Final author response for the manuscript gmd-2018-211: "Reconstructing climatic modes of variability from proxy records: sensitivity to the methodological approach" by Michel et al.

This final 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

- **Response to Anonymous Referee #1:**

(1) This paper presents new reconstructions methods and applies them to reconstruct the NAO using data primarily from the PAGES2k database. I think this is a good study that introduces some potentially useful new paleoclimate reconstruction methodologies.
(2) (3) We thank the reviewer for this overall positive evaluation of our work.

I have a number of comments, corrections, and requests for clarification below:

(1) p.1 l.7-9, p.4 l.18, p.20 l.10 These statements are too strongly worded. Not every mode of variability is reconstructable, some occur on too short of time scales to be captured in the paleoclimate record (e.g., monthly time scales) and some modes are in locations where there are poor covariances with available proxy records (e.g., the Southern Ocean).

(2) We agree with the reviewer that this claim was too strong.

(3) This statement is modified in the corrected manuscript to clarify that our method is not able to reconstruct every climate index but only the ones for which sufficient covariances between large-scale modes and proxy records are found and for which proxy records exhibit fine enough time resolution to resolve the main time scale of the considered variability mode. Furthermore, we will also highlight that our approach can be used to reconstruct other kind of climate variable time-series such as temperatures or precipitations for a given location.

(1) p.2 l.9-11 This sentence is unclearly worded, for example, "non-stationary variability" doesn't "ask" questions, people ask questions.

(2) We agree with the reviewer on this statement.

(3) We replaced "asks the questions of" by "highlights".

(1) Introduction: In general, the introduction takes a long time to get to the main points of the study. The authors might consider revising the introduction to cut down the length.

(2) (3) The introduction has been largely cut down by only keeping the most important informations relative to the topic of the manuscript.

(1) p.5 l.4-5 Linear interpolation of low resolution proxies artificially increases the influence of these records and introduces spectral artifacts in the proxy time series (e.g., Hanhijarvi, Tingley, Korhola 2013, doi: 10.1007/s00382-013-1701-4). This process also ignores dating uncertainty in such low-resolution proxies, which can be a significant source of reconstruction error. Have you accounted for these factors, particularly the dating uncertainty? What is the influence of using only annually resolved data?

(2) Indeed, we found that using interpolated low resolution proxy records results in overestimating their weights in our reconstruction because of the falsely high correlations they have with the NAO index. This is largely due to their respective high auto-correlations at the annual time-scale. Hence, as mentioned by the reviewer, using this kind of proxy record indeed brings a lot of reconstruction errors due to overestimated weights, dating uncertainties, but also, because they induce erroneous validation scores as the link between these proxy records and the NAO index is overestimated. Concerning the dating uncertainty, it is also present in annually-resolved proxy records and this aspect is not accounted for in the present version of the code.

(3) Following this comment we have updated our code, manuscript and data with the use of the 2017 version of the Pages 2k database as suggested by Reviewer 2. Then, using this new proxy database, and in order to address this comment, we decided to remove the proxy records that are not annually resolved. For dating uncertainties, this is certainly something to be considered in the next version of the code. We thus add a short discussion on this aspect in the discussion section, concerning potential outlooks for the next versions.

(1) Section 2.2 Do the methods estimate uncertainty in the reconstruction or just provide a single reconstruction? Are the ensembles of reconstructions discussed elsewhere a kind of uncertainty estimate of the mean reconstruction? These, or something like them, would be essential to use and display because without reliable uncertainty estimates, paleoclimate reconstructions are not useful.

(2) This was actually a major omission in the former version of the paper and we thank the reviewer to report it. The uncertainties we now provide are calculated as in Ortega et al. (2015) using the residuals calculated over the 50 training periods. These 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.

(3) 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) p.7 l.16-19 Using correlation as the only validation metric is problematic, especially when it comes to comparing reconstruction methodologies. You really must include additional metrics that account not just for the correlation, but the variance and bias as well. If the approaches provide uncertainty estimates, then the skill metrics need to also account for those (using, for example, the continuous ranked probability score).

(2) This comment was also highlighted by the other reviewer as well as in the short comment of Eduardo Zorita. We totally agree with this comment and we decided to add both the root mean squared errors and the Nash-Sutcliffe Coefficient of Efficiency (NSCE) as additional metrics. The NSCE calculates the ratio of the averaged quadratic distance between the reconstruction and the observations and the quadratic distance between the mean of the observations and the observations. This metric, defined between $-\infty$ and 1 indicates that the reconstruction is robust when NSCE>0. Otherwise, lower values mean that using the mean of the testing series is more robust than performing a reconstruction using the statistical model.

We thus believe that these two metrics adequately account for the bias and variance in the reconstruction, which should then improve the conservation of these properties in our reconstruction.

(3) The whole new manuscript now accounts for these two metrics and use the NSCE as the main decision metric.

(1) p.16 l.19-20 This statement is incorrect. Previous reconstructions almost never overlook this issue, but rather proxy network selection is integral to the reconstruction process. It is

very rare to have a reconstruction approach, especially one that is regression-based, that does not remove proxies because of insufficient correlation with the target climate variable.

(2) For climate index reconstructions we found at least two major studies that have not used proxy network selection to perform their reconstruction : Cook et al 2002 (NAO reconstruction) and Wang et al 2017 (AMV reconstruction).

(3) Nevertheless, we indeed found that these studies are particular cases and we modified this statement to clarify that we were referring mainly to these two studies.

(1) p.18 l.1-2 Or the "significant" correlation with the NAO could be spurious. Also note that non-stationarity violates one of the fundamental assumptions of these (and nearly all) reconstruction approaches.

(2) Indeed, we also ask ourselves if the significant correlations we found could be spurious but it is relatively difficult to determine whether they are or not. An indirect way to "verify" this significance of correlation is the location of the proxy records that have high correlations with the NAO. A way to rule out spurious correlation is the use of pseudo-proxies like in Ortega et al. (2015), but handling pseudo-proxies from different datasets was an arduous task for this multimethod paper. Nevertheless, the fact that most proxy records selected for the highest levels of correlation significance (i.e. Greenland, Arctic Canada, North America and Europe. See Fig. 6 in the last version of the manuscript) are located in the centers of action of the NAO (which has not been imposed *a priori*) (e.g. Casado et al. 2013) is a good indicator that most proxy records won't be spurious NAO predictors. The second comment about non-stationarity indeed highlights a problem that not only questions our study, but also all of the proxy based reconstructions studies.

(3) In the new version of the manuscript we remove the sentence concerning non-stationarity since this type of caveat has to be included in the discussion section. We also highlight that the location of most of the proxy records selected shows that our method seems to adequately select reliable predictors.

(1) p.19 l.12-15 I think this statement is too strongly worded given that you've only validated the reconstructions using correlation and haven't validated reconstruction uncertainties. How do the reconstructions compare given the uncertainties?

(2) (3) As mentioned above, in the revised version we use the coefficient of efficiency to validate our reconstructions and we include and discuss regression uncertainties in our main text and dedicated figures.

- **Response to Anonymous Referee #2:**

1 Scientific Comments

(1) 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.

(2) (3) 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 set of different alternative reconstructions of a given climate index for a given proxy record database. In addition, the code provides an ensemble of scores that evaluate the different reconstructions, each produced with different methodological choices. Thus the user of CliMoRec (see "Response to the short comment from Astrid Kerkweg: 'Executive Editor comment on gmd-2018-21' ") can finally 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 that CliMoRec 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, an information that is 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"

(1) 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.

(2) (3) 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.7 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.

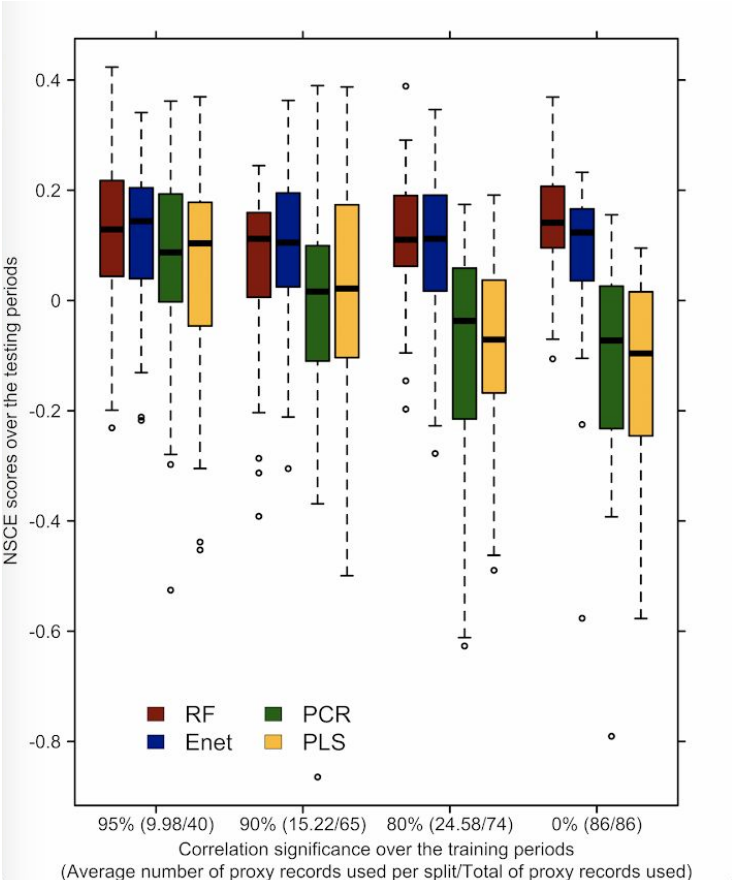

Fig. R1: Nash-Sutcliffe Coefficient Efficiency (NSCE) scores obtained for each method for the reconstruction period 1000-1970 and for different significance for the correlation test performed on the training periods: 95%, 90%, 80% and 0%. Red boxplots give the NSCE scores for the Random Forest method. 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

(1) 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.

(2) 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. 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 our approach and the approaches mentioned by the reviewer 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 implement and thus not accessible to all paleoclimatologists, necessarily discard other more simple statistical models. We believe that science can benefit from a variety of approaches, all together contributing to identify robust results.

(3) 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?

(1) 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).

(2) 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 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).

(3) We added regression uncertainties in a table and on the figures where the reconstructions are shown (Fig. 11 of the last version of the manuscript). Also, the code we deliver provide standard errors for each member of a given final reconstruction.

1.4 Statistical Models are Models too

(1) 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.

(2) 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.

(3) We have revised the manuscript so as to avoid such shortcuts following proposition described above.

1.5 Perfunctory Validation

(1) 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).

(2) (3) 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. We thank the reviewer to suggest this more sophisticated metrics that have been 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%).

(1) 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).

(2) (3) 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.

(1) 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.

(2) We agree that this is an important assumption to check.

(3) 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

(1) 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.

(2) 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.

(3) 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.

(1) Why use the PAGES2k version 1, and not PAGES 2k version 2 (PAGES 2k Consortium, 2017)?

(2) Pages 2k version 2 was not available when we started this study.

(3) We thank the reviewer for highlighting the updated version, which is now used in the new version of the manuscript.

(1) 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.

(2) 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.

(3) 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

(1) 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.

(2) We have worked hard for improving the language in this revised version and a native english colleague has agreed to review it before submission.

(3) 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

(1) 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.

(2) 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.

(3) 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).

1) 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.

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

(1) 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).

(2) 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.

(3) Thus, we have added a supplementary 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

(1) 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

(2) (3) We thank the reviewer for these corrections that have been added in the manuscript.

2.3 Unavailability

(1) I understand the need to protect data and code until the paper is published. However, 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.

(2) (3) When we submitted the paper we tested the Zenodo link (as advised by GMD), and it worked well. We figured out, thanks to this comment, that it is now broken, and we do not know since when neither why. 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

- **Response to the short comment from Eduardo Zorita: "Reconstruction Variance?"**

(1) The study uses one metric to evaluate the quality of the reconstruction methods : the correlation between observed and reconstructed index over a test period. However, other properties of the reconstructed indices may also be relevant, for instance, the variance. Many regression-based reconstruction methods underestimate past variability. This can be illustrated in a simple one-dimensional set up. Considering one proxy record P that reacts to variations of the NAO index:

$P(t) = \alpha NAO(t) + \varepsilon(t)$

where ε is random noise.

A simple, but widely used, reconstruction method is the statistical regression model:

$\hat{NAO}(t) = \beta P(t) + \eta(t)$

where η represents the variability not captured by the regression model. Using Ordinary Least Squares regression to estimate β leads to underestimation of the true value of β and, therefore, of the true NAO variance (see for instance Isobe et al 1990 Linear regression in astronomy for a review of different regression flavours and their properties).

This problem may or not be present in the methods used in this study. It would be useful if the authors could report in Table 4 also the variance of the reconstructed NAO index in the test period wrt. to the observations and also the variance of the reconstructed index over the full period.

(2) We thank Eduardo Zorita for this constructive and useful comment.

(3) We decided to add a table showing the variance for the best reconstructions from each method (i.e. the reconstructions presented in figure 11). The variance of the reconstructions is presented for the whole instrumental period, the testing period, the training period, the full reconstruction period and its portion before instrumental observations of the Jones et al. (1997) NAO index (the years before 1856 being excluded). We also add discussions in the main text of the manuscript about this well-known problem in paleoclimate reconstructions.

(1) Also, it would be informative if the time series in figure 11 were not normalized to unit variance (?), but showed the actual reconstructed variability.

(2) Normalizing to unit variance is a useful way to easily quantify NAO variability using standard deviations as unit. Nonetheless, as Eduardo Zorita is mentioning, it is actually hiding important informations about the reconstruction we performed.

(3) Thus, we decided to modify figure 11 in order to keep the actual reconstructed variability by our code. +1 and -1 standard deviation levels for each reconstruction have also been added in this figure in addition of their regression uncertainties (see response to 1.3 comment of "Response to Anonymous Referee #2").

- **'Response to the short comment from Astrid Kerkweg: "Executive Editor comment on gmd-2018-211"'**

(1) Dear authors,
in my role as Executive editor of GMD, I would like to bring to your attention our Editorial version 1.1:
http://www.geosci-model-dev.net/8/3487/2015/gmd-8-3487-2015.html

This highlights some requirements of papers published in GMD, which is also available on the GMD website in the 'Manuscript Types' section:
http://www.geoscientific-model-development.net/submission/manuscript_types.html

In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:
• "The main paper must give the model name and version number (or other unique identifier) in the title."

In order to simplify reference to your developments, please add the name of your software tool (e.g., "statistical toolbox") and its version number in the title of your article in your revised submission to GMD. The title could be something like "Reconstructing climatic modes of variability from proxy records using the statistical toolbox version 1.0: sensitivity to the methodological approach"

(2) We thank the executive Editor Astrid Kerkweg for reminding us the guideline for submission.

(3) We decided to attribute the name CliMoRec (Climate Mode Reconstruction) to our statistical toolbox and we have changed the name of the manuscript to: "Reconstructing

climatic modes of variability from proxy records using CliMoRec version 1.0: sensitivity to the methodological approach". Also, we have modified the references "statistical toolbox" to "CliMoRec version 1.0" in the main text of the manuscript.

Reconstructing climatic modes of variability from proxy records ~~: sensitivity to the methodological approach~~using CliMoRec version 1.0

S. Michel[1], D. Swingedouw[1], M. Chavent[2], P. Ortega[3], J. Mignot[4], M. Khodri[4]

26 avril 2019

1 : Environnements et Paleoenvironnements Oceaniques et Continentaux (EPOC), UMR CNRS 5805 EPOC-OASU-Universite de Bordeaux, Allee Geoffroy Saint-Hilaire, Pessac 33615, France.

2 : Institut National de la Recherche en Informatique et Automatique (INRIA), CQFD, F-33400 Talence, France.

3 : BSC, Barcelona, Spain.

4 : Sorbonne Universites (UPMC, Univ. Paris 06)-CNRS-IRD-MNHN, LOCEAN Laboratory, 4 place Jussieu, F-75005 Paris, France.

Abstract

Modes of climate variability strongly impact our climate and thus human society. Nevertheless, ~~their statistical properties~~ the statistical properties of these modes remain poorly known due to the short time frame of instrumental measurements. Reconstructing these modes further back in time using statistical learning methods applied to proxy records is ~~a useful way to improve~~ useful for improving our understanding of their behaviours~~and meteorological impacts~~. For doing so, several statistical ~~reconstruction~~ methods exist, among which the Principal Component Regression is one of the most widely used ~~. Additional predictive, and then reconstructive, statistical methods have been developed recently, following the advent of big data.~~ in paleoclimatology. Here, we provide to the climate community ~~a multi-statistical toolbox~~the computer device CliMoRec, based on four ~~statistical learning methods and~~ regression methods (PCR, Partial Least Squares, Elastic Net and Random Forest) and cross validation algorithms, that enables systematic reconstruction of ~~any climate mode of variability as long as~~ a given climate mode index. A prerequisite is that there are proxy records in the database that overlap in time with ~~the observed variations of the considered mode. The~~ its observed timeseries. The relative efficiency of the methods can vary, ~~depending on~~ according to the statistical properties of the mode and the ~~learning set~~proxy records used, thereby allowing to assess sensitivity related to the reconstruction ~~techniques. This toolbox is modular in the sense that~~ technique. CliMoRec is modular as it allows different inputs like the proxy database or the ~~chosen variability mode~~regression method. As an example, ~~the toolbox~~ it is here applied to the reconstruction of the North Atlantic Oscillation by using Pages 2K database. In order to identify the most reliable reconstruction among those given by the different methods, we ~~also~~ use the modularity of CliMoRec to investigate the sensitivity to the methodological setup to other properties such as the number and the nature of the proxy records used as predictors or the reconstruction period targeted. The best reconstruction of the NAO that we ~~thus obtain~~ obtain is using the Random Forest approach. It shows significant correlation with former reconstructions, but exhibits better validation scores. ~~.~~

1 Introduction

~~The climate system is composed of interdependent subsystems~~The interdependent components of the climate system, such as the atmosphere ~~that can vary at relatively fast timescales and is more chaotic as compared to the oceanor the cryosphere. As a result of the~~ and the ocean, vary at different timescales. The interactions between those components [Mitchell et al., 1966] ~~, the climate variability spectra is very large and ranges from hourly to~~ lead the climate to vary from the hourly to the multidecadal timescales. ~~In the absence of any modulations of the external forcings, such variability is still present, as evidenced in preindustrial control simulations with~~ Preindustrial control simulations of global coupled climate models ~~. This variability is~~ have evidenced that such a variability is still present without any modulation of the external forcings, which is frequently referred to as internal variability [Hawkins and Sutton, 2009]. ~~The variations and dynamics of the climatic system are also influenced by external~~ External factors such as volcanic aerosols [Mignot et al., 2011 ; Swingedouw et al., 2015 ; Khodri et al., 2017], ~~solar irradiance ; Seidenglanz et al.,~~ anthropogenic aerosols [Evan et al., 2009 ; Evan et al., 2011 ; Booth et al., 2012], solar irradiance [Swingedouw et al., 2011 ; Seidenglanz et al., 2012], and greenhouse gas concentrations [Stocker et al., 2013], ~~which alter the Earth ?~~also influence the variations and dynamics of the climate system by altering the Earth's radiation balance~~, and hence, deflect the mean climate state~~. By only considering the impact of external forcings which are not due ~~the of the~~ to the human activity, ~~one explores~~ we can characterise the so-called natural climate variability.

