the noise.

We put the method we use to calculate this significance in the supplement. Actually, we use a very similar approach than Rao et al. 2019 (for volcanic response but not fire response in their paper) which is the method used by Ortega et al. (2015), a Monte-Carlo approach. We first randomly select 1000 sets of 11 ”fake” volcanic eruptions and each is centered to 0 for the year N (year of the eruption). For each, a superposed epoch analysis of the 11 fake eruptions is performed. We then retain the 90% level of the N+1 response among the 1000 11-length composite of fake eruptions as the significance level. In Rao et al. (2019), their approach is more sophisticated as the significance level is calculated in the same way but for each time lag of eruption, which we are not doing here.

According to this interesting comment we decided to use and develop in the supplement the Rao et al. (2019) approach as suggested by the reviewer.

1.7 Feature selection

The conclusion states: We have shown that for Enet, PLS and particularly PCR which is frequently used in paleoclimatology, selecting proxy records with a strong correlation with the index to be reconstructed over the training periods is a good way to improve the NSCE scores, and hence it allows more reliable reconstructions (section 4.1.1). Contrarily, RF gives more reliable reconstructions using the whole set of records (section 4.1.1). This is entirely unsurprising for PCR and PLS, as they do are not designed to achieve feature selection (Friednman, Hastie & Tibshirani, 2008, chap 18). However, that is one of the purposes of RF, so it is entirely expected that it would not need additional screening prior to application. I'm a little more perplexed that Enet benefits from screening, as its L1 penalty encourages zero coefficients that effectively turn off features (here, proxies) that don't help prediction.

I suspect things might be different if the LASSO is used first for feature selection, and then ridge regression applied to minimize prediction error (with the correct parameter choice), as opposed to apply both at once. The extreme variance suppression of the Enet estimate in Fig 10 suggests that the parameter choice is not optimal, in this case at least.

Reviewer 3 also highlighted that he has some doubts about our choice of Elastic Net. He suggested to apply an Adaptive Lasso approach (Zou and Zhang, 2009, see comment 6 of reviewer 3). This response is then likely similar to the one we provided too reviewer 3 (see response to comment 1.7 of reviewer 3).

If Lasso+Ridge had provided better results than the methods presented in the former version of the manuscript, we would have certainly added it, and modified the figures accordingly, but given the short time available for resubmission, and the negative results, we decided not to. Fig S2 and S3 presented below show the results obtained for Fig 6 an Fig 7, but where PCR and PLS (outperformed by Enet and RF in this case) CE scores are respectively replaced by those obtained for Lasso+Ridge and adaptive Lasso CE scores.

