1 Scientific Comments

I’ll start with what I like about the paper: it applies several methods to the same dataset, and the results are fairly consistent among methods and with another recent reconstruction, in which one of the authors was involved (Ortegal et al, 2015). That’s about it.

We thank the reviewer for this positive comment. Nevertheless, as a general response to the main reviewer’s criticisms below, we would like to highlight that our study is proposing novel regression methods that have, to our knowledge, not yet been applied to climate signal reconstructions. In addition, we found in previous studies cited in this manuscript (that concerns the reconstruction of climate modes, but not of climate fields), several issues in the classical methodological approaches. Our objective here is to assist paleoclimate experts in making the best out of their proxy databases with valid and robust statistical assessments. More specifically, using a new metric that we discuss below, we show how to evaluate different reconstructions of the same climate index but with different methodological choices (regression method, proxy network, length of the period on which the regression model is built). The wide range covered by the scores shows that the selection of these inputs is an important step to obtain a reconstruction as robust as possible. Furthermore, to make the production of such reconstructions more straightforward and facilitate its use to potential users, we have developed a code that simply requires a few parameters as input and that provides a reconstruction of a given climate index for a given proxy record database. In addition, the code provides an ensemble of scores that evaluates the reconstruction. By varying the different methodological choices, the user of the code can then perform several reconstructions and pick the one that has the best scores. This is why we do not submit this paper to Climate of The Past, as we would like to make climate signal reconstructions more transparent and easily accessible and verified by the community.

Furthermore, we believe this statistical toolbox could be improved in the future by including further refinements in follow up versions which constitute an additional reason for which we prefer to submit this paper to GMD. Last but not least, we believe that providing sufficient level of details concerning the mathematical rationale behind our methods is very useful, while they are hidden in the appendix in journals like Climate of the Past, which are more focused on the scientific results.

1.1 This is no “big data”

Few things are more irritating than people pretending to do "big data" when they actually don’t. The authors only end up using a few dozen proxies, and only reconstruct a single index. Nothing wrong with that, but it’s not "big data" by any stretch of the imagination. In fact, except for the random forest method (which is only useful in the presence of hundreds or thousands of predictors, therefore not very useful here), all of the methods described are classic forms of linear regression. Anyone is free to call that "machine learning" (since most ML methods are regression in one form or another), but the larger problem is that this is a modeling journal, and I see very little in the way of statistical modeling here.

We entirely agree that what is done in this paper is not “big data” and we didn’t intend to claim we did it. The word “big data” was mentioned twice in the submitted text with the only aim of providing a context, once in the abstract (line 6) and once in the introduction (page 4, line 8). We are actually claiming that the emergence of big data that followed the innovation in technologies and data storage has led to the development of new regression methods in the 2000's, in particular elastic net regression and Random Forest (Breiman 2001; Zou and Hastie 2005). Those methods have indeed been developed in order to address high-dimensional problems (p>n), that Principal Components Regression and Partial Least Squares poorly deal with. However, since the word "big data" can be misleading, we have decided to remove it in the revised version.

Random Forests are indeed particularly useful for high dimensional data with numerous predictors such as boosting gradients or neural networks. However, in the new version of the code, by using the Nash-Sutcliffe Coefficient of Efficiency, we have found significantly better results for the Random Forest and the Elastic-net methods than for the PLS and the PCR methods (this is illustrated in the Fig. R1 that will replace Fig.5 of the previous manuscript), which shows that adding these methods even in a low-dimension study such as in ours can be more efficient than using classical forms of linear regression. Additionally the code we provide allows to choose the network of proxy records that is used for the reconstruction. As the number of available paleoclimate data is constantly growing (even if it does not reach hundreds of thousands yet), we claim that regression methods adapted to high-dimensional problems such as Random Forests will sooner or later, become particularly useful for climate index reconstructions. We have added a few words on this subject in the discussion of the manuscript.