~~An unequivocal~~

An unequivocal synchronous rise in both the greenhouse gas ~~composition~~ concentration in the atmosphere and the global mean temperature has been observed in instrumental measurements [~~Bradley, 2003 ;~~ ~~?~~Stocker et al., 2013]. However ~~, the nonstationary variability~~ for temperatures, fluctuations around this trend from a decade to another [Kosaka and Xie, 2013 ; Santer et al., 2014 ; Swingedouw et al., 2017] ~~, asks the question about the relative role of anthropogenic forcing relatively to that of~~ highlight the modulating role of natural variability at decadal to multidecadal ~~climate variations~~scales. Thereby, improving our knowledge about past natural climate variability ~~should allow improving our knowledge and better evaluate the~~ and its sources is essential to better understand the potential coming changes in climate~~in the near term future (decades, e. g. Hawkins and Sutton (Hawkins and Sutton)).~~ .

~~The physics~~

Physics driving the climate system induces large-scale variations, organised around recurring climate patterns with specific regional impacts and temporal properties. These variations are known as climate modes of variability~~, and their~~ . Their evolution is usually quantified by an index that can be calculated from a specific observed climate variable. These indices provide an evaluation of the corresponding climate variations and their regional impacts [~~?~~ Hurrell, 1995 ; Neelin et al., 1998 ; ~~?~~Trenberth and Shea, 2006]. As an example, the North Atlantic Oscillation (NAO), is the leading mode of atmospheric variability in the North Atlantic basin [~~?~~Hurrell et al., 2003]. Generally defined as the sea level pressure (SLP) gradient between the Azores high and the Icelandic low, the NAO describes large-scale changes ~~on~~ of winter atmospheric circulation in the northern hemisphere and controls the strength and direction of westerly winds and storm tracks across the Atlantic [~~?~~Hurrell, 1995]. A stronger than normal SLP gradient between the two centers of action induces a northward shift of the eddy-driven jet-stream. Such large scale changes in atmospheric circulation lead to precipitation and temperature variations in various regions (North Africa, Eurasia, North America and Greenland [~~?~~Casado et al., 2013]). Moreover, these meteorological impacts have major influences on many ecological processes, including marine biology [Drinkwater et al., 2003] as well as terrestrial ecosystems [Mysterud et al., 2001]. This mode also affects the oceanic convection in the Labrador

Sea and the Greenland-Iceland-Norwegian Seas through changes in atmospheric heat, freshwater and momentum fluxes [Dickson et al., 1996 ; Visbeck et al., 2003]. These changes may lead in turn to modifications in the Atlantic Meridional Overturning Circulation (AMOC) which ~~in turn affects~~ then affect the poleward heat transport and the related ~~SST~~ Sea Surface Temperatures (SST) pattern over the Atlantic [Trenberth and Fasullo, 2017].

The dynamics of these modes are still not fully understood due to the ~~small~~ relatively short duration of the instrumental records, which ~~is preventing~~ prevents robust statistical evaluation of their properties (e.g. spectrum, stability of teleconnections, underlying mechanisms~~...~~). To partly overcome this limitation, ~~numerous studies have reconstructed climate variations well~~ reconstructions of climate beyond the period of ~~climate measurements (since around 1870), based on proxy recordsCook et al., 2002 ; Mann et al., 2009 ; Ortega et al., 2015 ; Luterbacher et al., 2016 ; Wang et al., 2017~~direct measurements have been performed in numerous studies that combine appropriate statistical methods and information from proxy records. Proxy records provide indirect ~~observations of~~ estimates of past local or regional climate~~in the past, using~~ , derived from natural archives coming for instance from sediment cores, speleothems, ice cores or tree rings. ~~The different records have their own characteristics and limitations, which need to be considered when combined together to perform the reconstructions. For example~~According to its nature, each proxy record has a specific temporal resolution, from years to millennia, and ~~then covers~~ can cover a specific period: from hundreds to millions of years. New proxy records are continuously gathered extending the available datasets and allowing paleoclimatologists to build increasingly consistent reconstructions [Pages 2K Consortium, 2013 ; Pages 2K Consortium, 2017]. ~~The last millennium is a period extensively investigated as it contains the densest network of high-resolution proxy records. Mann et al., 2009 ; Luterbacher et al., 2016.~~

~~The last millenium is of a great interest to put in perspective and understand the recent climate variations. Indeed, before the early 19th century, the anthropogenic radiative forcing was negligible Hegerl et al., 2007 ; Hawkins et al., 2017. Moreover, proxy records reveal two contrasting climatic periods during that millennium, as identified by Lamb (1965). These periods are known as the Medieval Climate Anomaly (MCA) and the Little Ice Age (LIA) Mann et al., 2009, which correspond to an anomalously warm and cool period of north hemispheric mean temperature, respectively. Modes of climate variability can have diverse worldwide impacts (usually known as climate fingerprints), which can be recorded by different proxy time series. This can be thus combined to make reconstructions of their variability. The selected proxy records need to cover, at least partially, the observational period. That is an important requisite to make a robust calibration. Based on this assumption, several~~

Based on the assumption that climate modes such as the NAO affect climate conditions in different locations, some studies have used ~~statistical predictive methods to reconstruct different climatic modes on longer timescales Cook et al., 2002 ; Gray et al., 2004 ; Ortega et al., 2015 ; Wang et al., 2017. For instance, for the NAO, Cook et al. (2011~~regression-based methods on temperature and drought-sensitive proxy records to reconstruct the variability of these modes over the last thousand years. Cook et al. (2002) firstly proposed a complete methodology of nested Principal Component Regressions (PCRs) [Hotelling, 1957] using annually resolved proxy records bounding the North Atlantic to reconstruct ~~its~~ the NAO variability further back to 1400. ~~Several new proxy records have been documented since this date Pages 2K Consortium, 2017and the NAO reconstruction could probably be largely improved if it was updated to include these new data.~~ More recently, ~~Mignot et al. (2011~~Ortega et al. (2015) performed a NAO reconstruction from 1073 to 1969, also based on the PCR, using 48 proxy records that were significantly correlated with the ~~the~~ historical NAO index on their common time window. Instead of nesting reconstructions of different sizes, which ~~leads~~ can

lead to inhomogeneities between time windows using different proxy selections, this study used several random calibration/validation samplings of the overlap period of the NAO index and the proxy records to perform individual reconstructions on the same time frame. By repeating numerous times that sampling, several reconstructions were obtained through the different PCR results. This ensemble approach brings two advantages. The first is that since validation/calibration periods are not fixed, the validation/calibration skills do not depend on the particular way these periods are split. The second advantage is that the different reconstructions obtained can be aggregated by averaging each of them to isolate the coherent features among them. The standard deviation between the individual reconstructions is thereby reduced, as only the most emergent patterns are kept. Such kind of ensemble reconstruction, using nested PCR as in Cook et al. (2002), have been recently made by Wang et al. (2017), but for reconstructing the Atlantic Multidecadal Variability (AMV), a climate variability index characterising large-scale variations in North Atlantic SST Regression-based methods have also been used for reconstructing other climate modes indices than NAO, such as for instance El-Niño Southern Oscillation index [Trenberth and Shea, 2006Li et al., 2013] . The recent increasing amount of data is not specific to the paleoclimatology field. Indeed, since the past four decades, the advent of internet and technological innovation has allowed to store and manage exponentially growing data from various sources and the Atlantic Multidecadal Variability index [Wang et al.Gray et al., 2004 ; Wang et al., 2017]. Hence, the beneficial capacity of decision making through data analysis in several fields has been largely developed, using many predictive algorithms for all kind of data Tibshirani, 1996 ; Breiman, 2001 ; Zou and Hastie, 2005. That field of science, often referred as "big data", is based on several statistical and probability theories and is named Statistical Learning or Machine Learning which is a subpart of Artificial Intelligence Vapnik, 2000 ; Breiman, 2001 ; Zou and Hastie, 2005. Combined with cross validation algorithms, the PCR is one of the most efficient statistical learning regression methods Hotelling, 1957. It is still considered as a performant method in many fields, such as paleoclimatology. However, more recent algorithms provide alternative

More recent algorithms than PCR provide alternative regression methods that can also be used to reconstruct climate modes, and may possibly further improve the quality and the robustness of these reconstructions. In this paper, we provide a toolbox, using present the computer device CliMoRec (Climate Mode Reconstruction) version 1.0, which includes multiple statistical approaches, for reconstructing climate modes indices. It is based on four regression methods: the PCR, the Partial Least Squares regression (PLS), the Elastic-net regression (E-net) and the Random Forest (RF). The aim is to propose a systematic reconstruction approach through a computer device. This toolbox It communicates with a large proxy database. This database , that contains various types of proxy records distributed all over the Earth, and associated with worldwide and which are sensitive to different climate variables. Therefore, this toolbox allows reconstructing any climatic mode in the past CliMoRec is thus designed to reconstruct the past variability of different climate modes (Fig. 1). The confidence we have in the reconstruction is then evaluated through training-testing techniques. Some general statistical learning tools, such as the cross validation, It should be stressed that CliMoRec will only be useful with climate indices for which there are enough proxy records representing their regional climate imprints, and that have the appropriate time resolution to capture its preferred timescale of variability. Besides the climate modes, CliMoRec can also be used to reconstruct other kinds of climate time-series such as temperatures or precipitations in a given location.

In section 2, the database and some general statistical tools are first presented. The reconstruction methods, are then described in a mathematical formalism . We then compare in section 3. Section 4 compares these methods by reconstructing the NAO index over the last millenium . Finally, we investigate and investigates the reconstruction sensitivity to methodological choices such as the method used, the learning period, the proxy predictors selection and the size of the calibration samples. training samples. Final

section 5 presents a discussion including some outlooks for next version of CliMoRec and the conclusions of this study.

2 Data, notations and methodologies

2.1 Data

The assessment of our reconstruction techniques is investigated for the NAO index, as it is probably the mode of variability that has been observed for the longest time period. ~~Indeed, this index is~~ This index is indeed relatively simple to calculate from ~~instrumental records because it only needs two instrumental record locations for SLP~~SLP time series as it only requires two locations with instrumental records: one within the center of action of the Azores anticyclone (typically Gibraltar) and one within the Icelandic low ~~. Thus, because of this simplicity, the NAO index covers a longer instrumentally observed period than other indices.~~ (typically Reykjavik). The reference NAO index ~~calculated from SLP records in Gibraltar and Reykjavik starts in 1856 Jones et al., 1997. An extension to 1823 has been proposed , using new SLP series from Cadiz and San Fernando, approximately 100 kilometers from Gibraltar Vinther et al., 2003. That Gibraltar/Cadiz-Reykjavik index is chosen as our reference observational NAO index in this paper.~~ is then calculated as the normalized SLP difference between these two locations. Jones et al. (1997) have for example proposed an index spanning the whole period since 1856.

~~Our statistical toolbox is based on a set of proxy predictors essentially composed of the~~

In terms of proxies, we use the state-of-the-art Pages 2k ~~2014 version~~ database [~~Pages 2K Consortium, 2013~~Pages 2K Consortium ~~. However, some proxy records (*Arc*_38 to *Arc*5 9, following PAGES encoding)~~ in its latest 2017 version (P2k2017). The proxy records which resolutions are lower than annual have been removed~~because their resolution is longer than ten years, which may have an impact on the interpretation of annual to subdecadal climate processes in the reconstruction. All the proxy records with a greater than annual resolution are then linearly interpolated to that resolution~~. Even if these proxy records could be interpolated to a finer temporal scale and used for the reconstruction, their use is not recommended [Hanhijarvi et al., 2013], as the interpolated time series will present high auto-correlation coefficients, which could inflate the correlations with the NAO and thus their weight in the final reconstruction, potentially leading to spurious results. We also added ~~to this database 69~~ 44 annually-resolved proxy records used in ~~the Wang et al.~~ Ortega et al. (~~2017) and Ortega et al. et al. (2015)studies. All of the North American tree ring series in Pages 2K database have been truncated to 1200 as this is their oldest common year. 15 of these series extend further back in time and have been considered here in their full length. These series are encoded as *NAm – TR*_7, 13, 14, 15, 21, 28, 29, 30, 62, 76, 81, 109, 110, 127, 128 in the Pages 2K database 2014 version Pages2K2014. Thus, we~~ 2015) and not present in P2k2017 (see supplementary informations). We end up with ~~540 worldwide distributed proxy records, which can potentially allow to reconstruct any mode of variability. All of the proxy records which are not in the Pages 2K 2014 version are presented in the supplementary table 1. For the other proxy records, the reader can refer to the Pages 2K 2014 version database. We attribute an ID to each proxy records to make them recognizable by the users of the statistical tool (see supplementary table 1). Among the 540 proxy records, only those completely overlapping the reconstruction period are kept. The statistical tool that we propose adjust the proxy dataset depending on the reconstruction period targeted.~~ a database of 554 well-verified and worldwide distributed annually-resolved proxy records.

2.2 Methodology

The general reconstruction procedure follows ~~10 steps, all~~ 11 steps, among which the first three are inputs selection and the others are already implemented in ~~the statistical toolbox~~CliMoRec. These are applied sequentially as follows (Fig. 1):

1. An observational time series of the mode of variability is chosen to be used as the predictand
2. A target time period $\mathcal{T}$ for the reconstruction is selected
3. The statistical reconstruction method to be applied is selected
4. The proxy records that overlap with the selected reconstruction period are extracted to be used as predictors
5. The common period between the observed climate index and the selected proxy records is ~~extracted for fitting~~ identified and extracted for calibrating the reconstruction
6. This common period is randomly split in two, one for training the model (training period), and one for testing it (testing period). This is repeated R times to generate an ensemble of reconstructions
7. ~~For each member of the ensemble, the reconstruction~~ The proxy records that have a significant correlation at a given threshold with the climate index over the training period are selected to train the statistical model
8. Each of the R sets of periods and proxies is calibrated over the training ~~period~~ window for all the different statistical parameters ~~for a given method~~ of the given method selected in 3, and the best ~~one is identifiedd~~ performing set is identified
9. The corresponding optimal setup is then applied to extend the reconstruction over the target period ~~T~~ $\mathcal{T}$ for each ensemble member
10. ~~A validation~~ Avalidation score is computed for each member by comparing the ~~true~~ observation-based testing series and each individual reconstruction over the corresponding testing period
11. The final reconstruction is calculated as the average of all the individual R reconstructions

Thus ~~the toolbox~~ CliMoRec provides the mean reconstruction of the chosen mode with associated uncertainties and a vector with ~~en~~ an ensemble of R validation scores following different metrics as final outputs.

The number of proxy records and the reconstruction period are here fixed for the different training/-testing period sections, in contrast with some previous studies which used nested approaches [Cook et al., 2002 ; Wang et al., 2017]. ~~As the weight of each proxy record is unknown before performing the reconstruction, the nested approaches may attribute unrealistic weights to the proxy records that bear the longest temporal coverage. In addition, as we want to perform several reconstructions by changing the set of proxy records employed or the reconstruction period considered, using a nested approach would have a simultaneous impact on both factors, and may hinder the interpretation of the validation~~ We make this choice because the aim of this study is mainly focused on optimizing the methodological approach for the reconstruction and not the reconstruction itself. Nevertheless, CliMoRec can be used to perform reconstructions on different time windows which can be then aggregated to perform a nested reconstruction, with associated scores.

2.3 Mathematical formalism of empirical data

To ~~facilitate~~ simplify the mathematical notation, we make the assumption that the proxy record selection and truncation to the their common time window with the climate index have already been made (see

section 2.2, steps 4 and 5). It is important that all proxy records are truncated on the same time window to make them mergeable in the same matrix. Each record has to cover at least the chosen reconstruction time window $\mathcal{T}$ (section 2.2, step 2). Following these steps, the proxy record matrix does not contain missing values.

Fig. 2 illustrates how the proxy data are organised in the input matrix X. We denote $X^1 = (X^1_t)_{t\in\mathcal{T}}, \dots, X^p = (X^p_t)_{t\in\mathcal{T}}$, where t stands for the time (with N annual time steps), and p is the number of proxy records on the same period $\mathcal{T}$. X is thus a $N \times p$ matrix ~~where all these vectors are merged~~grouping the individual records: $X = [X^1, \dots, X^p]$. $Y = (Y_t)_{t\in T}$ is the target ~~mode of variability~~climate index, defined on the historical time window T ~~, containing~~ called the learning period, that contains n annual time steps. The period where Y is not known is denoted τ, containing m annual time steps (Fig. 2). Thus $\mathcal{T} = T \cup \tau$ is the entire reconstruction period, which contains ~~N~~ $N = n + m$ annual time steps. With these notations, the dimensions of the different matrices and vectors are: $X \in \mathbb{R}^{N\times p}$; $X_{(T)} \in \mathbb{R}^{n\times p}$; $X_{(\tau)} \in \mathbb{R}^{m\times p}$; $Y \in \mathbb{R}^n$. The ~~period T, on which all the predictors and the predictand are known and the training/testing splits are performed, is called the learning period. The period $\mathcal{T} = T \cup \tau$, covered by the predictors, is called the reconstruction period. The learning set is then~~ learning set is denoted $\{X_{(T)}, Y\}$, and the reconstruction set is denoted $\{X_{(\mathcal{T})}\}$.

2.4 Terms~~and~~ , notations ~~of learning theory~~and validation metrics

To build and validate the reconstruction of Y, the dataset of predictors X is split in two independent subsets as shown in section 2.2, one for the training (usually called training set), and another on which the model is tested (called testing dataset or first seen data).

Building a model consists in estimating all the parameters needed to reconstruct Y given the predictors $X^1, \dots, X^p$. As an example, building a PCR model consists in determining the Principal Component of the predictor matrix X and finding the best linear combination of them to reconstruct Y over the training period. Then, the reconstruction consists in projecting the first seen data on the orthogonal basis built, and applying the estimated regression coefficients to reconstruct Y over the whole time window $\mathcal{T}$.

We denote the chosen reconstruction method by $\mathcal{M}$. Each method is defined by a specific number of parameters q, contained in the vector denoted θ. As an example, the Principal Component Regression has a single parameter that is the number of Principal Component used as regressor [Cook et al., 2002; Gray et al., 2004; Ortega et al., 2015; Wang et al., 2017]. We can denote the function $\mathcal{M}$ as a function of: (*i*) a set on which the model is built ($\{X, Y\}$), (*ii*) observations of the predictors on the reconstruction period ($X_{(rec)}$), and (*iii*) a parameter vector (θ):

$$\mathcal{M} : (\{X, Y\}, X_{(rec)}, \theta) \to \hat{Y}_\theta \tag{1}$$

$$(\{\mathbb{R}^{n\times p}, \mathbb{R}^n\}, \mathbb{R}^{m\times p}, \mathbb{R}^{q\,s}) \to \mathbb{R}^m \qquad n, p, m, q\,s \in \mathbb{N} \text{ (not fixed)} \tag{2}$$

Hence, the $\mathcal{M}$ function gives an entire reconstruction of size $m \in \mathbb{N}$, depending on θ for given training/-testing periods.