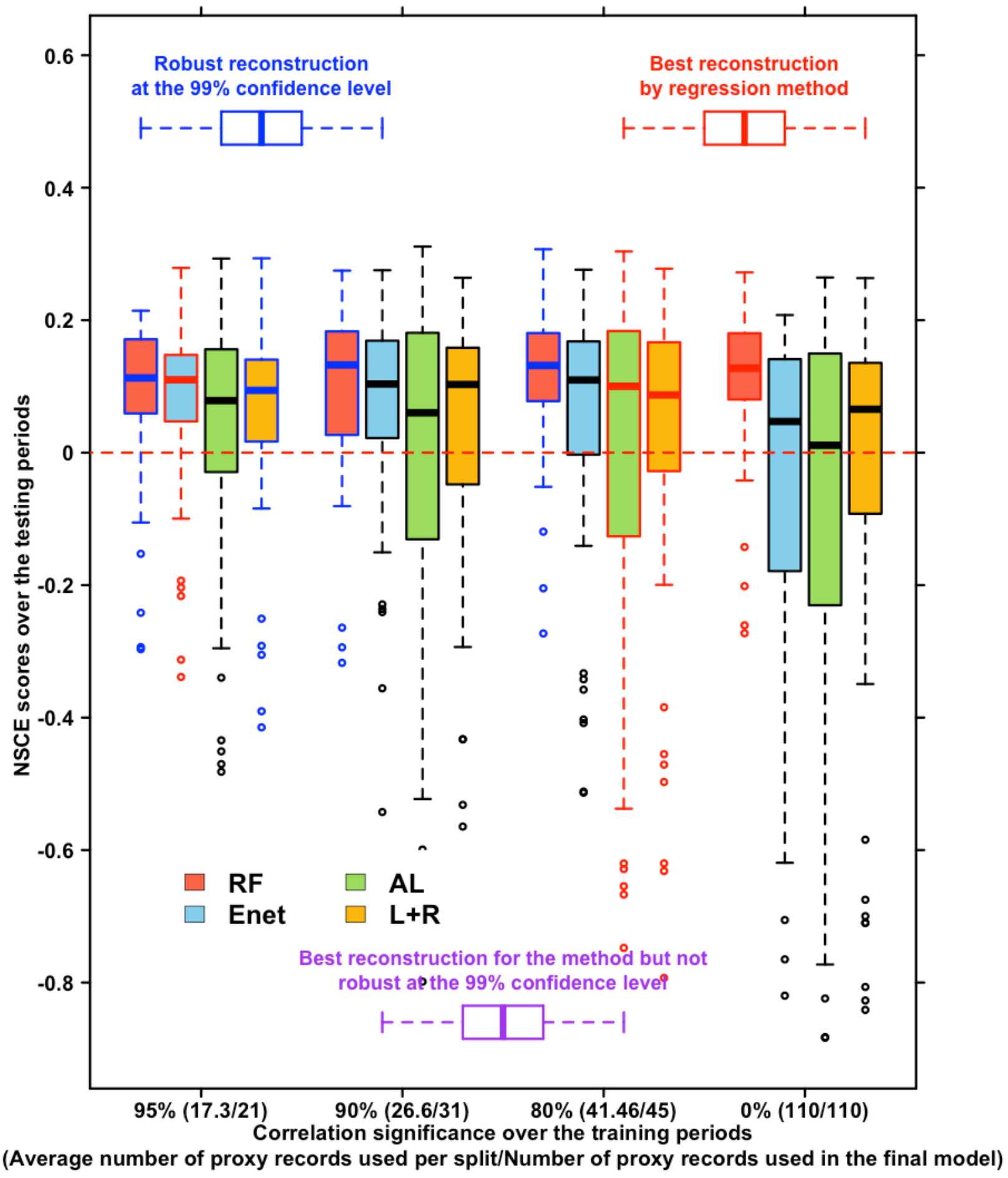

Figure S2: Same as Fig 6 of the manuscript but PCR method has been replace by adaptive Lasso (AL) and PLS has been replaced by Lasso+Ridge (L+R)