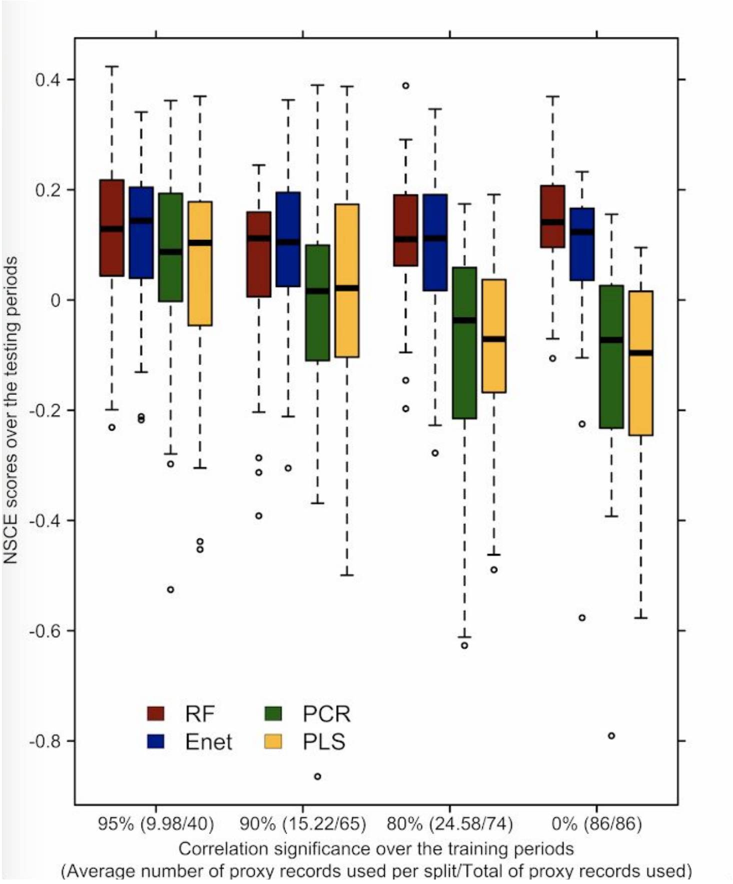

Fig. R1: Nash-Sutcliffe Coefficient Efficiency (NSCE) scores obtained for each method for the reconstruction period 1000-1970 and for different significance for the correlation test performed on the training periods: 95%, 90%, 80% and 0%. Red boxplots give the NSCE scores for the Random Forest method. Blue boxplots give the NSCE scores for the Elastic-net method. Green boxplots give the NSCE scores for the Principal Components Regression method. Yellow boxplots give the NSCE scores for Partial Least Squares method.

1.2 Suboptimal Methods

Furthermore, the chosen methods are unable to deal with missing data, forcing the authors to limit the calibration to a set of complete records, thereby jettisoning important information.

Meanwhile, at least three methods have been proposed to estimate past climates using discontinuous records:

1. The Expectation-Maximization algorithm (Dempster et al., 1977) and its regularized variants (Schneider, 2001; Guillot et al., 2015), as used by Mann et al. (2008) to reconstruct the global mean surface temperature, for instance.
2. Bayesian Hierarchical Models, that treat missing observations as extra parameters (Tingley and Huybers, 2010a,b; Tingley et al., 2012; Tingley and Huybers, 2013; Barboza et al., 2014).

3. Data assimilation approaches, for instance the Last Millennium Reanalysis framework (Hakim et al., 2016; Singh et al., 2018).

All of these methods have code that is publicly archived, often in open-source languages like R. Restricting themselves to antiquated regression methods forces the authors play a dubious game of optimization on the various training and verification sets, to offset the disadvantage of restricting the network to a gap-less training set. This is suboptimal on methodological and computational grounds.

In this study, we focus on climate variability modes, which is only a part of the global climate. We applied dedicated methods aiming at improving the reconstruction of these modes. Our techniques can certainly be further improved, but as it stands, we believe that they add new potentialities to the regression approaches currently at use.