We introduce S as the score function, or validation metric. This function is an indicator that estimates the quality of a ~~prediction~~ reconstruction $\hat{Y}$ ~~in comparison~~ with respect to the observed values $Y_{(obs)}$:

$$S : (Y_{(obs)}, \hat{Y}) \to s \tag{3}$$

$$(\mathbb{R}^m, \mathbb{R}^m) \to \mathbb{R} \tag{4}$$

In this paper, three kind of validation metrics will be considered. The first is a correlation function, the second is a root mean squared error (RMSE) function and the third is a Nash-Sutcliffe coefficient of efficiency ~~:~~ [Nash and Sutcliffe, 1970]:

$$S_{cor}(Y_{(obs)}, \hat{Y}) = Cor(Y_{(obs)}, \hat{Y}) \tag{5}$$

$$S_{RMSE}(Y_{(obs)}, \hat{Y}) = \|Y_{(obs)} - \hat{Y}\| = \sqrt{\sum_{i=1}^{m}(Y_{i\,(obs)} - \hat{Y}_i)^2} \tag{6}$$

$$S_{NSCE}(Y_{(obs)}, \hat{Y}) = 1 - \frac{\sum_{i=1}^{m}(Y_{i\,(obs)} - \hat{Y}_i)^2}{\sum_{i=1}^{m}(Y_{i\,(obs)} - \bar{Y}_{(obs)})^2}, \text{ with} \bar{Y}_{(obs)} = \frac{1}{m}\sum_{i=1}^{m} Y_{i(obs)} \tag{7}$$

~~The first~~ S_{NSCE} will be used to validate the reconstruction methods over the testing period, and ~~the second~~ S_{RMSE} will allow to determine the optimal parameters ($\hat{\theta}\theta$) for the reconstruction over the training period. We use S_{cor} because it is used in the last NAO reconstruction of Ortega et al (2015), with which we will compare our results. S_{NSCE} is a metric defined between $-\infty$ and 1, values lower than 0 mean that using the mean over the training period is better than the proposed statistical model [Nash and Sutcliffe, 1970]. Here, we will consider that a final reconstruction is robust and reliable when its R NSCE scores are significantly positive at the 99% confidence level using a Student test. As the possible values of the NSCE score is not symmetric around 0, the best reconstruction is identified as the one that has the higher median of NSCE scores.

2.5 ~~Parameter tuning~~ Final reconstruction and ~~model comparison~~parameter tuning

2.5.1 ~~Parameter tuning by leave-one-out cross validation~~

We split the initial learning period T in R partitions of two subsets: $\left\{T^{(r)}_{(train)}, T^{(r)}_{(test)}\right\}, \forall 1 \leq r \leq R$. For a given method $\mathcal{M}$, R reconstructions are build on the R training samples. R is arbitrarily chosen, but larger R tends to produce reliable ensemble reconstruction by decreasing the variance of the R individual reconstructions made on the training samples. $\forall 1 \leq r \leq R$, we denote $\left\{X^{(r)}_{(train)}, Y^{(r)}_{(train)}\right\}$ the training set, and $\left\{X^{(r)}_{(test)}, Y^{(r)}_{(test)}\right\}$ the test set. At each step, the columns of X, $X_{(train)}$ and X_{test} are normalized to the mean and the standard deviation of the respective columns of $X_{(train)}$.

To estimate the optimal set of parameters θ_{opt} on a given training set $\{X_{train}, Y_{train}\}$, we use the K-fold cross validation approach (KFCV ; section 2.2, step ~~7 and~~ 8 and 9) [Stone, 1974 ; Geisser, 1975]. Cross Validation (CV) methods, are in general, widely used as parametrization and model validation techniques [Kohavi, 1995 ; Browne, 2000 ; Homrighausen and McDonald, 2014 ; Zhang and Yang, 2015]. ~~As presented in Fig. 3, the particularity of the LOOCV is that it uses a single observation for verification and the $n-1$ other observations as calibration set Stone. Here it is used to determine an empirically~~ Here, it is used as an optimization method to empirically determine an optimal set of parameters for θ. ~~$\forall 1 \leq i \leq n$, we denote $\{X_{(i)}; Y_{(i)}\}$~~ As presented in Fig.3, the KFCV splits the observations into a partition of n groups of same sizes (or approximately same sizes if the length of the training set is not divisible by K). $\forall 1 \leq k \leq K$, we denote $\{X_{(k)}, Y_{(k)}\}$, containing only information for the ~~i^{th} time step~~k^{th} drawn sample. Then, ~~$\{X_{(-i)}; Y_{(-i)}\}$~~ $\{X_{(-k)}, Y_{(-k)}\}$ is the set containing all the ~~initial observations, except the i^{th}~~K-1 other sets. For all possible values of θ contained in Θ, we scan the ~~n~~ K models based on the sets ~~$\{X_{(-i)}; Y_{(-i)}\}_{1 \leq i \leq n}$~~s $\{X_{(-k)}, Y_{(-k)}\}_{1 \leq k \leq K}$:

The empirical optimal set of parameters is obtained by minimizing the averaged S_{RMSE} functions on the ~~n splits regarding~~ K splits by considering all possible combinations of θ [Stone, 1974]. Mathematically, the optimal ~~LOOCV~~ KFCV set of parameters θ_{LOO} θ_{KF} is determined by:

$$\theta_{LOOKF} = \arg\min_{\theta\in\Theta} \frac{1}{n}\frac{1}{K}\sum_{i=1}^{}{}_{k=1}^{K} S_{RMSE}(Y_{(i)(k)}, \mathcal{M}(\left\{X_{(-i)(-k)}, Y_{(-i)(-k)}\right\}, X_{(i)(k)}, \theta)) \tag{8}$$

Using this approach, we retain the empirical estimation of the optimal set of parameters $\hat{\theta}_{opt} = \theta_{LOO}$ $\hat{\theta}_{opt} = \theta_{KF}$ for the given method $\mathcal{M}$ and a given learning set $\{X, Y\}$.

2.5.1 ~~Final reconstructions and validation correlations~~

~~In order to find the most performant method for a given dataset, we split the initial learning period T in R partitions of two subsets: $\{T^{(r)}_{(train)}, T^{(r)}_{(test)}\}$, $\forall 1 \leq r \leq R$. For all the methods, R reconstructions are build on the R training periods. R is arbitrarily chosen, but larger R tends to produce reliable ensemble reconstruction by decreasing the variance of the R individual reconstructions made on the training samples Browne, 2000. $\forall 1 \leq r \leq R$}, we denote $\{X^{(r)}_{(train)}, Y^{(r)}_{(train)}\}$ the training set, and $\{X^{(r)}_{(test)}, Y^{(r)}_{(test)}\}$ the test set.~~ In this study, KFCV method will be used on every training sets in order to perform each individual reconstructions according to the different training/testing splits.

~~LOOCV~~ KFCV is applied to build a unique optimized reconstruction for every training sets and any given method. Then, for all the corresponding and independent testing periods, the associated testing series ~~$Y^{(r)}_{(test)}$~~ $Y^{(r)}_{(test)}$ are compared to the individual reconstructions using the ~~S_{cor}~~ S_{NSCE} function. This way, R ~~validation correlations~~ NSCE scores are obtained for ~~the four methods~~ $\mathcal{M}$. In section 4, the distributions of the ~~validation correlations~~ NSCE scores will be used as a metric to compare different reconstructions. Fig.~~4 shows the whole procedure~~ 4 shows the the calculation to get the ~~validation correlation vectors~~ NSCE scores for a given method ~~M~~ $\mathcal{M}$.

3 ~~Statistical regression~~ Regression methods

We present each method in two steps: model fitting (for training) and reconstruction (for testing). ~~We also identify the number of parameters and their mathematical meaning.~~ For each method the proxy ~~predictor set~~ matrix is denoted as $X \in \mathbb{R}^{n\times p}$ the proxy predictor set and the ~~target~~ climate index as $Y \in \mathbb{R}^n$. ~~In this section,~~ $X_{(rec)} \in \mathbb{R}^{m\times p}$ is the testing dataset ~~on~~ from which a $\mathbb{R}^m$ reconstruction ~~vector is evaluated on the testing period. Y and each column of X are here normalized on their own time period.~~ is build using the regression method

3.1 Principal Component Regression (PCR)

3.1.1 Modeling

The Principal Component Regression [Hotelling, 1957] method consists in finding the best linear combination between Y and the Principal Component of X. The Principal Component Analysis (PCA) consists in applying an orthogonal transformation of an initial set of variables, potentially correlated between them, into another set of linearly uncorrelated variables: the Principal Component [Pearson, 1901; Hotelling, 1933].

The first step consists in building an orthogonal basis where X will be projected. We define $S \in \mathbb{R}^{p\times p}$, as the empirical estimator of the covariance matrix of X:

$$S = \frac{1}{n} X^T X \in \mathbb{R}^{p\times p} \tag{9}$$

~~The idea is to~~ We calculate the orthogonal basis formed by the vectors $v_1, \dots, v_p$ by diagonalizing S:

$$v_1 = \arg\max_{\substack{v\in\mathbb{R}^p \\ \|v\|=1}} v^T S v \tag{10}$$

$$v_2 = \arg\max_{\substack{v\in\mathbb{R}^p \\ \|v\|=1 \\ \langle v^T v_1\rangle=0}} v^T S v \tag{11}$$

$$\dots \tag{12}$$

$$v_p = \arg\max_{\substack{v\in\mathbb{R}^p \\ \|v\|=1 \\ \langle v^T v_1\rangle=0 \\ \dots \\ \langle v^T v_{p-1}\rangle=0}} v^T S v \tag{13}$$

$$\tag{14}$$

where $\|v\| = \sqrt{\sum_{j=1}^{p} (v^j)^2}$, $\forall v \in \mathbb{R}^p$. ~~This procedure~~ It is equivalent to maximizing step by step the empirical variance of the projection of X on each orthogonal axis. Indeed, $\forall v \in \mathbb{R}^p$:

$$v^T S v = \frac{1}{n-1} v^T X^T X v = \frac{1}{n-1} (Xv)^T (Xv) = Var_{emp}(Xv) \tag{15}$$

The vectors $(v_k)_{1\le k\le p}$ are called the Empirical Orthogonal Functions (EOFs). ~~Since the columns of X represent the proxy records, it means that each EOF, which~~ It corresponds to the eigenvectors of the covariance matrix ~~, contains a certain~~ and each contains a given part of the spatial variability of the ~~dataset. Hence, we attribute them~~ proxy dataset. We attribute them the eigenvalues $(\lambda_k)_{1\le k\le p}$, which corresponds to the initial variance of X translated by each orthogonal projection in the new basis:

$$\lambda_k = Var(Xv_k) = v_k^T S v_k \qquad \forall 1 \le k \le p \tag{16}$$

The Principal ~~Component~~ Components $(u_1, \dots, u_p)$ are then the projections of X on the EOFs. We denote $V = (v_1, \dots, v_p)$. We then calculate the Principal Component matrix $U = (u_1, \dots, u_p)$, defined as:

$$U = XV \in \mathbb{R}^{n\times p} \tag{17}$$

Now, we regress Y on the $q \le p$ (see subsection 3.1.3) first Principal Component. These q Principal Component are merged in a submatrix of U: $\mathcal{U} = (u_k)_{1\le k\le q}$. The model is given by:

$$Y = \mathcal{U}\beta + \epsilon \tag{18}$$

Where ϵ is a white noise vector of size n.

The best estimator for $\beta = (\beta_1, \dots, \beta_q)$, is given by the Ordinary Least Squares (OLS) estimator which minimizes $\|\hat{\epsilon}\| = \|Y - \hat{Y}\|$:

$$\hat{\beta}_{OLS} = (\mathcal{U}^T \mathcal{U})^{-1} \mathcal{U}^T Y \tag{19}$$

3.1.2 Reconstruction

~~Using the testing data~~ We project the testing matrix $X_{(rec)}$ ~~(see section 2.4), we project the former~~ on the pre-calculated orthogonal basis V:

$$U_{(rec)} = X_{(rec)}V \in \mathbb{R}^{m\times p} \tag{20}$$

We then obtain the ~~prediction~~ reconstruction by applying the estimated coefficient vector on the sub-matrix $\mathcal{U}_{(rec)} = (U^1_{(rec)}, \dots, U^q_{(rec)}) \in \mathbb{R}^{m\times q}$:

$$\hat{Y}_q = \mathcal{U}_{(rec)}\hat{\beta}_{OLS} \in \mathbb{R}^m \tag{21}$$

3.1.3 Parameters

Here, q is the unique tuning parameter. The choice of that parameter clearly affects the reconstruction and then the ~~validation NSCE~~ NSCE scores. Here the parameter vector θ is unidimensional and takes its values in the discrete set $\{i\}_{1\leq i\leq p}$.

~~To our knowledge, this is the first time that a PCR uses the KFCV method to tune the number of Principal Component used at each split in paleoclimatological reconstruction. Previous studies used different criteria to define the number q of Principal Component $U_1, \dots, U_q$ to be kept. For example, Gray et al. (2004) retained all Principal Component for which the cumulated eigenvalues weights just exceeds 66% of the initial variance. Wang et al. (2017), selected the q Principal Component for which $\lambda_k > 1, \forall k \in \{1, \dots, p\}$. Also, Ortega et al. (2015) used the Preisendorfer's rule N Preisendorfer, 1988. In our case, the use of KFCV as our parameter selection method is preferred, as it is also valid for the other reconstruction techniques.~~

3.2 ~~The~~ Partial Least Squares Regression

The ~~Principal Component Analysis~~ PCA keeps most of the initial variance in X in a lower number of vectors. ~~The major problem of the PCR in a predictive or reconstructive purpose, is that the~~ But EOFs $v_1, \dots, v_p$ are constructed without taking into account any information about the predictand Y. Another possible approach is thus to determine the orthogonal basis in which the empirical covariance between Y and the projection of X on that former is maximized. This is the Partial ~~Lest~~ Least Squares regression (PLSr) method [~~Zou and Hastie, 2005~~Wold et al., 1984].

The first latent variable (LV), denoted $\xi_1 = \sum_{j=1}^p v_{1,j}X^j = Xv_1$, where $X \in \mathbb{R}^{n\times p}$ and $v_1 \in \mathbb{R}^p$ is the linear combination of the initial variables $X^1, \dots, X^p$ such as:

$$v_1 = \arg\max_{\substack{u\in\mathbb{R}^p \\ \|v\|=1}} Cov(Y, Xv), \tag{22}$$

In a similar approach to the PCR, the second LV is $\xi_2 = \sum_{j=1}^p v_{2,j}X^j = Xv_2$, orthogonal to ξ_1, such as:

$$v_2 = \arg\max_{\substack{v\in\mathbb{R}^p \\ \|v\|=1 \\ \langle\xi^1, Xv\rangle=0}} Cov(Y, Xv) \tag{23}$$

And so on, until we have $r \leq p$ LVs. The LV matrix is denoted $\Xi = [\xi_1, \dots, \xi_p]$. Here, $v_1, \dots, v_p \in \mathbb{R}^p$, are analogous to the EOFs in PCA, and are here called loadings. The latent variables $\xi_1, \dots, \xi_r$ respectively correspond to the projection of X on the r loadings.

Finding the loadings is not as trivial as for PCR. ~~This is due to the fact that the~~ Indeed the empirical covariance matrix is not necessary definite positive and thus cannot be ~~inversed~~diagonalized. We solve this problem by using the algorithm 1 named PLS1. Analogously to the PCR, the method provides various alternative reconstructions depending on the value of r, which corresponds to the number of LVs kept as regressors.

Algorithm 1

procedure PLS1
  $X_0 \leftarrow X$
for $h = 1, \dots, r$
  $v_h \leftarrow \frac{X_{h-1}^T Y}{\|X_{h-1}^T Y\|^2}$
  $\xi_h \leftarrow X_{h-1} v_h$
  $X_h = X_{h-1} - \frac{\xi_h \xi_h^T}{\|\xi_h\|^2} X_{h-1}$ *(deflation phase)*
end procedure

Now we regress Y on the $r \leq p$ first LVs. These r LVs are merged in a submatrix of Ξ: $\Psi = (\xi_k)_{1 \leq k \leq r}$. The model is given by:

$$Y = \Psi\beta + \epsilon \tag{24}$$

Where ϵ is a white noise vector of size n.

The best estimator for $\beta = (\beta_1, \dots, \beta_q)$, is given by the Ordinary Least Squares (OLS) estimator which minimizes ~~$\|\hat{\epsilon}\| = \|Y - \hat{Y}_{qLOO}\|$~~$\|\hat{\epsilon}\| = \|Y - \hat{Y}_{qkf}\|$:

$$\hat{\beta}_{OLS} = (\Psi^T \Psi)^{-1} \Psi^T Y \tag{25}$$

3.2.1 Reconstruction

The ~~prediction~~ reconstruction is done in the same way as for PCR. Using the first seen data matrix $X_{(rec)}$ (section 2.4), we project the latter on the pre-calculated orthogonal basis V:

$$\Xi_{(val)(rec)} = X_{(val)(rec)} V \in \mathbb{R}^{m \times p} \tag{26}$$

~~We then obtain the prediction~~ The reconstruction is obtained by applying the estimated coefficient vector on the sub-matrix $\Psi_{(rec)} = (\xi^1_{(rec)}, \dots, \xi^r_{(rec)}) \in \mathbb{R}^{m \times r}$:

$$\hat{Y}_r = \Psi_{(rec)} \hat{\beta}_{OLS} \in \mathbb{R}^m \tag{27}$$

3.2.2 Parameters

For the PLSr method, r is the unique tuning parameter. Analogously to the Principal Component Analysis, the tuning of that latter is obtained by KFCV.

3.3 ~~The~~ Elastic Net regression

3.3.1 Modeling

Without using orthogonal transformation of the initial variables as in PCR and PLSr, the most simple predictive model is the multiple linear regression model:

$$Y = X^1\beta_1 + \cdots + X^p\beta_p + \epsilon \tag{28}$$

Where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $Cov(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$.

The ~~prediction~~ reconstruction of Y, given p proxy records $X^1, \ldots, X^p$ is obtained by the equation:

$$\hat{Y} = X^1\hat{\beta}_1 + \cdots + X^p\hat{\beta}_p \tag{29}$$

$\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ are the regression coefficients, which are obtained by the OLS predictor. However, this usual regression model is known to ~~present frequently a poor prediction~~ often result in a poor reconstruction accuracy due to the several assumptions made on the original data [Poole and O'Farrell, 1971], which are often not verified~~:~~, such as homoscedasticity and errors normality. Several studies developed regularized (or penalized) regression methods to overcome the OLS defaults. Here we focus on the Elastic Net regression [Zou and Hastie, 2005], which is a combination of the Ridge regression [Hoerl and Kennard, 1970] and the Lasso regression [Tibshirani, 1996]. All these methods have been developed to avoid the high variability of the OLS predictor when the number of predictors is relatively high. The Ridge regression shrinks towards zero the estimated coefficients associated to predictors unlinked to the predictand. No predictor selection is made by this method, but the shrunken estimated coefficients modulate the importance of these in the model. By contrast, the ~~lasso~~ Lasso also reduces the variability of the estimates, but in this case by shrinking to zero the estimated coefficients associated to unreliable variables. Hence, a selection is made by rejecting variables associated to coefficients shrunk to zero.

The idea of a regularized (or penalized) regression is to add a threshold constraint using the l_k norm of β: $\|\beta\|_k^k = \sum_{j=1}^k \left|\beta_j\right|^k$. With $k = 1$ in Lasso regression, and $k = 2$ in Ridge regression. The penalized loss functions are given by:

$$L^{ridge}(\beta) = \|Y - \sum_{j=1}^{p} \beta_j X^j\|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{30}$$

$$L^{lasso}(\beta) = \|Y - \sum_{j=1}^{p} \beta_j X^j\|^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| \tag{31}$$

$$L^{enet}(\beta) = \|Y - \sum_{j=1}^{p} \beta_j X^j\|^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{32}$$

λ_1 penalizes the sum of the absolute values of the regression coefficients while λ_2 penalizes their summed squares. Here, $\lambda_1, \lambda_2 > 0$.