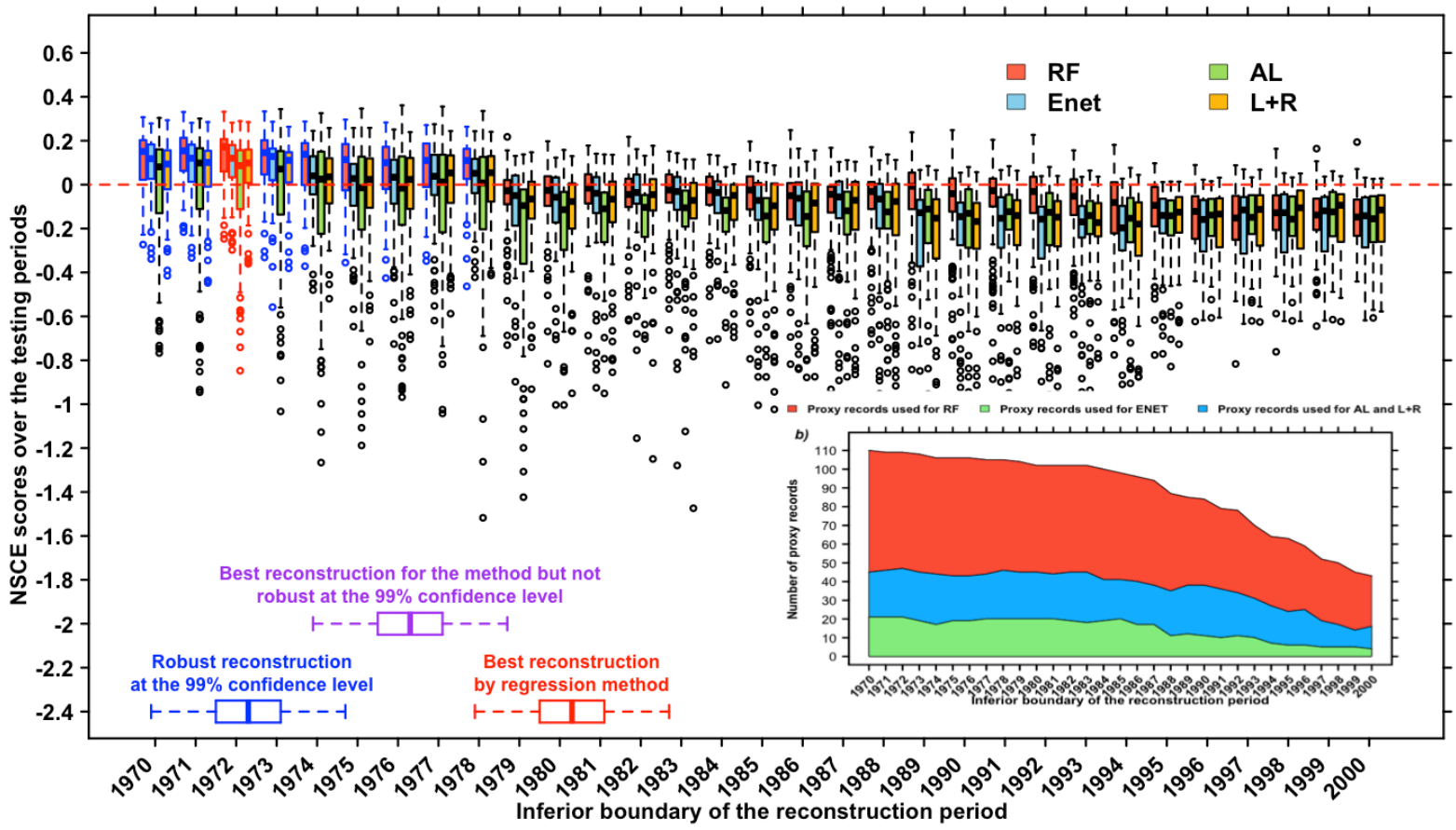

Figure S3: Same as Fig 7 of the manuscript but PCR method has been replace by adaptive Lasso (AL) and PLS has been replaced by Lasso+Ridge (L+R)

Fig S2 and S3 show that the adaptive lasso does not provide significantly better results than Elastic Net that are already worse than Random Forest CE scores. Of course, this might not be true for the reconstruction of other climate indices and the potential use of adaptive lasso in future climate reconstructions might be relevant. In addition, we found that the best adaptive lasso reconstruction (the one having the best scores on Fig S3) has a correlation of 0.98 with the one obtained using the elastic model optimized using nested cross validations, which, as mentioned above, has higher validation scores. As mentioned above, given that none of the two provides an improvement to the 4 former methods for our target index (e.g. NAO) none of the two methods have been included in the paper. But they have been integrated in the new online version of the code (ClimIndRec1.1.r in Zenodo and Github), which is an updated version of the one presented in the manuscript (ClimIndRec 1.0).

In terms of screening, the reviewer can see Fig S2 that the best way for using Lasso+Ridge in our case uses a pre-screening, but for proxy records significantly correlated at the 80% confidence level, which is less constraining than the one we use for Elastic Net (95%).

1.8 Uncertainties

Buried in the supplement is the definition of how uncertainties are calculated with CliMoRec; it turns out to be the standard error of residuals, a perfectly reasonably choice when the number of predictors stays constant over time, but an otherwise suboptimal one. Indeed, as proxy density decreases back in time, so does the information available, and therefore error bars should widen back in time.