Many paleoclimate studies (cited in the manuscript), such as our study, are focusing on reconstructions of global climate indices in time while others (e.g. the studies the reviewer mentioned here) reconstruct a particular climate variable (usually temperatures) in space and time (e.g. Climate Field Reconstruction methods). This paper is actually clarifying and adding methodological clue and gives an accessible tool to help paleoclimatologists to build more robust climate index reconstructions.

Although both approaches aim at reconstructing past climate, the question and focus of the paper is not to show if one is better than the other, but to try to further develop one of them.

Concerning data assimilation methods, we certainly agree that these are very useful methods, but we do not believe that these methods, difficult to develop within paleoclimates, necessarily discard other more simple statistical models.

We believe that science can benefit from a variety approaches, since it contributes to build robustness of its results. Therefore, we acknowledge the existence of the three methods depicted by the reviewer, and discuss them shortly in our manuscript, but we do not think there are decisive arguments showing that our approach is necessarily weaker, although this is not the scope of this paper to prove it at this stage.

1.3 How uncertain?

An even more serious issue is that the authors do not provide any measure of uncertainty for their reconstructions. They could do so via any defensible method that has been applied in paleoclimate investigations, e.g. parametric or non-parametric bootstrap, jackknife, or maximum-entropy bootstrap (Vinod and de Lacalle, 2009).

We thank the reviewer for pointing out this major omission (also mentioned by Anonymous Reviewer 1): that is the importance of assessing the reliability of our reconstruction.

The uncertainties we now provide are calculated as in Ortega et al. (2015) using the residuals calculated over the 50 training periods. These regression uncertainties are represented by the standard errors (s.e.) of the regression, calculated as the root of the sum of the squared residuals divided by the degree of freedom over the training periods divided by the degree of freedom:

$$s.e = \sqrt{\frac{\sum_{i=i}^{n_{train}} (Y_{train} - \widehat{Y}_{train})}{n_{train} - 2}}$$

Where n_{train} is the length of the training sample, Y_{train} the true values of the NAO index over the training period, and $\widehat{Y}_{train}$ the fitted NAO by the regression model over the training period.

An uncertainty band 2*s.e. is calculated for each of the 50 individual reconstructions and the envelope of this 2*s.e. uncertainty bands is our estimate of the total uncertainty range of the final reconstruction (as a sum of the regression uncertainty plus the parameter uncertainty).

We added regression uncertainties in a table and on the figures where the reconstructions are shown. Also, the code we deliver provide standard errors for each member of a given final reconstruction.

1.4 Statistical Models are Models too

I feel compelled to point out that this is a journal about models, so it would be desirable to discuss the advantages of the methodological choices on modeling grounds: each of them models the data and uncertainties in various ways, and it would seem natural for such modeling assumptions and choices to be discussed here (more so than say, Climate of the Past, where the current manuscript would be a better fit in present form). 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). Again, some of the methods mentioned above can deal with that, and the authors should consider using them.

We have explained before our motivation for submitting the paper to this journal rather than to *Climate of the Past*: the idea is to propose a statistical modelling tool, which will be available to the community and could be further developed in a transparent way, rather than to only propose a new NAO reconstruction. We have been encouraged for this by the editorial guidelines of GMD which include 'statistical models'. Nevertheless, we leave it to the editor to decide whether our study is suited for GMD or not.

Regarding the modelling assumption: stating that the NAO is a linear combination of the proxy data is something about which we have been unclear in the manuscript but this is not what we have meant literally. "NAO index can be reconstructed from a linear combination" would be a more suited sentence. We hope that the reviewer agrees with this one and we have revised the manuscript so as to avoid such shortcuts.