Let $w = (w_j)_{1 \leq j \leq p} = (sgn(\beta_j))_{1 \leq j \leq p}$, where sgn is the sign function. The loss functions can then be

denoted as:

$$L^{ridge} = \|Y - X\beta\|^2 + \lambda_2 \beta^T \beta \tag{33}$$

$$L^{lasso} = \|Y - X\beta\|^2 + \lambda_1 w^T \beta \tag{34}$$

$$L^{enet} = \|Y - X\beta\|^2 + \lambda_1 w^T \beta + \lambda_2 \beta^T \beta \tag{35}$$

The estimated regression coefficients obtained by minimizing the Lasso and the Ridge loss functions are:

$$\hat{\beta}^{lasso} = (X^T X)^{-1}(X^T Y - \frac{\lambda_1}{2} w) \tag{36}$$

$$\hat{\beta}^{ridge} = (X^T X + \lambda_2 I)^{-1} X^T Y \tag{37}$$

The Elastic Net regression coefficients are then estimated by minimizing L^{enet}:

$$\hat{\beta}^{enet} = (X^T X + \lambda_2 I)^{-1}(X^T Y - \frac{\lambda_1}{2} w) \tag{38}$$

An alternative way to write this equation as a linear combination of $\hat{\beta}^{lasso}$ and $\hat{\beta}^{ridge}$ is:

$$\hat{\beta}^{enet} = (X^T X + (1-\alpha)\lambda I)^{-1}(X^T Y - \frac{\alpha\lambda}{2} w) \tag{39}$$

where $\alpha \in [0, 1]$. If $\alpha = 1$, a Rigde regression is applied, and if $\alpha = 0$, we apply a Lasso regression.

3.3.2 Reconstruction

The ~~prediction~~ reconstruction is obtained by applying the estimated regression coefficients $\hat{\beta}^{enet}$ on the validation variables $X^1_{val}, \ldots, X^p_{val}$:

$$\hat{Y}_{\lambda,\alpha} = \sum_{j=1}^{p} X^j_{(val)} \hat{\beta}^{enet}_j \tag{40}$$

3.3.3 Parameters

For Enet method, the tuning parameters are λ and α. The latter controls the relative balance between the ~~lasso and ridge~~ Lasso and Ridge regularization, while the former controls the overall intensity of regularization as λ_1 (resp. λ_2) in ~~lasso~~ Lasso (resp. ~~ridge~~ Ridge regularization). A high α suggests a dense model with many but small non-zero coefficients. A low α suggests a sparse model with many zero coefficients. In our case, since we want a general methodology performant for each random split, we apply two simultaneous KFCV to find the best estimated pair $(\hat{\lambda}, \hat{\alpha})$.

Since λ and α take respectively their values in the continuous sets $\mathbb{R}^p$ and $[0, 1]$, we have to discretize their respective intervals for the parameter estimation. The finer these discretizations are, the more reliable the parameters will be, ~~but the longer the required computational timewill be~~at the expanse of the computational time.

3.4 Random Forest regression

The random forest has been introduced by Breiman (2001) as a learning method for regression. The method relies on using randomization to minimize the ~~prediction~~ reconstruction uncertainty given by regression trees. Random forests encompass a large variety of regression methods [Breiman, 2001]. Here, we present the most classical kind of random forests known as random-input random forests [Breiman, 2001].

3.4.1 Modeling

First we have to define regression trees. We denote each set of predictand/predictors by $\{Y_i, X_i)_{1\leq i\leq n}\}$ where $X_i = (X_i^1, \ldots, X_i^p)$, is the ensemble of proxy records for the i^{th} time step, and Y_i the corresponding values of the climate index at the same time step, $\forall 1 \leq i \leq p$. All the observations, $\{Y_i, X_i)_{1\leq i\leq n}\}$, $\forall 1 \leq i \leq p$, are put on the root of the tree. The first step consists in cutting that root in two child nodes. A cut is defined as:

$$\left\{X^j \leq d\right\} \cup \left\{X^j \geq d\right\} \tag{41}$$

where $j = \{1, \ldots, p\}$ and $d \in \mathbb{R}$. Cutting a node with $\left\{X^j \leq d\right\} \cup \left\{X^j \geq d\right\}$ means that all observations with a j^{th} variable lower than d are placed in the left child node. Hence, all observations with a j^{th} variable greater than d are placed in the right child node. The method selects the best pair (j, d) which minimize a loss function. Here, we aim at minimizing the variance of the child nodes. The variance of a given node t is defined as:

$$\sum_{i:X_i\in t} (Y_i - \bar{Y}_t)^2 \tag{42}$$

where $\bar{Y}_t$ is the averaged Y_i in the node t.

The same procedure is then applied recursively to the next child nodes using the same variables until a certain stop criterion is reached. The procedure automatically stops if each node contains a unique observation. Hence, the maximal depth of a regression tree is $n - 1$. An illustration of such tree is presented in Fig. 5.

A random-input regression tree is used here. This is a particular case of regression trees, in which a set of $m < p$ variables is randomly preselected before applying the regression tree. A large number K of random-input trees is computed. For each tree, we randomly select $m < p$ variables with probability $\frac{1}{p}$ and we apply the method until it reaches its maximal depth.

3.4.2 Reconstruction

The ~~prediction~~ reconstruction is obtained by splitting each testing series in the different trees~~previously constructed~~. In each tree, the estimation attributed to an observation is the empirical average of Y inside the node where the corresponding observation ends up, given the cut made on the corresponding predictors. For each testing series, the K reconstructions are averaged to give the final ~~prediction.~~ reconstruction.

3.4.3 Parameters

A priori, this method requires the optimization of two parameters: the number of trees K and the number of variables selected for each tree m. In practice K does not require ~~to be tuned~~tuning, as long as the number of trees is sufficiently high given p, which guarantees convergent results for any value of m [Breiman, 2001]. m is then the only parameter to optimize. The KFCV is then applied on m with a high K (here set to 1000), to select empirically the most efficient model.

4 Results

4.1 Methodological sources of uncertainty in the reconstruction

We apply ~~the former methods~~ CliMoRec with the four methods presented above to the reconstruction of the NAO. In the following, each reconstruction is obtained by averaging $R = 50$ individual reconstructions performed for R training/testing random draws. ~~Validation scores (based on correlations over the testing periods)~~ NSCE scores are also produced, and stored in a vector of R elements. This vector will thus be used as a quality metric to characterize the methodological uncertainty in the reconstruction. The following actions were undertaken to minimize the reconstruction uncertainty, and estimate its sensitivity:

1. Pre-selecting the most relevant proxy records
2. Choosing the most appropriate training/testing window length
3. Selecting the best learning period

These three steps are described below, before assessing the reconstruction itself.

4.1.1 ~~Proxy pre-selection~~

~~In order to investigate the sensitivity related to the selection of the initial set of predictors, we set the reconstruction period to $\mathcal{T} = [\![1000, 1970]\!]$, and the learning period to $T = [\![1823, 1970]\!]$, with $n = 148$. In addition, the training window length is set at $n_{train} = 111$, which gives $n_{test} = 37$. Only 122 of the 540 proxy recors of the initial dataset are covering this reconstruction period. We~~

4.1.1 Proxy pre-selection over the training periods

Among the previous climate reconstruction studies, Ortega et al. (2015) have performed a proxy selection over the training periods at the 90% confidence level using the correlation test from McCarthy et al. (2015) while Cook et al. (2002) and Wang et al. (2017) have selected their proxies by focusing on the regions affected by the modes they respectively reconstructed. Here we run 4 ~~different reconstructions ,~~ reconstructions of $R = 50$ individual members for each method~~, each based on a different proxy group chosen according to a correlation significance test with the original NAO index on the period T. The first group contains all the available proxy records on the~~ . These reconstructions are respectively performed with different significance levels for the proxy selection by correlation over the training periods. These levels are 0% (which means that all the records are used at each training/testing split), 80%, 90% and 95%. The reconstructions are performed for the reconstruction period ~~$\mathcal{T}$ (122 proxy records). The three other groups respectively contain the proxy records significantly correlated with the NAO index at the confidence levels 80% (61 proxy records), 90% (35 proxy records) and 95% (18 proxy records). The proxy records, and their respective correlation significance level with the NAO index are presented in Fig. ??. Fig. 6 gives the validation scores related to each reconstruction and each proxy selection. First, it appears~~

that for each method, the validation scores are improved when we use the most significantly correlated $\mathcal{T} = [\![1000, 1970]\!]$ and the learning period $\mathcal{T} = [\![1856, 1970]\!]$ encompassing 110 available proxy records with the NAO index over the historical period (Fig. 6). In addition, not all the methods have the same sensitivity to the proxy pre-selection. Indeed, Enet, PCR and RF methods have better validation results than PLS when all of the available proxy records are used as predictors. $n = 115$. In this section the training periods length is set to 92 and testing periods length to 23.

Our results suggest that enhancing the spatial coverage of the proxy records is not a necessary condition to improve the reconstruction. Indeed, we showed that using the densest proxy network (i.e., all of the available proxy records on $\mathcal{T}$) does not lead to better validation scores, due to the noise introduced by predictors that covary weakly with the target index (Fig. ?? and 6). Among the previous reconstruction studies, this kind of investigation have often been overlooked at the expense of increasing the spatial density of the proxy

Fig. 6 shows that RF method, particularly useful for large datasets, is more efficient using the whole set of proxy records Cook et al., 2002 ; Gray et al., 2004 ; Wang et al., 2017. Ortega et al. (2015) already showed the advantage of subsampling the proxy records more significantly correlated (i. e. 90%) with with $med(S_{NSCE}) = 0.18$ (med is the NAO. The validation NSCE obtained in their study are weaker than those we obtained here by using PCR on the 35 proxy records significantly correlated with the NAO index at the 90% confidence level , from which 19 are the same in both studies. Here, the best score ($\bar{r} \simeq 0.46$ on average) are obtained for median function), even if using proxy records uncorrelated with the NAO or not located in regions affected by NAO variations. On the other hand, the PLS method when only the proxy records significantly correlated with the NAO index at the 95% confidence level are kept (16 proxy records) . These results are better than those obtained by Ortega et al. (2015), 3 other regression methods are more adapted when the finest proxy selection (95%) is applied, as highlighted by Ortega et al. (2015) for the calibration constrained reconstruction ($r_{val} \in [-0.14; 0.58]$; $\bar{r} \simeq 0.24$) as well as for the model constrained reconstructionPCR. Fig. 6 is also evidencing that the widely used PCR methods and PLS have to be employed cautiously with a statistically-based proxy selection over the training periods in further studies. Indeed the reconstructions performed with these methods are only significantly robust at the 99% confidence level (see section 2.4) by using the most constraining pre-selection of proxies. In addition, even their best NSCE scores (for 95%) are relatively weak, with their first quartile slightly under 0. On the opposite, for RF and Enet methods, the proxy selection is not affecting the statistical robustness of the reconstruction, with reconstructions significantly robust at the 99% confidence level ($r \in [0.14; 0.64]$; $\bar{r} = 0.42$) (see Ortega et al., 2015)see section 2.4) for every choice of proxy selection.

Overall, RF gives the best NSCE scores and also provides the best reconstruction. Nevertheless, it should be noted stressed that these results have been obtained for a particular length in the training/testing windows of (11192/3723). The sensitivity to this will be is assessed in the next section.

4.1.2 Sensitivity to the length of training and testing periods

To estimate the sensitivity of the reconstruction performance to the length of the training and the testing windowperiods, we set again the reconstruction period to $\mathcal{T} = \{1000, \ldots, 1970\}$, and the learning period to $\mathcal{T} = \{1823, \ldots, 1970\}$, with $n = 148$$\mathcal{T} = \{1856, \ldots, 1970\}$, with $n = 115$. Based on the findings of the previous sectionsubsection, we only keep the proxy records which are significantly correlated with the NAO index at the 95% confidence level (18 proxy records, see section 4.1.1 and Fig. ??over the training periods for PCR, PLS and Enet and we use the whole set of proxy records at each split for RF (110 records). We run R $R = 50$ reconstructions with different window sampling for each method by gradually increasing

the length of the training ~~window: from 5% to 95%~~ periods: from 30% to 90% of the initial size of the learning period, with a step of ~~5%. Fig. 7 shows the validation NSCE obtained for these simulations. Small training window, this may related to an overlook of the general information in the data, which translates into negative and non-significant validation NSCE (Fig. 7). On the contrary, using a very long training window gives very high validation NSCE close to 1, but it also give negative ones (Fig. 7), i.e. a very wide range of validation scores , suggesting that the testing period is too short to robustly validate the reconstruction.~~ 5%. Training periods lengths out of this interval gives extreme negative scores and have thus not been considered.

~~Between these two extremes~~

According to the NSCE metric we find a large ~~window where~~ range of splitting periods for which validation scores are relatively similar and significantly positive for RF and Enet (from around ~~30% to 70%) . To assess the best reconstruction, we search the score vector which has the most significant positive validation NSCE and which is the highest on average. Following this rule, the optimal window split is 70%~~ 35% to 85%) (Fig. 7). For PLS the only acceptable setup is obtained by using the split 80% of the total size of the learning period for the training (n_{train} ~~= 104;~~ n_{test} ~~= 44)for PLS and PCR . For RF, the optimal split is 45%~~ n_{train} = 92; n_{test} = 23). The only optimal window split is 70% of the total for the training for PCR (n_{train} ~~= 67;~~ n_{test} ~~= 81), while for Enet, the optimal split is 65%~~ (n_{train} ~~= 96;~~ n_{test} ~~= 52) . Overall method which gives the highest validation NSCE on average is the PLS, closely followed by PCR and Enet~~ (n_{train} = 80; n_{test} = 35) (Fig. 7). ~~We now address the degree of uncertainty associated to the way the training/testing windows are partitioned. Fig. ?? shows the correlation between the reconstructions in the optimal window split (identified above), and the other alternative partitions. All correlation values thus obtained are particularly high, specially for training windows length representing at least a 45% of the total period, for which correlations are greater than 0.96, regardless of the method , except RF, for a training window length of correlations of 85% of the length of the initial periods . This suggests that the choice of the training period is not an important methodological source of uncertainty for the reconstruction .~~ Here, we have shown again that classical regression methods such as PLS and PCR are not producing the best reconstructions of the NAO. For this set of reconstructions, the method which gives the highest NSCE scores and provide the best reconstruction is again RF (Fig. 7).

4.1.3 Sensitivity to the reconstruction period

In this section, we ~~focus on the most efficient method (PLS) with~~ keep for each method the optimal selection of proxy records over the training periods (see section 4.1.1) and the optimal training/testing windows length (n_{train} ~~= 104,~~ n_{test} ~~= 44,~~ see section 4.1.2)~~and we~~ . We explore the impact of the reconstruction period~~, and hence, the learning period and the proxy set. Changing this period~~ . This affects the final reconstruction in two different ways, both related to the final proxy selection. ~~First, by modifying~~ Firstly it modifies the initial set of proxy records considered (as they need to cover the whole reconstruction period). Secondly, ~~by changing~~ it changes the period of overlap with the observations ~~, which lead to different correlations between the proxy records and the NAO index, which would affect their significances and therefore the final proxy selection. Indeed, a proxy record significantly correlated with the NAO index at a given confidence level on a given time window, can be non-significantly correlated with the NAO index with the same confidence level, but on another time window. This may be induced by physical processes that modifies the stationarity of the NAO and its teleconnections.~~ and then the randomly drawn training/testing splits.

We run the reconstruction ~~on 36~~ for 31 periods $\mathcal{T}$: from ~~1000-1965~~ 1000-1970 to 1000-2000, with an

increment of one year. By doing so, the number of available proxy records is not the same for each of the periods ~~. Each reconstruction is performed by using only proxy records significantly correlated at the 95% confidence level with the NAO on the corresponding learning period.~~ (see Fig. 8). Fig. 8~~shows the evolution of the proxy predictor set and the validation NSCE~~ .a) shows the NSCE scores obtained for the different ~~reconstructions and~~ reconstruction/learning periods. Using the ~~validation NSCE as a quality~~ NSCE metric, we find that the best reconstruction time window is ~~1000-1967 (19 proxy records used ; Fig. 8). Indeed, the associated validation NSCE ($\bar{r} = 0.48$; $r \in [0.11, 0.68]$) are on average significantly greater than all of the others at least at the 95% confidence level. In addition, we observe two significant drops in validation NSCE at the 95% confidence level, depending on the size of the reconstruction period: One from 1978 to 1979 and one from 1994 to 1995 (Fig. 8). Both can be associated to important changes in the number and the nature of proxy predictor sets (Fig. 8). For the other methods, we found that the optimal reconstruction period for Enet and RF is~~ 1000-1973 ~~(not shown), while the optimal reconstruction period for PCR is 1000-1970 (not shown).~~ for the RF method and 1000-1970 for the 3 others.

Following the optimal setup for each method from section 4.1.1 and 4.1.2, RF uses 110 records, PCR uses a total 65 records with 16.28 records selected per training/testing split on average. Enet and PLS use a total of 60 records with 17.26 records selected per training/testing split on average. Among these four optimized reconstructions which are the final ones of this study, the RF gives the highest NSCE scores with $med(S_{NSCE}) \simeq 0.18$ and $S_{NSCE} \in [-0.33, 0.39]$ (Fig. 8).

In contrast with the ~~length of the training periods, the~~ previous setups investigated in this study, the four methods are strongly affected by the choice of the reconstruction period~~appears as an important source of reconstruction uncertainty. This parameter strongly affects the reconstruction by modifying directly or indirectly the predictors.~~ . Thus, we recommend to determine this period carefully with numerous simulations on different time windows, following the approach we presented here. Overall, this study shows that for each optimisation, PCR and PLS are less reliable to reconstruct the NAO than RF and Enet (section 4.1.1, 4.1.2 and 4.1.3).

4.2 Reconstructions assessment

We now compare and assess the best reconstructions obtained for each of the methods. The four optimized reconstructions are obtained by maximizing the ~~validation NSCE~~ NSCE scores on the training/testing period (see section 4.1.2) and the total reconstruction period (~~PLS:~~ see section 4.1.3~~; other methods: not shown~~), using the ~~proxy records~~ full set of proxy records for RF and only using the proxy records significantly correlated at the ~~95%~~ 95% confidence level with the NAO ~~on the corresponding learning period~~ (index over each training periods for the other methods (see section 4.1.1~~and 4.1.3~~).