Indeed, we do not perform a nested reconstruction (such as in Wang et al. 2017 for instance) in this toolbox and we use methods that does not deal with missing data, thus there is no time-dependent uncertainties for the moment.

While it is true that this point remains depressingly under-appreciated in the paleoclimate community, some methods can deal with this, like BARCAST (see Tingley and Huybers, 2013, , Fig. 1) and LMR (Hakim et al., 2016). In the regression context, this can be taken care of with frozen network analysis, or bootstrapping, as done in PAGES 2k Consortium (2017) (see their Figs. 7 and 8). At the very least, the

authors should flag that this choice neglects changes in proxy availability over time, highlighting potential improvements for future versions of the toolbox.
This issue is now discussed in the new section 2.1 mentioned section 1.4 of this review, and we will highlight that this may be present in the future version of the toolbox as we also aim at implementing methods that deal with missing data.

2 Editorial Comments

Gallicisms have almost entirely disappeared; nice job! One remains: “contrarily” should be replaced by “on the contrary”.
It partly almost disappeared thanks to the reviewer suggestions of the last review. We thank the reviewer for his very helpful suggestions.
As pointed out before: The description of methods is incredibly tedious. Sections 3.1.2, 3.2.1, 3.3.2 explain the obvious step of linear model prediction as a matrix multiplication. None of this is useful in any way as long as the code is shared. I maintain that the mathematical details of these regression methods is of limited use: people with a statistical background already know them, and people without a statistical background are unlikely to read them. This section should be moved to an appendix, so that the few readers who really need the details can find them, but it doesn’t clutter the narrative.
The present arguments put forward by the reviewer indeed seem indisputable to us and finally convinced us, after discussions between the authors, to move this section to the supplement of the paper.
What would be interesting is to discuss the modeling assumptions underlying these methods (as requested above), but that is not what is done here. Progress has also been made in that the authors are now using the up-to-date version of the PAGES 2k database. Yet, they insist on calling it Pages 2K. Not a deal-breaker, but it would be nice to use the correct spelling (PAGES 2k).

We apologize that we did not correct it the right way in the last version of the manuscript.
The next version of the manuscript is now spelling the database as PAGES 2k.

Re: code, the GitHub link works, but Zenodo registration is still a good idea to encourage code citation.
The GitHub link has been created: https://github.com/SimMiche/ClimIndRec with the new changes
The code is now available on the following Zenodo link:
https://zenodo.org/record/3372760#.XVxQGy2B288

Response to review comment 3

This author response is organised for each comment as follows:
(1) Comment from referee/public
(2) Author's response
(3) Author's change in the manuscript and code

The authors provide an overview of four statistical methods together with code for the development of proxy reconstructions. I agree with the first reviewer that the paper (and code) provide a valuable contribution to the literature, particularly by making the statistical tools easily accessible and comparing their performance. However, I also share the second reviewer's concerns about the content, presentation, and format of the paper. In particular, the authors need to highlight the value-added of their manuscript compared to existing work already documenting the implementation of existing software packages.

Main Comments
1. Calling R-code a 'computer device' seems strange. Why not release a formal R-package on CRAN? Or alternatively, call it 'software' or just 'R-code'.

We did not propose a R-package because we decided to develop a tool for which the use is close (in a very simpler way) to the use of climate models for instance. The development of an additional R package may take a lot of time because this requires a specific and minitious formatting and presentation and a systematic testing, especially since the tool currently developed also works with bash scripts and not only R codes. However, we do not exclude the idea of adapting this tool to an R package in the future. Unfortunately, we do not think we have time for this round of review but it is something we are going to look at seriously, as it has already been suggested by one of the co-authors (Marie Chavent).

In the new version of the manuscript, we specify that this tool acts as a software rather than a computer device.