1.5 Perfunctory Validation

Another major problem is that the authors carry out a very perfunctory validation using a metric (correlation) that is known to only reward phase coherence (Wang et al., 2014). At the very least, the authors should explore the Reduction of Error and Coefficient of Efficiently (Nash and Sutcliffe, 1970) statistics, which have been used for more than 25 years in the

dendrochronological literature (Cook et al., 1994). Another useful measure for point forecasts is the Continuous Ranked Probability Score (Gneiting and Raftery, 2007).
We agree that the results may be sensitive to the choice of the calibration/validation metric. Thus, we have also calculated Root Mean Squared Errors as a new validation score. It gives very similar results than correlations. We thank the reviewer to suggest this more sophisticated statistics that will be added and used as the main metric in the manuscript on top of the correlations and RMSE: The Nash-Sutcliffe Coefficient of Efficiency (NSCE). The NSCE scores is indeed helping us in many ways. It shows that all the reconstruction made using the Vinther et al (2003) NAO index are not reliable since their NSCE scores are not significantly different to 0 (following student test on the scores obtained from the individual reconstructions). However, using the Jones et al (1999) index (which is exactly the same as Vinther et al (2003) index on their common period) we obtain more robust validation scores (i.e. significantly higher than 0 at 95%).

If the authors were making interval forecasts, which they should, the sharpness of their prediction bands should be evaluated by an Interval Score (Gneiting and Raftery, 2007).
We thank the reviewer for this interesting comment. Nevertheless we should confess that what the reviewer is requesting here is not very clear to us even after carefully reading the reference mentioned. As a response, we can say that in the revised version of the manuscript, we are now properly computing uncertainties (cf. point 1.3) and notably for the validation scores, which correspond to the "forecast section" from our methodology.

Finally, an obligatory measure of any statistical forecasting is to inspect the quality of residuals: since regression relies on residuals being Gaussian, independent and identically distributed, any statistics book (e.g. Wilks, 2011) says that the residuals should be tested for these features. This should at least be present in an Appendix.
We agree that this is an important assumption to check. We have then check this assumption for the best reconstruction of each method (presented Fig. 11 of the previous manuscript) and we have added a figure showing the p-values of Shapiro-Wilk tests obtained for the 50 individual reconstructions for each of them (which have the best NSCE scores on average). Also, we have updated the code to provide the p-values of this test as an additional output.

1.6 Double dipping
The authors pre-screen the proxy network for correlation to the NAO index. What isn't clear is whether that is done as part over the model training, or whether this is done over the entire instrumental era (or the parts of it that overlap with each proxy series). If the latter, this is an example of "double-dipping", whereby information from the test set is used as part of training, leading to overoptimistic results. I could not ascertain this from the paper, so a clarification is necessary.
This comment is very useful firstly because we have been unclear on this point and secondly because it helps us to actually find out that we were doing double-dipping. Indeed, as the proxy records are selected over the entire instrumental era, the model built over the training

period uses proxy records that are, at least partially, coherent with the NAO index over the testing period, which is supposed to be independent. To correct this issue, we decided that the subselection of proxy records based on correlation test with the NAO has to be made always on training period, which means that there is no *a priori* information about the coherence between the NAO index and the selected proxy records made over the overlap period with the NAO. We have modified the code and all the results following this improvement in our approach. This does not affect much our results in the end, but is clearly an improvement in the coherence and rigor of our method, for which we thank the reviewer again.

Why use the PAGES2k version 1, and not PAGES 2k version 2 (PAGES 2k Consortium, 2017)?
Pages 2k version 2 was not available when we started this study. We thank the reviewer for highlighting the updated version, which is now used in the new version of the manuscript.