4.2.1 Comparison with previous work

Fig. 9 shows the different reconstructions of the NAO, including the Ortega et al. (2015) calibration constrained reconstruction (only proxy-based), and Tab. 1 exhibits the paired correlations between the 5 reconstructions. ~~All the reconstructions are significantly correlated with each other at the 99% confidence level on their overlap periods even if they were performed with different proxy groups and learning periods (Tab. 1) . As they also have been optimised for multiple sources of sensitivity for the reconstruction, these results strongly supplies that the reconstructions we propose are reliable to translate the variations of the NAO index over the past millenium.According to the validation scores, the best reconstruction that we found has been obtained using the PLS method on the reconstruction period 1000-1967, using the 19~~

proxy records significantly correlated with the NAO index on this period. The averaged validation scores attributed to each of the best reconstruction for each method are: $\bar{r}_{PCR} = 0.41, \bar{r}_{RF} = 0.41, \bar{r}_{RF} = 0.43, \bar{r}_{PLS} = 0.48$. The correlation coefficient between the original NAO index and The regression uncertainties (see supplementary informations) are also shown for the PLS reconstruction is about 0.63 (p<0.01)on the time window 1823-1967 while four reconstructions of this study on Fig. 9. The normality of the residuals for the fours methods has been verified as demonstrated in Fig. 11. Tab. 1 and Fig. 9 shows that the NAO reconstruction based on RF is distinguishable from the four others including Ortega et al. (2015). Indeed its correlation with the Ortega et al. (2015) reconstruction is about 0.45 (p<0.01) on the time window 1823-1969. Furthermore, as the PLS validation NSCE are strongly greater than those from Ortega et al. (2015). This is true in comparison to the calibration constrained reconstruction ($r_{val} \in [-0.14; 0.58]; \bar{r} = 0.24$) and the model constrained reconstruction ($r \in [0.14; 0.64]; \bar{r} = 0.42$) (see Ortega et al., 2015). To understand this difference it is important to note that the best performing reconstruction in Ortega et al. (2015) has a substantially weaker other indices is between 0.69 and 0.79 (Tab. 1) while the paired correlations obtained between the others are greater than 0.95. Additionally Fig. 10 shows that the RF reconstruction has a higher correlation with the observed NAO in their overlap period (Jones et al. (1997) NAO index than the other indices: r=0.42, 0.96 (p<0.01)that all the NAO reconstructions discussed here for the different methods ($r \in [0.56, 0.63]$, $p < 0.01$) . The 5 reconstructions, including Ortega et al. (2015) do not show a predominant positive NAO phase during the MCA, contrary to the hypothesis formulated by Trouet et al. (2009) . The different optimizations performed on the different methods allowed us to find the optimal reconstruction contrary to other NAO reconstructions. Hence, we , while Ortega et al. (2015) reconstruction has a correlation of 0.45 (p<0.01). The RF reconstruction that uses 108 proxy records (22 common proxies with Ortega et al. 2015) presented in Fig 12, has the best NSCE scores ($med(S_{NSCE}) = 0.18; S_{NSCE} \in [-0.33, 0.39]$) and its correlation scores ($med(S_{cor}) \simeq 0.47; S_{cor} \in [0.09, 0.81]$) are significantly higher at the 99% confidence level than Ortega et al. (2015) calibration constrained reconstruction ($S_{cor} \in [-0.14; 0.58]; med(S_{cor}) \simeq 0.24$) and model constrained reconstruction ($S_{cor} \in [0.14; 0.64]; med(S_{cor}) \simeq 0.43$). We thus statistically verified that the best reconstruction from this study is more robust and reliable than those in from Ortega et al. (2015). This improvement in performance may arise from the inclusion of new relevant proxy records into the reconstruction, but also the using from the use of a new statistical regression methods . The PLS reconstructionuses 19 different proxy records, 12 of them have been used in the last NAO reconstruction study for climate index reconstruction: the RF. Finally, it has to be stressed that the 5 reconstructions presented in Fig. 9, including Ortega et al. (2015)(see Fig. ??). Among the 7 proxy records we added, there is an Asian proxy recorded on tree rings, with a medium negative weight in the reconstruction. This proxy record belongs from , do not show a predominant positive NAO phase during the MCA, contrary to the hypothesis formulated by Trouet et al. (2009).

4.2.2 Response to external forcing

No significant correlation is found between the NAO reconstruction based on RF method and the Total Solar Irradiance (TSI) reconstruction from Vieira et al. (2011)($r \simeq -0.09; p \simeq 0.23$). The same is true for the best reconstruction of the Pages2K database 2014 version Pages 2K Consortium, 2013, and no reference are given. To refer to this proxy record, the reader can have a check to the proxy encoded " Asi_221" in the Pages2K 2014 database version Pages 2K Consortium, 2013. The six other proxy records are located in the Arctic area: three of them have been recorded from Greenland ice cores Vinther et al., 2010, two have been recorded in North Canada Vinther et al., 2008, Meeker and Mayewski, 2002, and the last one has been record in Northern Young et al., 2012(Fig. ??). For the other proxy records, the weight we attributed to them are consistent with those found in other methods (not shown) and Ortega et al. (2015). None of

the reconstructions (including Ortega et al. (2015)

4.2.3 ~~Response to external forcing~~

~~We now focus on the response of the NAO to external forcing: volcanic aerosols, Total Solar Irradiance (TSI) , and~~ shows clear negative phases during the Maunder and the Spörer minima as suggested by some model simulations [Shindell et al., 2004]. In addition, no significant correlation on the pre-industrial era has been found with the CO_2 ~~concentration. Indeed,~~ reconstruction based on a Law Dome (East Antarctica) ice core [Etheridge et al., 1996] (not shown), indicating that the NAO is not linearly associated with CO_2 variations over this time frame.

Ortega et al. (2015) suggested that a positive NAO phase is triggered two years after strong volcanic eruptions, a response that is not reproduced over the last millennium by model simulations [Swingedouw et al., 2017]. ~~By applying composite analysis of the NAO response to the~~ We use the 10 large volcanic eruptions selected in Ortega et al. (2015) and a second selection (see supplementary informations) of the 11 ~~strongest volcanic eruptions which occurred during the last millenium, and using dates from 4 different reconstructions of the last millenium volcanic activity Gao et al., 2008 ; Crowley and Unterman, 2013 ; Sigl et al., 2015 ; Ortega et al., 2015, we obtained the consistent results for the four regression methods developed here: a positive NAO response 2 and 4 years following the eruptiononset~~ largest volcanic eruptions from the well-verified reconstruction of Sigl et al. (2015). By using a superposed epoch analysis and Monte-Carlo approach as in Ortega et al. (2015) we find that using the same set of eruptions than Ortega et al. (2015) leads to the same result: a significant positive response of the NAO two years after the eruption. However, for RF this result is not significant with a p-value just above 0.1 (Fig. 13). ~~By using a Monte-Carlo approach as in Ortega et al. (2015) , we obtain significance levels greater than 99% for all methods, all volcanic reconstructions and all composites, except for the composite RF based on the volcanic activity reconstruction from Gao et al., 2008 and Sigl et al., 2015 (Fig. 13). The table giving the volcanic activity reconstructions is presented in appendix.~~ On the opposite, by using the Sigl et al. (2015) 11 largest volcanic eruptions, we find a significant response at the 90% confidence level for Enet, PLS and PCR, but one year after the eruption with a p-value under 0.05. For RF, the positive NAO response is significant 1 to 3 years after the eruption. Here again, the significance for the RF composite is smaller than for the other methods while this reconstruction is the most robust. Nevertheless, individual response analysis shows that for the RF reconstruction, this result is particularly significant for the 2 larget eruptions of the millennium (Samalas, 1257 and Kuwae, 1458) and not so clear for the 9 others (not shown). This result suggests that the positive NAO response might be mainly associated to volcanic eruptions with very large and rare intensities such as Samalas or Kuwae eruptions and concerns less eruptions with weaker intensities

5 Discussion and conclusion

5.1 Discussion, caveats and outlooks

The results presented above regarding the NAO have all been obtained using CliMoRec. Indeed, they require advanced programming and statistical knowledge to ensure a good estimation of the robustness of the reconstruction performed. This is possible in CliMoRec that proposes an integrated package through which parameters and methods can be efficiently tested and compared, together with advanced validation metrics such as the NSCE. Nevertheless, the methodology proposed in CliMoRec could be further improved in different ways.

~~On the other hand, we did not find any significant correlation with any of the TSI reconstructions available~~

Firstly, CliMoRec does not deal with missing data in proxy records. This implies selecting exclusively the proxy records that entirely cover the reconstruction period, which thus excludes some existing proxy records. Also, proxy records with gaps are not used in the present version of CliMoRec as their use in an interpolated version would artificially increase their weight in the reconstruction and thus possibly induce spectral artefacts in the reconstruction [~~Crowley, 2000 ; Vieira et al., 2011~~Hanhijarvi et al., 2013]~~. Moreover, none of the reconstructions (including Ortega et al., 2015)shows clear negative phases during the Maunder and the Spörer minima as some model simulations were suggesting~~ . Secondly, except RF which is a bootstrap aggregating approach, the proposed methods are classical regression approaches. In next versions of the device, it would be interesting to test other methods such as Gaussian Processes regression [~~Shindell et al., 2004. In addition, no significant correlation on the pre-industrial era~~ Stein, 1999], Expectation-Maximization algorithm [Dempster et al., 1977] and its regularized variants [Schneider, 2001 ; Mann et al., 2008 ; Guillot et al., 2015] or Bayesian Hierarchical models [Tingley and Huybers, 2010a ; Tingley and Huybers, 2010b Tingley, 2012 ; Tingley and Huybers, 2013 ; Cahill et al., 2016] that can deal with missing data and compare the reconstructions obtained with the four methods already included. Another point that is limiting the capacities of CliMoRec is that it is based on the assumption that teleconnections of the reconstructed mode are stationary in time, while they may depend on the state of the climate system. This is a classical limit for statistical climate reconstructions but it can be evaluated by use of pseudo-proxy methods (e.g. Lehner et al., 2012, Ortega et al. 2015). On this aspect, more complex methods like data assimilation can clearly overcome this weakness by combining model and data. The use of such approaches for last millennium remains nevertheless very complex primarily because of their computational cost and the lack of data. They are however emerging (e.g. Hakim et al., 2016 ; Singh et al., 2018). Data assimilation techniques can be very model dependent as highlighted for the ocean over the recent period (Karspeck et al., 2015) so that their reconstruction of a given regional climatic modes can suffer from interferences with reconstructions of other aspect of the climate. Thus, dedicated approaches like the ones developed here can be seen as very complementary approach and may increase our confidence in the reconstructions. Indeed, if different approaches provide very similar results, this can be interpreted as a source of robustness for a given result or reconstruction.

Another caveat concerns the fact that the present version of CliMoRec does not account for dating uncertainties in proxy records. Future developments of CliMoRec may allow to take into account these uncertainties and to provide their estimation along time. For doing so, deeper investigations for each proxy record are needed as these sources of uncertainty are not exhaustively provided in P2k2017. Also, we found that the reconstructions performed by CliMoRec provide a clear loss of variance over the learning period and the reconstructed period (before 1856) (see supplementary table 4). The RF method is the only one that reproduces adequately the NAO amplitude only over the learning period but also provide a significant loss of variance over the reconstructed period. This indicates that the loss of variance over the reconstruction period could partly be due to the proxy records themselves and not only to the statistical approach.

A key aspect that has been found ~~with a CO_2 reconstruction based on a Law Dome (East Antarctica) ice core~~ within this study is the sensitivity of the results to the validation metric used. Indeed, we also used correlation as the main score for the test period. It appears that this metric was mainly capturing the phasing of the modes in their reconstruction (not shown) [~~Etheridge et al., 1996~~Wang et al., 2014]~~, indicating that the NAO is not linearly associated with CO_2 variations.~~ . By using NSCE, we improved the strength of our reconstruction since other aspects than the synchronisation were accounted for. This latter

metric, which is more classical in prediction evaluation further highlights that the RF method outperforms most of the others methods, and notably the PCR which is a classical method used in paleoclimatology [Cook et al., 2002 ; Gray et al., 2004 ; Ortega et al., 2015 ; Wang et al., 2017]. Other metrics of prediction validation exist (e.g. Continuous Ranked Probability Score, Gneiting and Raftery, 2007) so that a more extensive analysis of the sensitivity of the reconstruction to other metrics for the validation period might be very useful. Thus, the development of other validation metrics in next versions of CliMoRec appears as an interesting avenue to explore.

6 ~~Conclusions~~

5.1 Conclusions

We have proposed and described here four statistical methods for reconstructing ~~some~~ modes of climate variability and have compared them for a particular example: the reconstruction of the NAO. By ~~investigating~~ identifying and minimizing the sources of reconstruction uncertainty ~~,~~ due to the method used (sections 3, 4.1.1, 4.1.2 and ~~4.2.1~~4.1.3), the time frame considered (section 4.1.3) and the proxy selection (sections 4.1.1 and 4.1.3), we found the optimal NAO reconstructions~~, all providing better validation and calibration results than previous studies~~ . It was obtained for the RF method over the time frame 1000-1973 using the 108 proxy records available on this time frame (section 4.2.1) ~~. All the reconstructions show a positive NAO response the year 2 and 4 following volcanic eruptions, in agreement with Ortega et al. (2015). Moreover they also presents low-frequency negative phases at the multi-decadal scale (section 4.2.1) , which may induce cold winter conditions in Europe during these periods (e.g. 11^{th}, 12^{th} and 15^{th} centuries).~~ with a training sample length of 80% of the length of the learning period. This method has not been used yet to our knowledge for climate index reconstructions and seems thus promising. The reconstruction we obtained is distinguishable from the Ortega et al. (2015) reconstruction but remains significantly correlated with it ($r=0.47$; $p<0.01$ over the period 1073-1855).

~~We have showed that using~~

We have shown that for Enet, PLS and particularly PCR which is frequently used in paleclimatology, selecting proxy records with a strong correlation with the index to be reconstructed over the ~~overlapping period~~ training periods is a good ~~means for improving the validation~~ way to improve the NSCE scores, and hence ~~allow~~ it allows more reliable reconstructions ~~. Among the 540 available proxy records collected, containing the PAGES 2K database 2014 version~~ (section 4.1.1). Contrarily, RF gives more reliable reconstructions using the whole set of records (section 4.1.1). This may be due to the fact that it has been mainly developed for large datasets [~~Pages 2K Consortium, 2013~~Breiman, 2001]~~, which is a well-verified high resolution proxy collection, only 19 covers the reconstruction period 1000-1967 and are significantly correlated with the NAO index (at the 95% confidence level) on the period 1823-1967. Gathering new proxy records , significantly correlated with the NAO, ~~. For both cases, gathering new proxy records to the 554 available proxy records collected, may be a reliable source of reconstruction improvement. The inclusion of new NAO-sensitive proxy records in the future may thus lead to better reconstructions. ~~The toolbox we developed in this paper~~ CliMoRec should allow to easily perform such new reconstructions~~, thanks to a devis made available to the community.~~ .

In order to extract the most robust reconstruction, numerous simulations are needed. To ~~facilitate it, the statistical tool we developed~~ simplify it, CliMoRec performs a reconstruction by considering several entries: an index of the climate mode, the reconstruction period, the length of the training window (in proportion

of the total length of the learning window), the number of training/testing period ~~samplings~~splits, and a threshold confidence level for the correlation between the proxy records and the target index~~(appendix 1). This modular statistical tool~~ . CliMorRec is an opportunity to reconstruct quickly and with quantified reliability several climate modes. This may ~~allows~~ allow us to improve our understanding of the last millennium ~~large scale~~ large-scale climate variations, such as the MCA and the LIA, as well as the interactions between the modes, which will be analysed in future studies.

6 ~~Acknowledgements~~

Author Contribution

Simon Michel has integrally coded CliMorRec and used it to produce the results of this study. Simon Michel has been the main writer of the manuscript, including figures production. Didier Swingedouw has contributed to develop the main features of CliMoRec and has supervised the manuscript writing during the whole process. Pablo Ortega, Juliette Mignot and Myriam Khodri has contributed to write the manuscript and to discuss about results. Marie Chavent has contributed to write the manuscript, with a particular focus on section 2 and 3.

Acknowledgements

This research was partly funded by the Universite de Bordeaux. It is also funded by the LEFE-IMAGO project. To develop the statistical tool and analyse its outputs, this study benefited from the the IPSL Prodiguer-Ciclad facility, supported by CNRS, UPMC Labex L-IPSL. Finally, this study used the ~~PAGES~~ Pages 2K database ~~2014 version~~version 2.0, available online and supported by the ~~PAGES~~ Pages group.

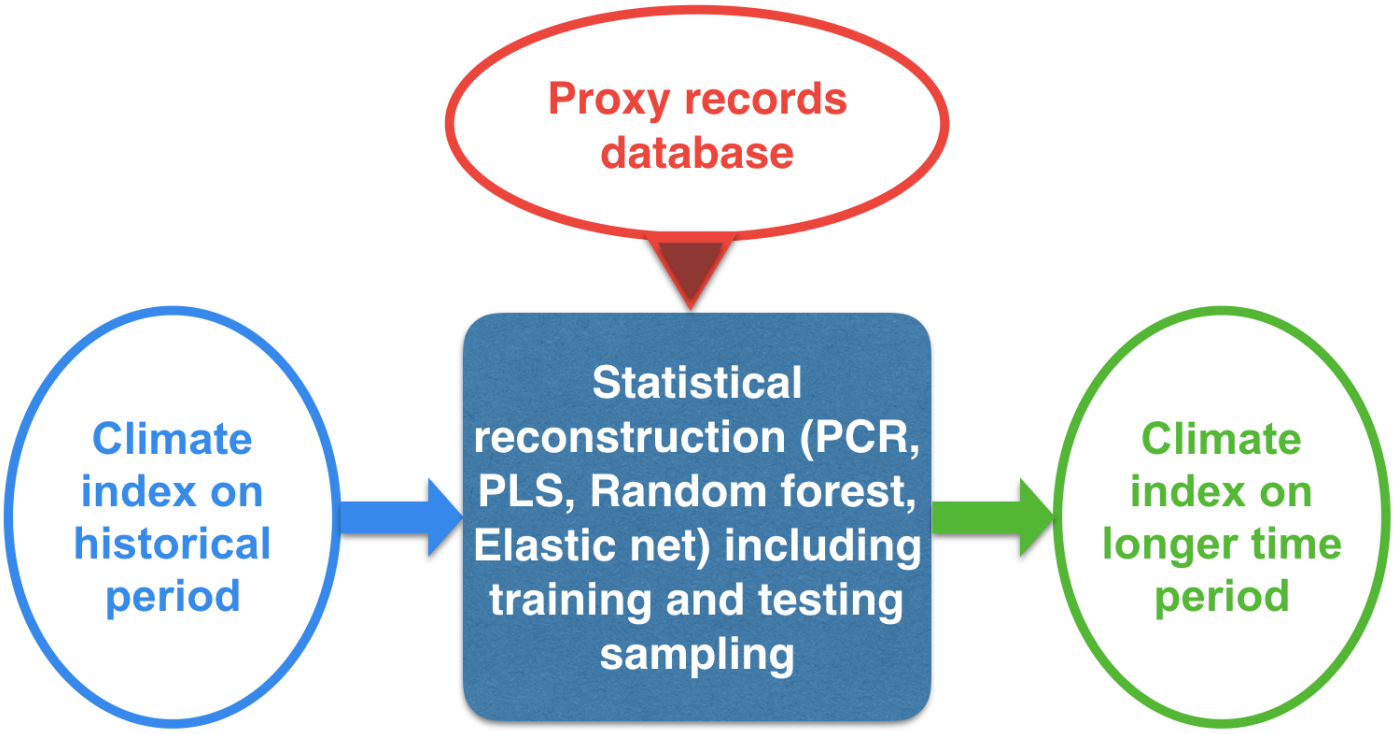

FIGURE 1 – Scheme summarising the main features of ~~the proposed statistical toolbox~~CliMoRec.