2. Please clarify more carefully what the value-added of the paper is. There already exist R-packages (which are used within CliMoRec) that run PCR, Lasso, Elastic Net etc. What additional benefits does CliMoRec provide over the existing packages and their standard implementations? This would be important to highlight for potential users who have to choose between just using e.g. the 'glmnet' package for Elastic Net, vs. CliMoRec.
The packages mentioned by the reviewer allow to use the methods individually, each following a specific format and is not necessarily tailored for climate reconstruction. In the paper we identify that using different regression methods, set of proxies and set of observation years strongly impact the final reconstruction obtained. In the manuscript and the code, we then provide a complete methodology allowing to perform the reconstructions consistently for all the four methods, to evaluate the robustness of the given reconstructions, to compare them and to extract the most robust among them. This is something a user can not do by just using "glmnet" and "randomForest" packages as they are for instance. The user will need a lot of coding to perform evaluated reconstructions as we do and we develop in the manuscript (the reviewer can see it in the code on Github).

However, to reinforce this point on the manuscript, we added discussions about why ClimIndRec (its name changed following the suggestions of reviewer 1) is a powerful and simple tool, that partly uses advanced R packages to perform robustly evaluated reconstructions.

3. Please provide a bit more detail on important tuning parameters. For example, the section on choosing the penalty terms lambda and alpha (in Elastic Net) is very brief and it is not clear for practitioners whether the particular form of cross-validation employed is generally applicable.
We might have not been clear about hyperparameters tuning, particularly for the Elastic Net method that needs two parameters defined in continuous sets to be tuned. The tuning is performed by using a double cross-validation by shuffling each combination of α and λ according to discretized versions of the continuous sets in which they are defined. In section 2 of the last version of the manuscript, the rationale of K-Fold cross-validation is implicitly indicating that the dimension of θ can be higher than 1, which is the case for Elastic Net with $\theta = (\alpha, \lambda)$.

In the next version of the manuscript, the description of K-Fold cross-validation is now explicitly specifying that the dimension θ can be greater than 1 which means that they need to be cross-tuned by shuffling and computing different combinations of their possible values. Following reviewer 1 comment, the section 3 of the previous version of the manuscript where the regression methods are developed and the hyperparemeters are identified, has been moved to appendix 1. However, for the Elastic Net section, we now specify that two simultaneous cross-validations are applied to tune α and λ. We also specify the discretized version of the continuous in which they are defined that is used by default in ClimIndRec. However, it can be easily modified by its user in the ClimIndRec main script.

4. More generally, please clarify why the particular model selection procedures are chosen? Why PCR, Elastic Net and Random Forests? There are many other alternatives (see e.g. adaptive Lasso, general-to-specific selection, etc.). The manuscript would benefit from being placed in the wider context of other methods being available as alternatives.
It is implicit in the last version of the manuscript. The choice of the methods, even if not exhaustive, is not purely arbitrary. We decided to add PLS because of its strong closeness of the PCR. Elastic Net has been proposed in order to introduce regularized regression methods, barely used in paleoclimate science. And we finally proposed Random Forests because it is now a very commonly used and well verified regression method that has shown its strength in many fields. Of course we could add other methods but we think presenting these four methods is enough as we provide a tool and a methodology to compare them that can thus be applied to alternative reconstruction methods. Next versions of ClimIndRec will be dedicated to the implementation of new methods which is already specified in section 2 and conclusions (see response to reviewer 1). More generally, we believe that our choice, even if not complete, it is still valuable because we :

i) Cannot compare every regression methods that exists

ii) Provide a methodology to compare methods which can be extrapolated to others than those which are investigating in the paper

iii) Specify that other methods will be implemented in next versions while the methodology developed in the text to compare them will stay the same.

iv) Outperform the commonly used method (PCR) with Enet and RF and thus show the relevance of the chosen methods.

In the actual version of the manuscript we specify why we chose these procedures in section 2. In conclusions we explain that our method evaluation can be extrapolated to alternative methods, among which some will be implemented in future versions of ClimIndRec.