Also, the forcing of Gao et al. (2008) is known to contain many errors, which have been corrected by the vastly more complete dataset of Sigl et al. (2014). This could explain the very weak signals observed in the paper's Superposed Epoch Analysis. I recommend using the best available data.
We thank the reviewer for pointing us to the potential errors present in the Gao et al. (2008) reconstruction. Indeed, this reconstruction is now quite old, and we agree that the more recent reconstructions may have corrected some of the errors from former ones. Thus, we have removed the use of the Gao et al. (2008) reconstruction and only kept Sigl et al (2014) and Crowley et al. (2013) in the analysis of the manuscript. The inclusion of Gao et al. (2008) in the submitted manuscript was aiming to better explore potential uncertainties, but we agree with the reviewer that since Sigl et al. (2014) built on the reconstruction of Gao et al (2008) trying to improve it, this latter one has been superseded.

2 Editorial Comments

The manuscript reads like a literal translation of a chapter from a French PhD thesis. That means it is 1) overloaded with tedium intended to show that the main author knows what (s)he is talking about; (b) chock full of gallicisms.
We have worked hard for improving the language in this revised version and a native english colleague has agreed to review it before submission.
As mentioned by Anonymous reviewer 1, the introduction of this paper was very heavy and difficult to read, with a lot of technical details that were not always useful. The introduction of the paper has been largely reduced.

2.1 Tedious writing

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. Also, an entire appendix is devoted to a user's guide, which should really be a readme file on GitHub. Please do not waste the readers' disk space and printer ink with this.

While writing the first version of the manuscript we indeed hesitated to put section 3 in the main text and not in the appendix. We believe that it may be useful to have all the necessary details in the main text, which was one of the reasons why we choose GMD. We have asked the editor about this issue and she supports our choice since it may improve clarity for people that are non-expert in statistical models. Indeed, we acknowledge that the reviewer is a great expert in statistical modelling, but our aim here is to gather a larger audience, and notably the paleoclimate record experts, who may be interested in having further details to precisely follow our methodology. Thus, we believe that this level of details is useful and this explain why we chose GMD instead of Climate of the Past.

Following the reviewer's advice, the user's guide has been removed from the appendix and is now available in a readme file on GitHub, where codes and data are also available (see section 2.3 of this response).

One of the most tedious parts is that the *PAGES 2k Consortium* (2013) paper is consistently referred to as "the Pages 2K database 2014 version". Since it was published in 2013, why insist on calling it 2014? Also, the consortium's name is "PAGES 2k", not Pages 2K.

As we have updated the database in our code, we now call it the "P2k-2017 database" in the new version of the manuscript.

In section 3.1.3, several approaches are mentioned to choose the truncation parameter
(none, it should be said, with the aid of any statistical theory), but they are not used. Either leave them unsaid, or mention them and use them (e.g. by comparing what choice is obtained with those methods vs cross-validation).

We have actually tested them for the Principal Components Regression because they only are specific to this method. Results show that cross-validation gives better results but we decided not to show it in the manuscript as it was already quite dense. Nevertheless, we agree that it should be shown or mentioned. Thus, we have added a figure and a supplementary table in order to show that the use of cross validation provides better results than previous methods (only for PCR).

2.2 Gallicisms

The manuscript is generally well organized, but the writing suffers from many gallicisms. Since I happen to know a little French, here is an attempt at translating them:

- page 6, line 11: facilitate → simplify
- page 8, line 11: most performant → best-performing
- page 11, line 16: inversed → inverted
- page 12, line 23: to present frequently a → to often result in a • page 15, line 15: require to be tuned → require tuning

We thank the reviewer for these corrections that have been added in the manuscript.

2.3 Unavailability

I understand the need to protect data and code until the paper is published. How- ever, acting like they are public, and linking to a non-functional Zenodo link (https: //zenodo.org/record/1403146#.W4UMUGaB2qA) is bad form. Either give a complete link or mention that the data/code will be shared upon publication.

When we submitted the paper we tested the Zenodo link, and it worked well. We figured out, thanks to this comment, that it is now broken, and we do not know since when. We did not mean to protect our code nor our data and we actually are glad to share it as we have worked hard to build it. Codes and data can now be found on the following GitHub link: https://github.com/SimMiche/CLIMOREC