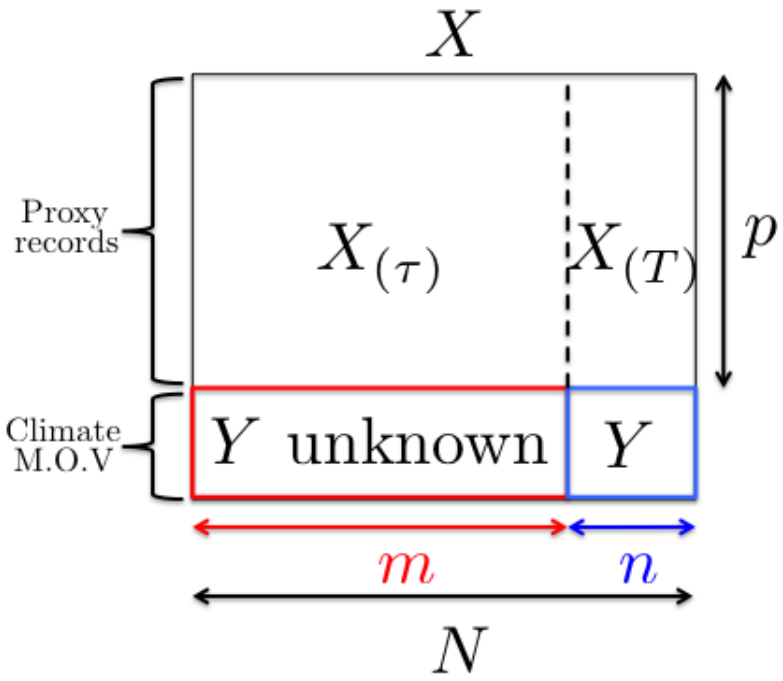

Figure 2 – Scheme of the initial data. X and Y are respectively the proxy records matrix and the index of the considered mode of variability. N is the size of the common period of all proxy records. n is the size of the common period of all proxy records and the index of the mode of variability. m is the size of the common period of all proxy records, where the mode of variability is not known. p is the number of proxy records. $X_{(T)}$ is the sub-matrix of X where the mode of variability is known. $X_{(\tau)}$ is the sub-matrix of X where the mode of variability is not known.

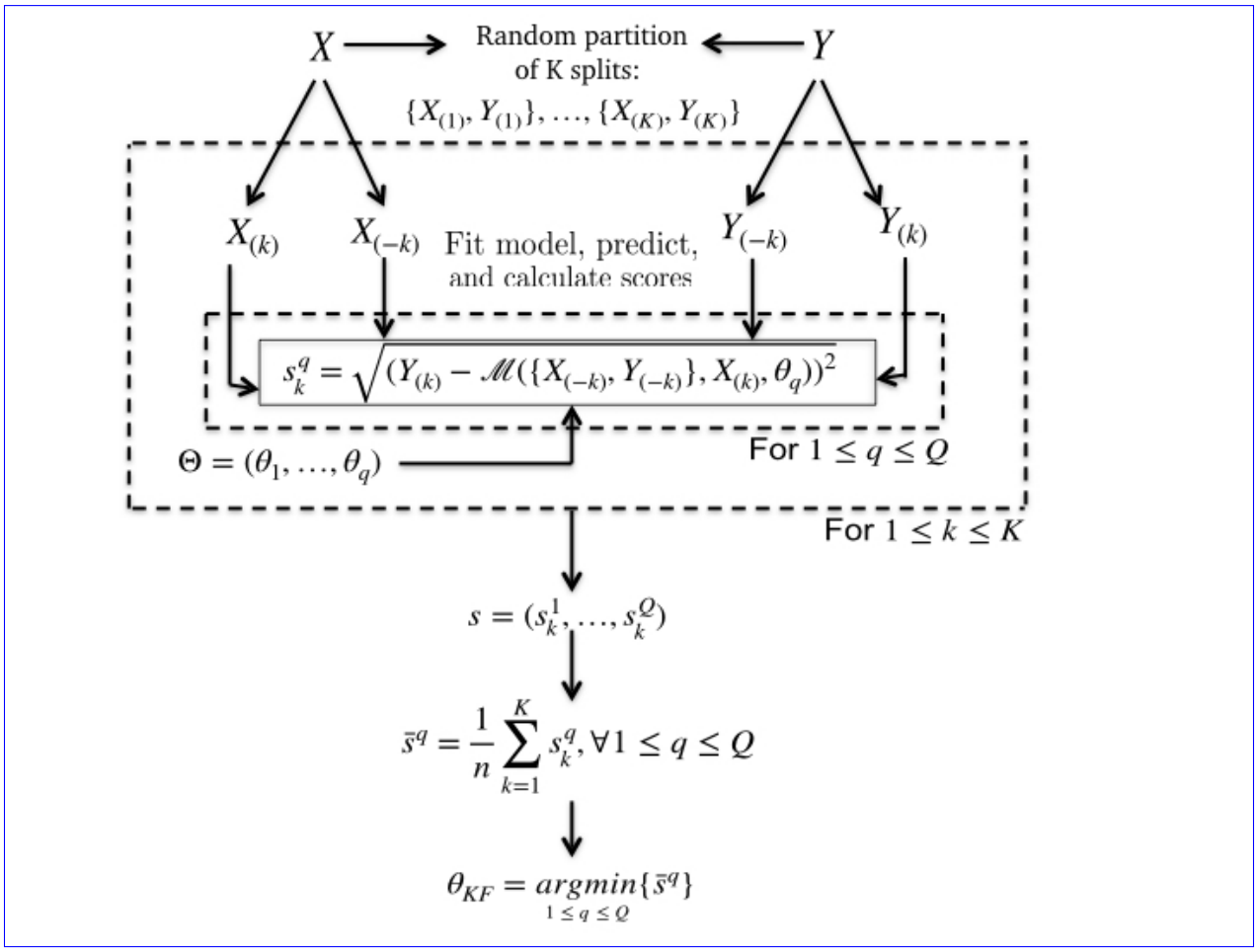

Figure 3 – Scheme of a ~~leave-one-out~~ K-Fold cross validation procedure to select the optimal parameter of a specific learning method $\mathcal{M}$. X is the input set of predictors and Y the corresponding variability mode index. $\forall 1 \leq i \leq n$, ~~$\{X_{(i)}, Y_{(i)}\}$~~ $\{X_{(k)}, Y_{(k)}\}$ is the ~~i^{th}~~ k^{th} randomly drawn group of observation and ~~$\{X_{(-i)}, Y_{(-i)}\}$~~ $\{X_{(-k)}, Y_{(-k)}\}$ contains all observations except the i^{th}. ~~$\Theta = (\theta_1, \ldots, \theta_K)$~~ $\Theta = (\theta_1, \ldots, \theta_Q)$ is the ensemble of possibles values of ~~$\theta \in \mathbb{R}^q$~~ $\theta \in \mathbb{R}^s$.

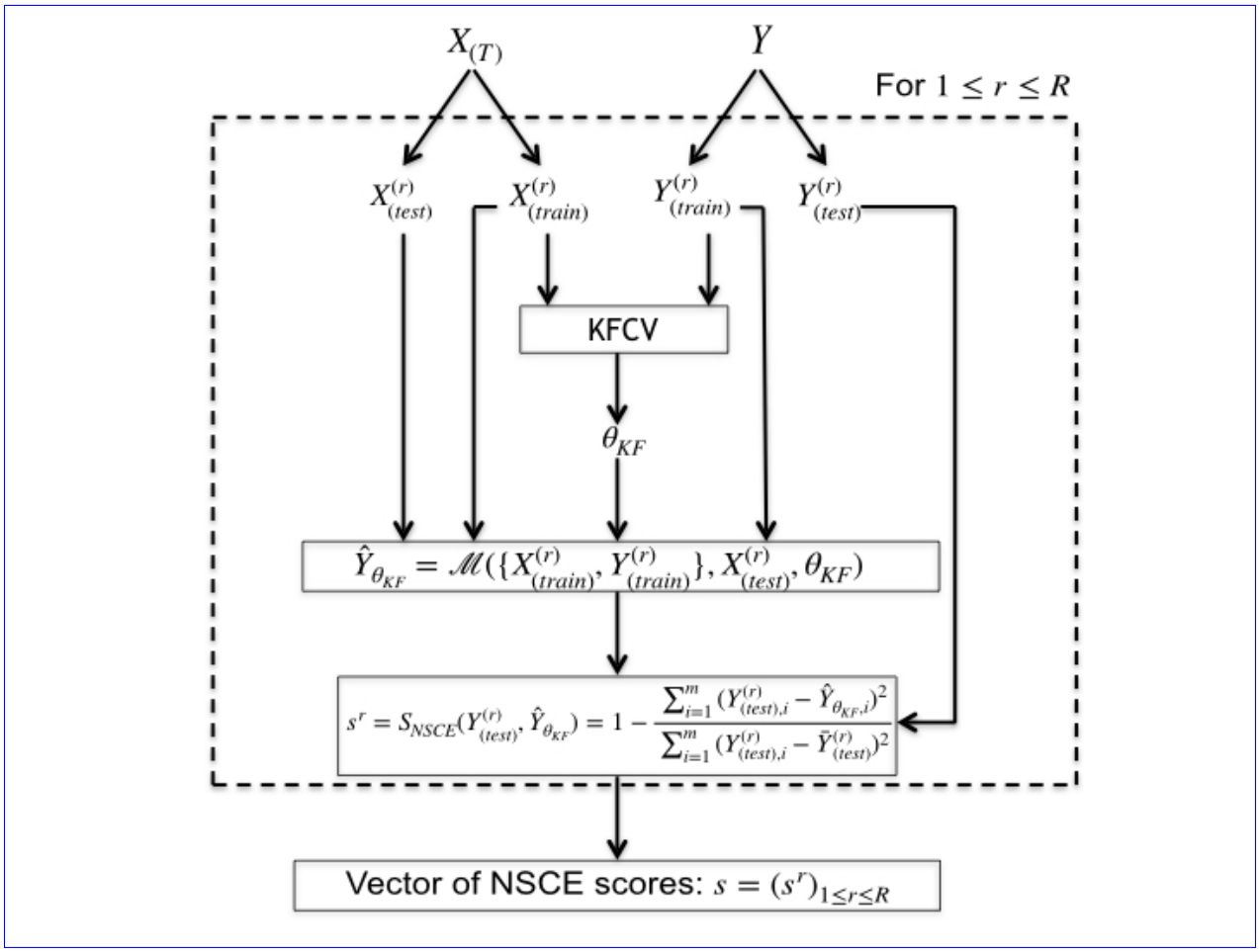

Figure 4 – Scheme of the whole procedure for scores calculation for a given method $\mathcal{M}$. Y is the index of the chosen mode of variability. $X_{(T)}$ is the proxy dataset restricted to the period where Y known. $\{X^{(r)}_{(train)}, Y^{(r)}_{(train)}\}$ is the r^{th} training sample and $\{X^{(r)}_{(test)}, Y^{(r)}_{(test)}\}$ is the r^{th} testing sample. ~~θ_{LOO}~~ θ_{KF} is the empirically optimal set of parameters obtained by applying the KFCV (Fig. 3; section 2.5.1)

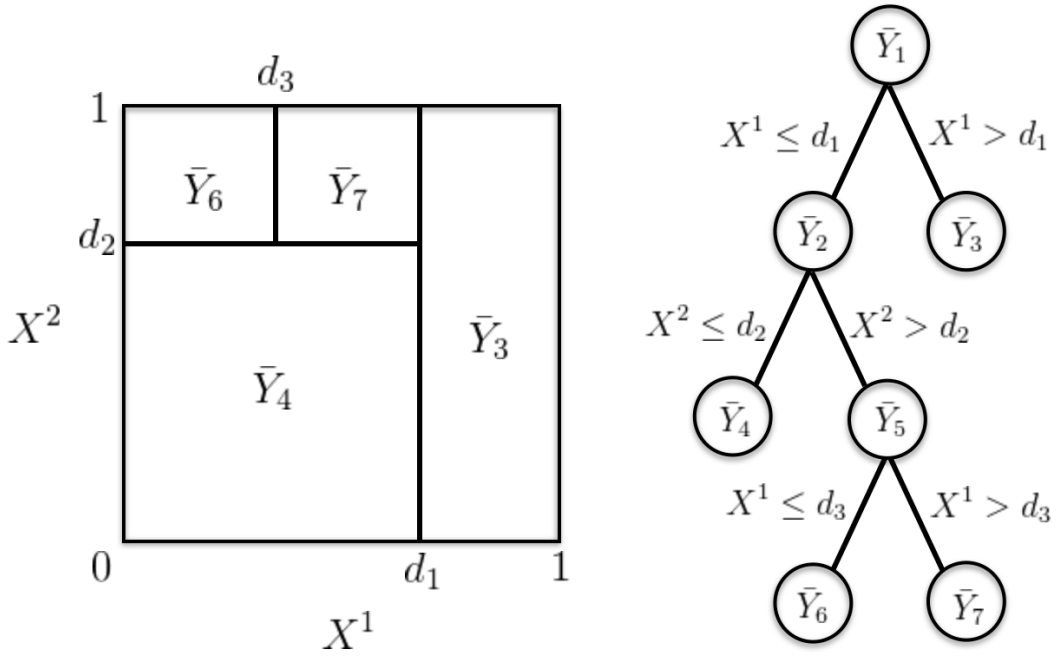

Figure 5 – Dyadic partition of the unit square (left) and its corresponding regression tree (right). *Y is the predictand and X^1, X^2, X^3 are the predictors. d_1, d_2 and d_3 are the optimal thresholds of the three steps respectively.*

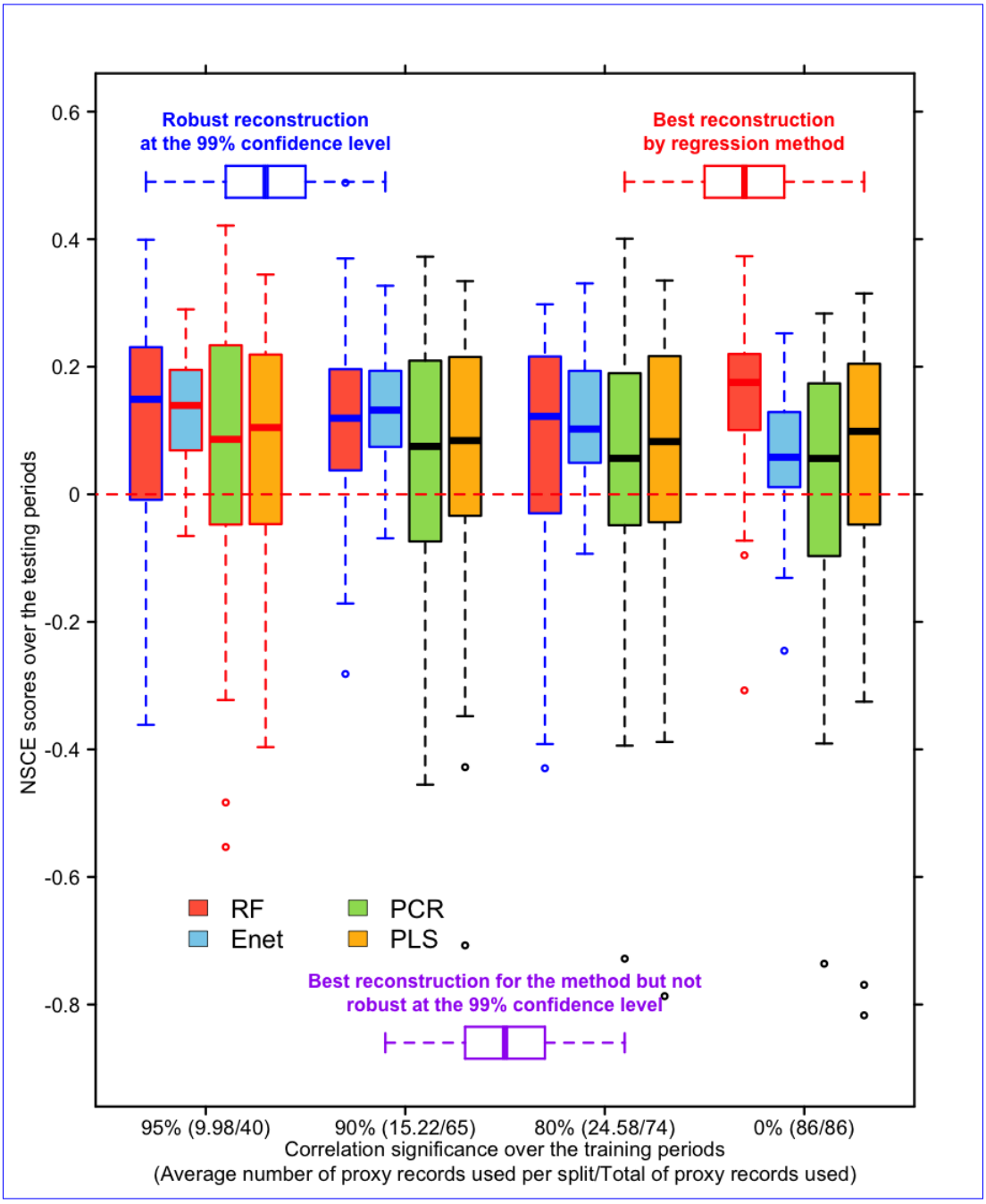

Figure 6 – ~~Geolocation, types and correlation confidence level between the 122 available proxy records for the period 1000-1970, and the NAO index on the period 1823-1970~~Boxplot of ~~validation~~ NSCE scores obtained for the four methods and different groups of proxy records by reconstructing the NAO index on the period 1000-1970 with $R = 50$ ***~~validation~~training/~~calibration~~ testing randomly drawn samples. ~~Calibration~~ Training samples size is*** $n_{train} = ~~111~~ n_{train} = 92$***, and ~~validation~~ testing samples size is*** $n_{test} = ~~37~~ n_{test} = 23$***. Green boxplots are the ~~validation~~ NSCE scores obtained for the PCR method. Yellow boxplots are the ~~validation~~ NSCE scores obtained for the PLS method. Red boxplots are the ~~validation~~ NSCE scores obtained for the RF method. Blue boxplots are the ~~validation~~ NSCE scores obtained for the Enet method. The first cluster of boxplots is the ~~validation~~ NSCE scores obtained by using all the available proxy records over the period (~~122~~ 110 proxy records). The second cluster of boxplots is the ~~validation~~ NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 80% confidence level ~~(61 proxy records)~~over the training periods. The third cluster of boxplots is the ~~validation~~ NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 90% confidence level ~~(35 proxy records)~~over the training periods. The fourth cluster of boxplots is the ~~validation~~ NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 95% confidence level ~~(18 proxy records)~~over the training periods. Boxplots with blue edges are the scores significantly positives at the 99% confidence level. Boxplots with red edges correspond to the scores associated with the best reconstruction for each method.***