In comment 6 of the actual review we mention that a 1.1 version of ClimIndRec is ready (the one investigated in the new version of the manuscript being version 1.0 of ClimIndRec) and it includes

adaptive Lasso and Lasso+Ridge (see comment 1.7 of reviewer 1) approaches. In section 6, we show that both of the methods do not provide significantly higher scores than Enet while they provide significantly lower scores than RF, which supports the relevance of the chosen methods for this study (more details in comment 6 of this review).

5. CliMoRec appears to use R-code based on existing packages implementing PCR, Lasso, Elastic Net, etc. rather than developing these functions itself. These existing packages have to be cited and credited in the methods sections (such as 'glmnet' in R for elastic net). Not citing the software packages is poor practise and particularly important for a paper that discusses software. Most packages used in CliMoRec have a corresponding *Journal of Statistical Software* (JSS) paper that should be cited. For example, 'glmnet' used by the authors (in the first line of code of CliMoRec on Github) is documented in Friedman, Hastie, and Tibshirani (2010) https://www.jstatsoft.org/article/view/v033i01 and should be cited as such.
If there is no JSS paper available, then the R-packages should be cited through CRAN.
We thank the reviewer for pointing out this. Each reference of the different packages used in the R code provided are now cited and included in the bibliography of the manuscript.

6. On Lasso and Elastic Net: Lasso is not a consistent model selection method with oracle properties, instead, the authors may want to refer readers to the Adaptive Lasso and Elastic Net. See e.g. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429., or Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. Annals of statistics, 37(4), 1733.
Reviewer 1 also highlighted that he has some doubts about our choice of Elastic Net. He suggested to first apply a lasso for variable selection then applying a rige model using the kept variables. This response is then likely similar to the one we provided too reviewer 1 (see response to comment 1.7 of reviewer 1).

After reading Zou and Zhang (2009), we have decided to implement this approach using the "glmnet" R package in a 1.1 version of ClimIndRec (the actual version being 1.0) to investigate if it potentially provides a better approach than the nested cross-validation we perform for Elastic Net.

Adaptive Lasso consists in first building a ridge model where λ is optimized using K-fold cross validation then using the inverse of the estimated ridge regression as penalty factors for building a lasso regression model (Zou and Zhang, 2009) where λ is optimized using K-fold cross-validation.

If adaptive lasso had provided better results than the methods presented in the former version of the manuscript, we would have certainly added it, and modified the figures accordingly, but given the short time available for resubmission, and the negative results, we decided not to. Fig S1 and S2 presented below show the results obtained for Fig 6 an Fig 7, but where PCR and PLS (outperformed by Enet and RF in this case) CE scores are respectively replaced by those obtained for Lasso+Ridge and adaptive Lasso CE scores.

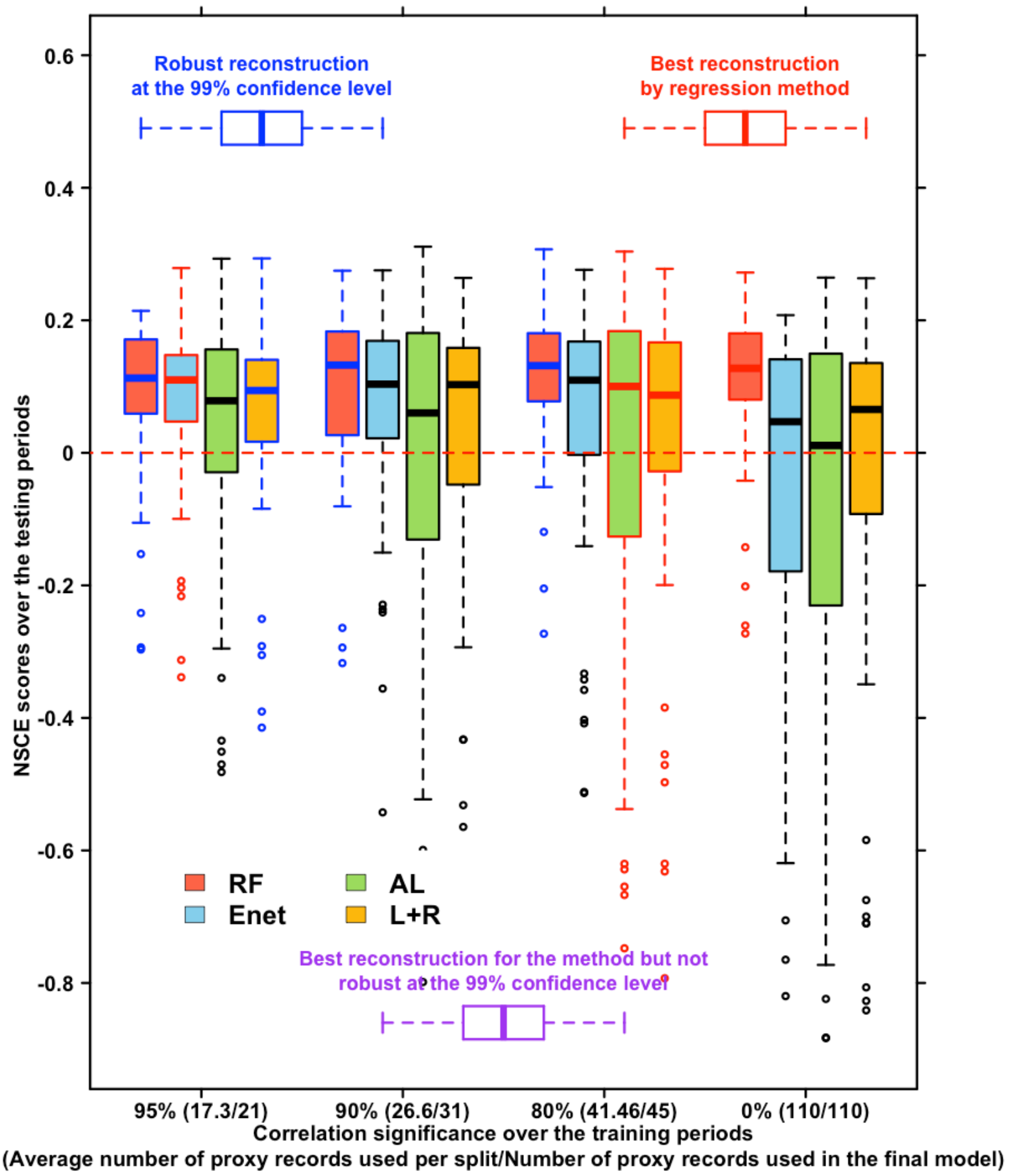

Figure S1: Same as Fig 6 of the manuscript but PCR method has been replaced by adaptive Lasso (AL) and PLS has been replaced by Lasso+Ridge (L+R)