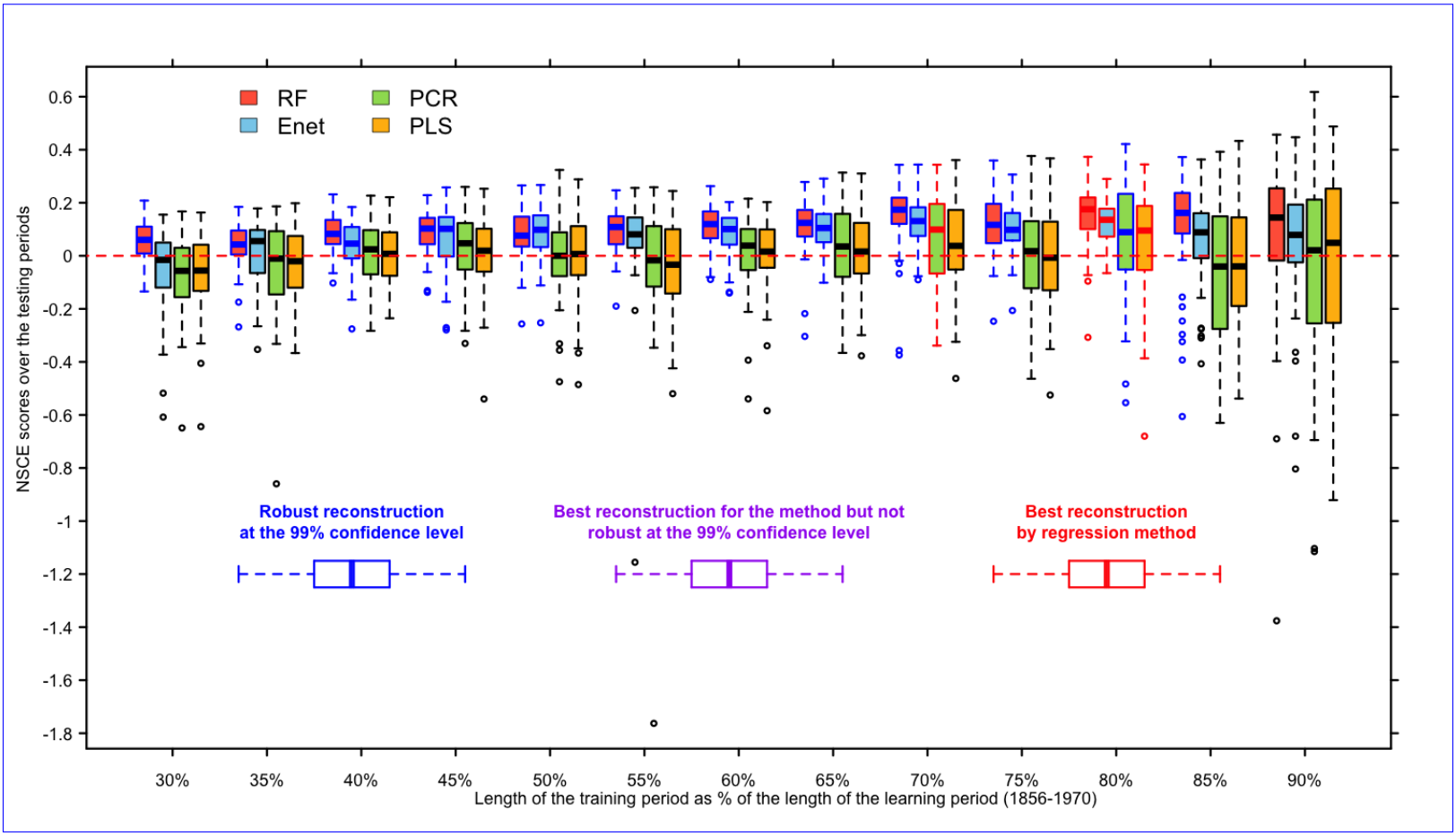

Figure 7 – ~~validation~~ NSCE scores obtained for different sizes of the ~~calibration~~ training samples: from ~~5%~~ 30% to ~~95%~~ 90% of the length of the learning period (n = 148) with a 5% step. All of the reconstructions are performed for period 1000-1970. Red boxplots are ~~validation~~ NSCE scores obtained by ~~100~~ 50 training/testing sampling using the RF method. Blue boxplots are ~~validation~~ NSCE scores obtained by ~~100~~ 50 training/testing sampling using the Enet method. Yellow boxplots are ~~validation~~ NSCE scores obtained by ~~100~~ 50 training/testing sampling using the PLS method. Green boxplots are ~~validation~~ NSCE scores obtained by ~~100~~ 50 training/testing sampling using the PCR method. ~~All of the~~ RF reconstructions are ~~made~~ performed using the ~~reconstruction period 1000-1970~~ whole set of available proxy records (110,section 4.1.1). Enet, PLS and PCR reconstructions are performed by selecting the ~~18~~ proxy records significantly correlated with the NAO ~~index~~ at the ~~95%~~ 95% confidence level over the ~~learning period 1823-1970.Correlations between the best reconstruction of each method given the calibration samples size and those obtained from all of the investigated calibration samples size: from 5% to 95% of the size of the learning period (n = 148) with a 5% step. The best PCR and PLS proportion for the~~ training ~~samples length is 70% of the learning period~~ periods (~~n_{train} = 104; n_{test} = 44~~section 4.1.1)~~while the RF best calibration samples size is 55% of the size of the learning period (n_{train} = 81; n_{test} = 67). Red line gives~~ Boxplots with blue edges are the ~~corresponding correlations for~~ scores significantly positives at the ~~RF method~~99% confidence level. ~~Blue line gives~~ Boxplots with red edges correspond to the ~~corresponding correlations for~~ scores associated with the ~~Enet method. Yellow lines gives the corresponding correlations~~ best reconstruction for ~~the PLS~~ each method.~~Green line gives the corresponding correlations for the PCR method.~~

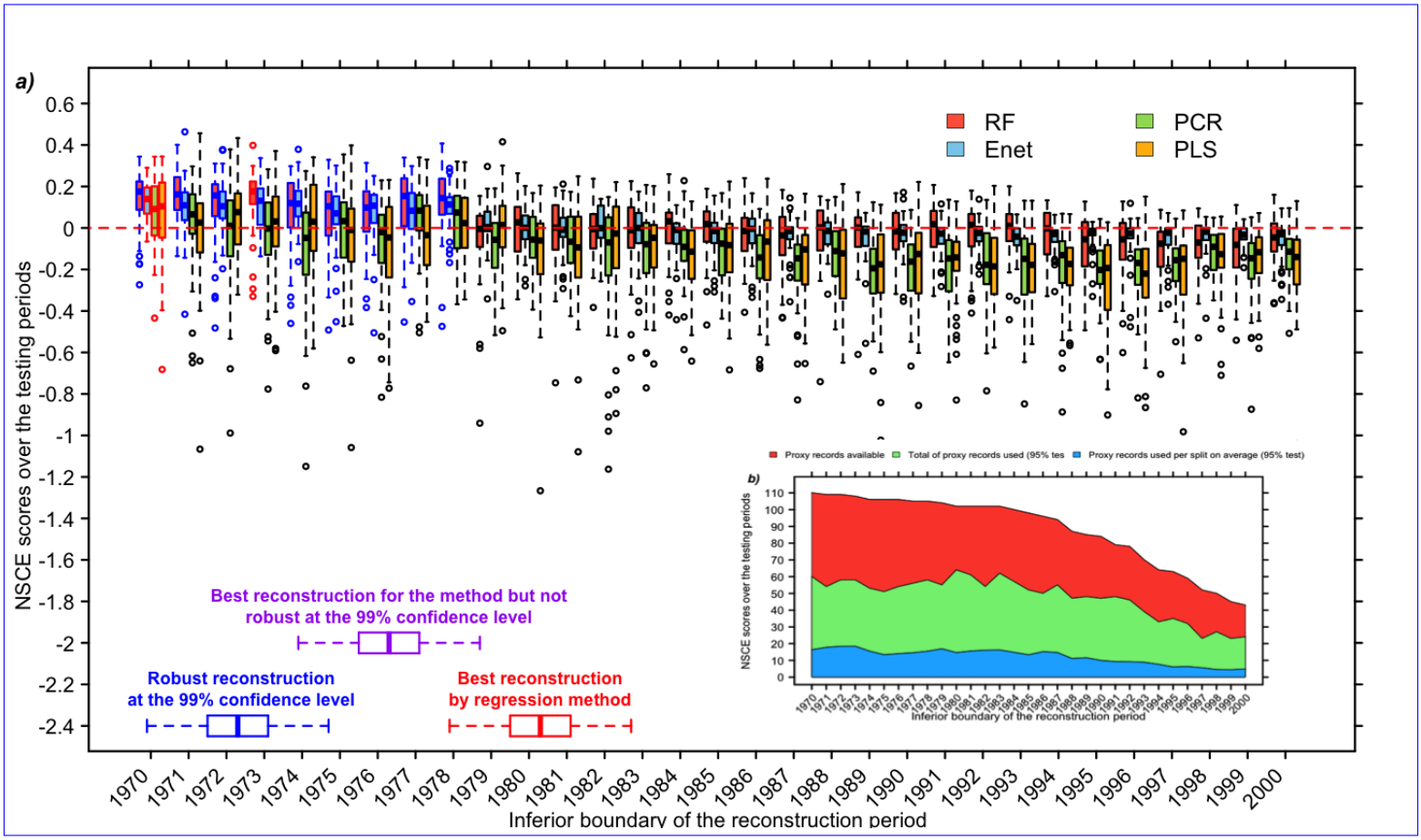

Figure 8 – All of the reconstructions are ~~made by~~ performed for $R = 50$ ~~sampling calibration~~randomly drawn training/~~validation, using the PLS method~~testig samples. The proportion of the length of the training samples is fixed to 70% *for PCR and and only the proxy significantly correlated with the NAO index at the* 95% *confidence level on the learning period are used for reconstruction. ~~The yellow boxplots~~ Reconstructions are ~~the validation NSCE obtained for each of the 36~~ performed using 31 reconstruction period: from ~~1000-1965~~ 1000-1970 to 1000-2000 by moving the superior born by 1. ~~Filled areas: Evolution of~~ RF reconstructions are performed using the ~~proxy predictor~~ whole set ~~. For each reconstruction period~~of available proxy records (110, section 4.1.1) with training samples length of 80% of the ~~selected~~ length of the learning period (section 4.1.2). PCR reconstructions are performed using by selecting the proxy records ~~are those which cover~~ significantly correlated at the ~~reconstruction~~ 95% confidence level with the NAO over the training periods (section 4.1.1) with training samples length of 70% of the length of the learning period (section 4.1.2). PLS and Enet reconstructions are performed using by selecting the proxy records significantly correlated at the 95% confidence level with the NAO ~~index~~ over the training periods (section 4.1.1) with training samples length of 80% of the lengt of the learning period (section 4.1.2). a) Red boxplots are the NSCE scores obtained using RF method. Blue boxplots are the NSCE scores obtained using Enet method. Red green are the NSCE scores obtained using PCR methodh. Yellow boxplots are the NSCE scores obtained using PLS method. Boxplots with blue edges are the scores significantly positives at the ~~95%~~ 99% confidence level~~on corresponding~~ . Boxplots with red edges correspond to the ~~learning period~~scores associated with the best reconstruction for each method. ~~Cyan area: proxy~~ b) Proxy records ~~finishing before 1970 included~~available/used by reconstruction period. Red area ~~:~~ gives the number of available proxy records ~~finishing after 1970 excluded and before 1980 included~~which is typically the number of records used for the RF reconstructions. Green area: ~~proxy~~ total of records ~~finishing after 1980 excluded~~ used for Enet, PCR and ~~before 1990 included~~PLS for each reconstruction period. Blue area: number of proxy records ~~finishing after 1990 excluded~~ used per training/testing splits on average for Enet, PCR and ~~before 2000 included~~PLS methods.~~Purple area: proxy records finishing after 2000 excluded.~~*

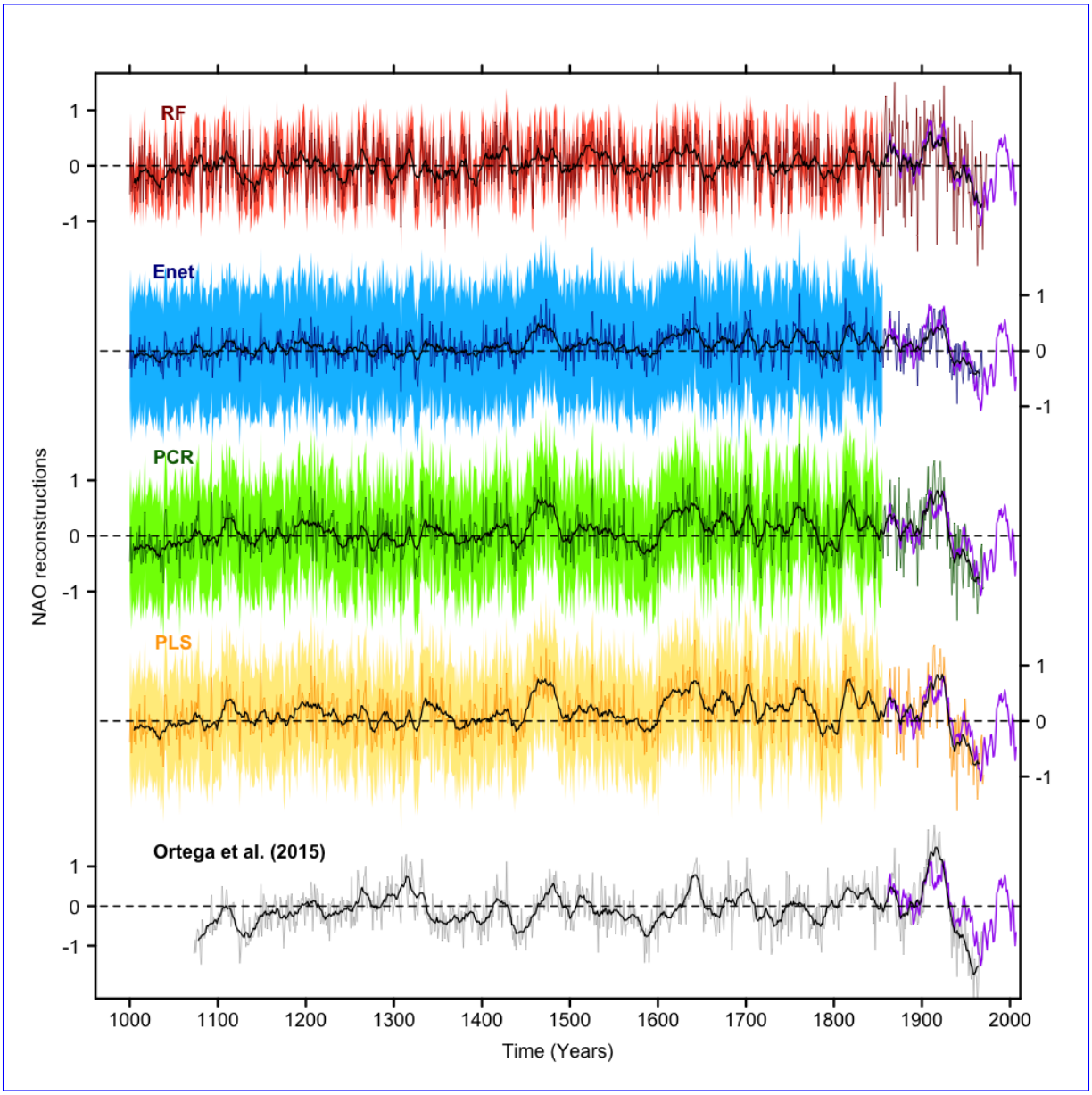

Figure 9 – Red line: RF reconstruction on the period 1000-1973 (section 4.1.3), using ~~18~~ the whole set of available proxy records ~~significantly correlated at the 95% confidence level~~ (110, section 4.1.1) with ~~a proportion~~ training samples length of 80% of the length of the ~~training samples of 45%~~ learning period (section 4.1.2). ~~Dark red line: ten years low-pass filter of the RF reconstruction.~~ Blue line: Enet reconstruction on the period ~~1000-1973, using 18~~ 1000-1970 (section 4.1.3) by selecting the proxy records significantly correlated with the NAO index at the ~~95%~~ 95% confidence level ~~,~~ over the training periods (section 4.1.1) with ~~a proportion~~ training samples length of 80% of the length of the ~~training samples of 65%~~ learning period (section 4.1.2). ~~Dark blue~~ Green line: ~~ten years low-pass filter of the Enet~~ PCR reconstruction ~~. Yellow line: PLS reconstruction~~ on the period ~~1000-1967, using 19~~ 1000-1970 (section 4.1.3) by selecting the proxy records significantly correlated with the NAO index at the ~~95%~~ 95% confidence level ~~,~~ over the training periods (section 4.1.1) with ~~a proportion~~ training samples length of 70% of the length of the ~~training samples of 70%~~ learning period (section 4.1.2). ~~Dark yellow~~ Orange line: ~~ten years low-pass filter of the RF~~ PLS reconstruction ~~. Green line: PCR reconstruction~~ on the period 1000-1970 ~~, using 19~~ (section 4.1.3) by selecting the proxy records significantly correlated with the NAO index at the ~~95%~~ 95% confidence level ~~,~~ over the training periods (section 4.1.1) with ~~a proportion~~ training samples length of 80% of the length of the ~~training samples of 70%~~ learning period (section 4.1.2). ~~Dark green~~ Black line (tiny): ~~ten years low-pass filter of the RF reconstruction~~Ortega et al. ~~Grey line: Calibration~~ calibration constrained ~~NAO~~ reconstruction [~~Ortega et al., 2015~~Ortega et al., 2015]~~on the period 1073-1969~~. Red area: Regression uncertainties (see supplementary) for RF reconstruction. Blue area: Regression uncertainties for Enet reconstruction. Blue area: Regression uncertainties for PCR reconstruction. Orange area: Regression uncertainties for PLS reconstruction. Heavy black ~~line: ten years low-pass filter of~~ lines are the ~~calibration constrained~~ corresponding 11-year filtered reconstructions for each method. Purple lines: superposed 11-years filtered Jones et al. (1997) NAO ~~reconstruction Ortega et al., 2015~~ index.