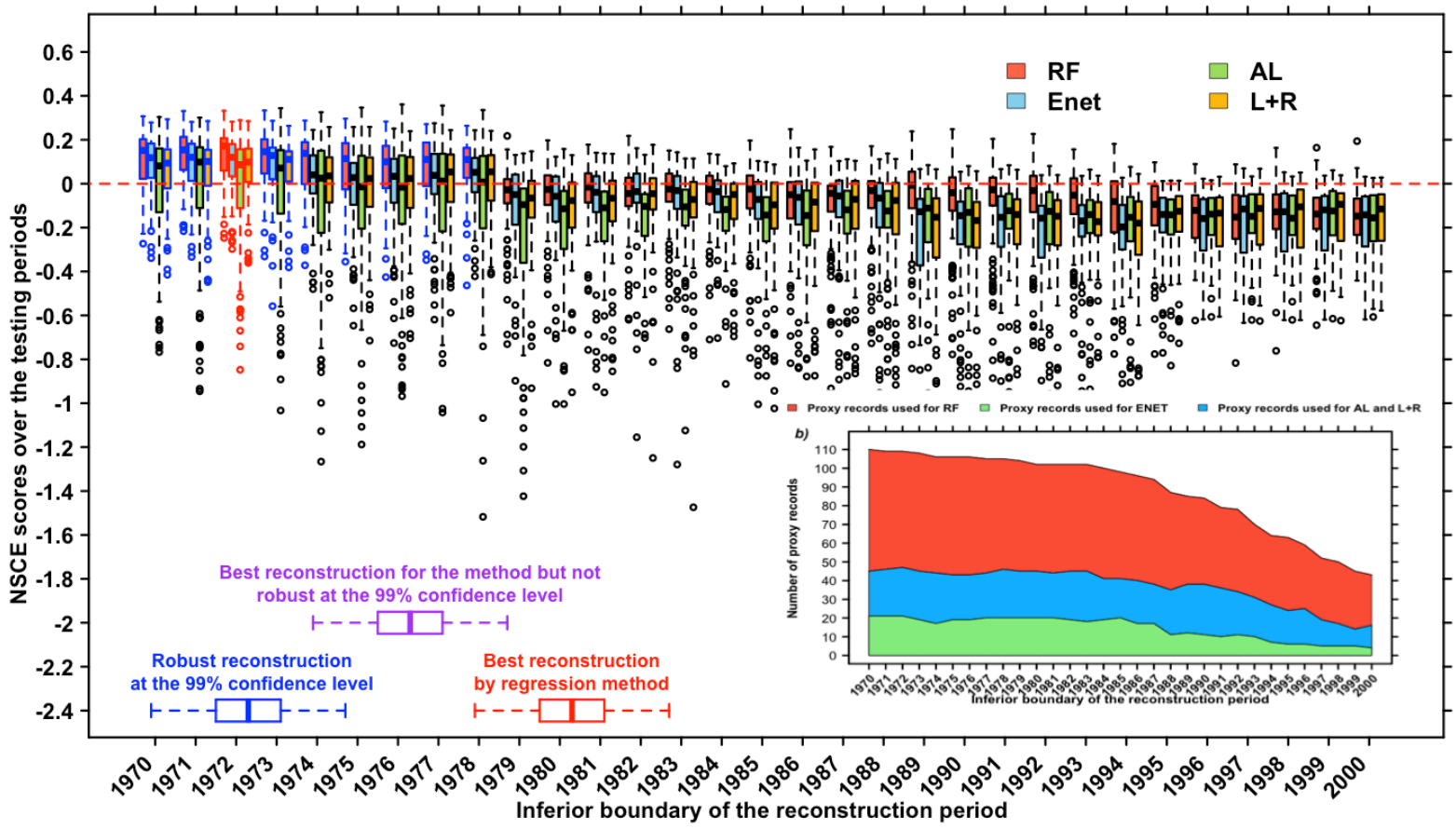

Figure S2: Same as Fig 7 of the manuscript but PCR method has been replaced by adaptive Lasso (AL) and PLS has been replaced by Lasso+Ridge (L+R)

Fig S1 and S2 show that the adaptive lasso does not provide significantly better results than Elastic Net that are already worse than Random Forest CE scores. Of course, this might not be true for the reconstruction of other climate indices and the potential use of adaptive lasso in future climate reconstructions might be relevant. In addition, we found that the best adaptive lasso reconstruction (the one having the best scores on Fig S2) has a correlation of 0.96 with the one obtained using the elastic model optimized using nested cross validations, which, as mentioned above, has higher validation scores. As mentioned above, given that none of the two provides an improvement to the 4 former methods for our target index (e.g. NAO) none of the two methods have been included in the paper. But they have been integrated in the new online version of the code (ClimIndRec1.1.r in Zenodo and Github), which is an updated version of the one presented in the manuscript (ClimIndRec 1.0).

Minor Comments

1. P14. Line 1: “most simple” replace with “simplest”
2. Section 2.3: the title “Mathematical Formalism” seems strange and not entirely clear.
3. P 13, the sentence “For Enet method” is missing a word, maybe “For *the* Enet method”?

We thank the reviewer for pointing out these errors which will help us to provide a more readable version of the manuscript.