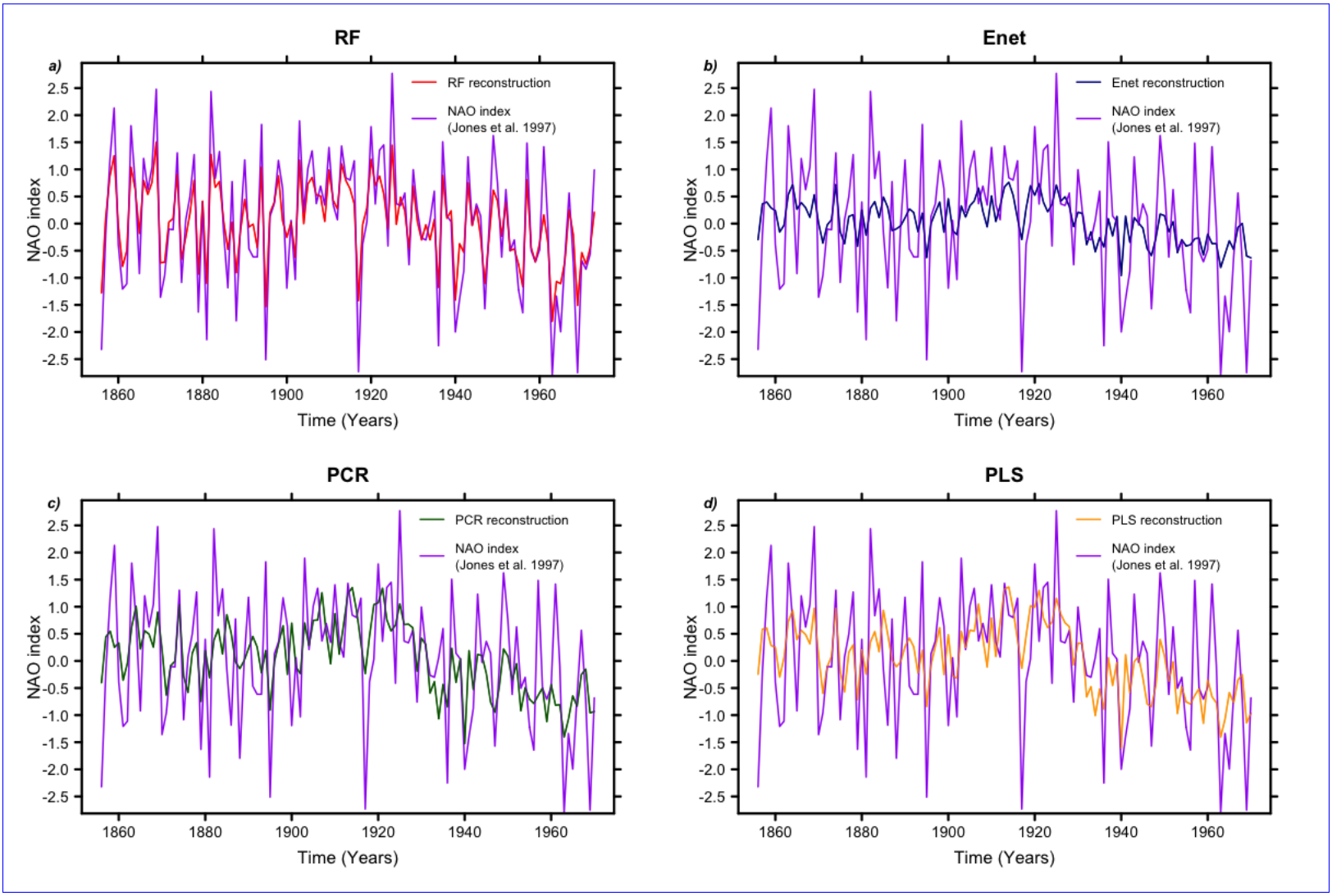

Figure 10 – ~~Map and weights~~ Comparison of ~~the 19 proxy records significantly correlated~~ reconstructions from this study with the original Jones et al. (1997) NAO index ~~on the time window 1000-1967~~over their common period. ~~These weights are~~ a) RF reconstruction. b) Enet reconstruction. c) PCR reconstruction. d) PLS reconstruction. Purple

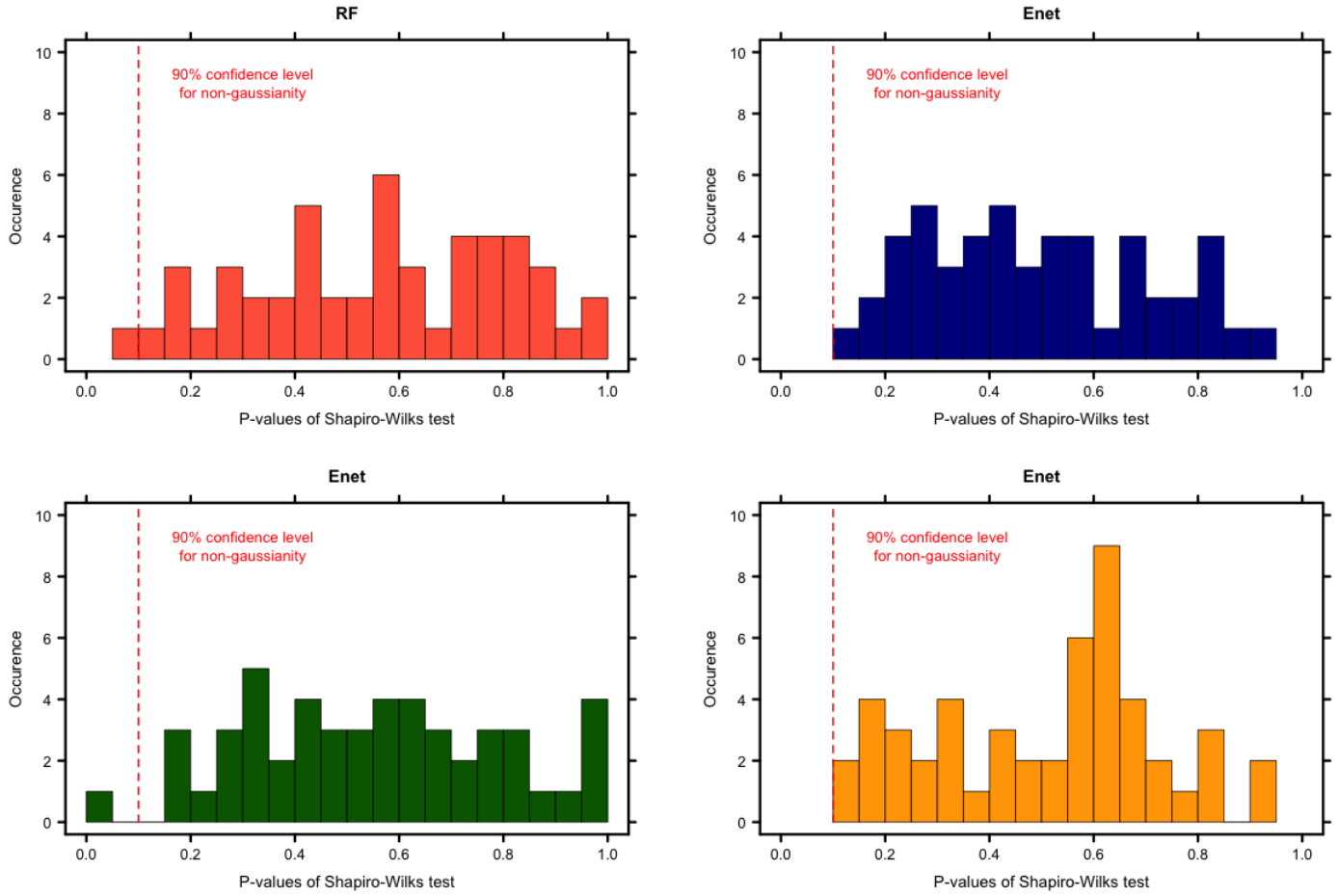

Figure 11 – P-values obtained from Shapiro-Wilk normality tests on the ~~PLS method~~ residuals from each reconstruction of Fig. 9. For a), b) c) and d), the repartition of the 50 p-values obtained for each training/testing split are ~~calculated by projecting regression coefficients on~~ presented. Red dashed lines indicates the ~~loadings~~ 90% confidence level for non-normality. For $0 \leq \alpha \leq 1$, if p-value<=α, it means that the residuals distributions is significantly not gaussian at the $1 - \alpha$ confidence level (see ~~Cook et al., 2002 and section 3.2~~shapiro.test R documentation)

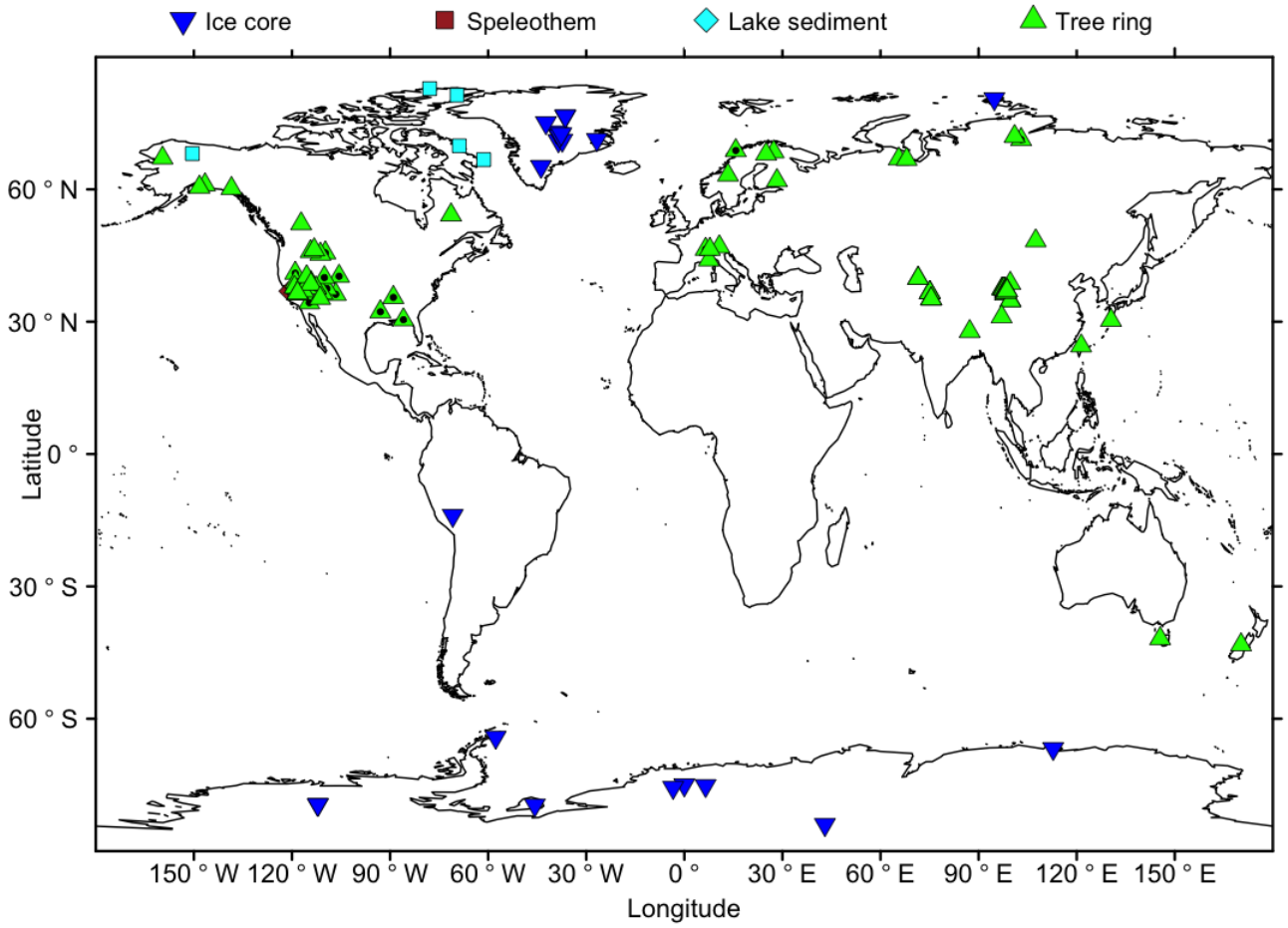

FIGURE 12 – Map of the 108 proxy records used for the reconstruction of the NAO index form Jones et al. ~~The shapes marked by~~ (1997) on the time window 1000-1973 using the RF method. Points with a black ~~circle~~ point are the proxy records also used in Ortega et al. (2015

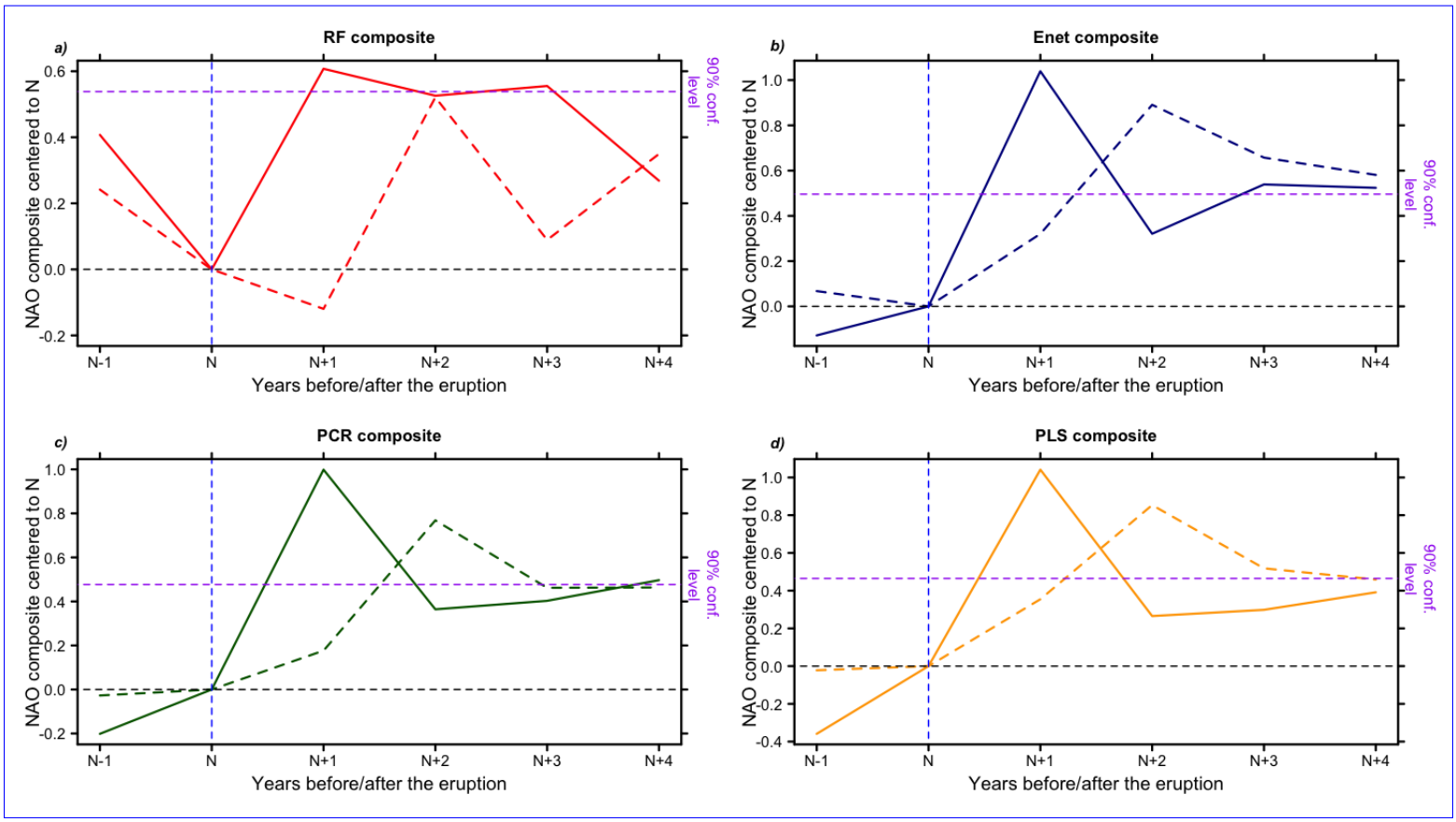

Figure 13 – ~~Composite~~ Superposed epoch analysis of the NAO response from two years (~~N-2~~N-1) before to five years after (N+~~5~~4) ~~ten strong~~ to the largest volcanic eruptions ~~considering 4 volcanic activity reconstructions~~used by Ortega et al. ~~Red lines: Composites from the RF NAO reconstruction. Blue lines: Composites from the Enet NAO reconstruction. Yellow lines: Composites from the PLS NAO reconstruction. Green lines: Composites from the PCR NAO reconstructions. Light lines: composites determined using the Gao et al. (2008~~2015) ~~volcanic activity reconstruction. Light dashed lines: composites determined using the Sigl et al. (2015~~10 eruptions) ~~volcanic activity reconstruction. Heavy lines: composites determined using~~ and the ~~same volcanic activity reconstruction than Ortega et al. (2015)~~11 largest from Sigl et al. ~~Heavy dashed lines: composites determined using the Crowley and Unterman (2013~~2015)~~volcanic activity reconstruction~~. All of the composites are ~~centered~~ centred to their values at the year of the volcanic eruption occurrences. For each method a ~~99%~~ 90% confidence level ~~have~~ has been calculated by Monte-Carlo simulations using 1000 composites of eleven sampled ~~8~~ 6 years long sub-series. The confidence born is calculated as the 99^{th} percentile of the 1000 differences between the ~~5^{th}~~ 4^{th} and the ~~3^{rd}~~ 2^{rd} values of the sample composite series (i.e between N+2 and N). ~~Black dashed lines indicate for each method the 0 level and the 99% confidence level.~~ All of the composite series have been ~~centered~~ centred to the values at the time N. a) Red line: Composite for RF reconstruction response to Sigl et al. (2015) volcanic eruptions. Dashed red line: Composite for RF reconstruction response to Ortega et al. (2015) volcanic eruptions. Dashed purple line: Monte-Carlo 90% confidence level. b) Blue line: Composite for Enet reconstruction response Sigl et al. (2015) volcanic eruptions. Dashed blue line: Composite for Enet reconstruction response to Ortega et al. (2015) volcanic eruptions. Dashed purple line: Monte-Carlo 90% confidence level. c) Green line: Composite for PCR reconstruction response Sigl et al. (2015) volcanic eruptions. Dashed green line: Composite for PCR reconstruction response to Ortega et al. (2015) volcanic eruptions. Dashed purple line: Monte-Carlo 90% confidence level. d) Orange line: Composite for PLS reconstruction response Sigl et al. (2015) volcanic eruptions. Dashed orange line: Composite for PLS reconstruction response to Ortega et al. (2015) volcanic eruptions. Dashed purple line: Monte-Carlo 90% confidence level.

	RF	Enet	~~PLS~~ PCR	PLS	Ortega
RF	1.00	~~0.88~~ 0.79	~~0.83~~ 0.73	~~0.61~~ 0.69	0.52
Enet	~~0.88~~ 0.79	1.00	~~0.82~~ 0.96	~~0.90~~ 0.96	0.68
~~PLS~~ PCR	~~0.79~~ 0.73	~~0.82~~ 0.96	1.00	~~0.88~~ 0.98	~~0.52~~ 0.73
~~PCR~~ PLS	~~0.83~~ 0.69	~~0.90~~ 0.96	~~0.88~~ 0.98	1.00	~~0.66~~ 0.73
Ortega	~~0.61~~ 0.52	~~0.68~~ 0.65	~~0.52~~ 0.73	~~0.66~~ 0.73	1.00

Table 1 – Table of correlations between five reconstructions: Ortega et al. (2015) reconstruction; RF reconstruction on the period 1000-1973 with a proportion of the length of the training samples of ~~55%~~80% ***; Enet reconstruction on the period*** ~~*1000-1973*~~ *1000-1970* ***with a proportion of the length of the training samples of*** ~~70%~~80% ***; PLS reconstruction on the period*** ~~*1000-1967*~~ *1000-1970* ***with a proportion of the length of the training samples of*** ~~70%~~80% ***; PCR reconstruction on the period 1000-1970 with a proportion of the length of the training samples of*** 70%.

Code and data availability: CliMoRec's code and the proxy records database are available at the link: https://github.com/SimMiche/CLIMOREC

Références

Andersen, K., Ditlevsen, P., Rasmussen, S., Clausen, H., Vinther, B., Johnsen, S., and Steffensen, J.: Retrieving a comon accumulation record from Greenland ice cores for the past 1800 years, Journal of geophysical research, 111, D15 106, doi: 0.1029/2005JD006765, 2006.

Andersen, K. K., Bigler, M., Buchardt, S. L., Clausen, H. B., Dahl-Jensen, D., Davies, S. M., Fischer, H., Goto-Azuma, K., Hansson, M. E., Heinemeier, J., Johnsen, S. J., Larsen, L. B., Mischeler, R., Olsen, G. J., Rasmussen, S. O., Röthlisberger, R., Ruth, U., Seierstad, I. K., Siggaard-Andersen, M.-L., Steffense, J. P., Svensson, A. M., and Vinther, B. M.: Greenland Ice Core Chronology 2005 (GICC05) and 20 year means of oxygen isotope data from ice core NGRIP, URL `https://doi.org/10.1594/PANGAEA.586838`, 2007.

Björklund, J. A., Gunnarson, B. E., Seftigen, K., Esper, J., and Linderholm, H. W.: Blue intensity and density from northern Fennoscandian tree rings, exploring the potential to improve summer temperature reconstructions with earlywood information, Clim. Past., 10, 877–885, doi: 10.5194/cp-10-877-2014, 2014.

Booth, B. B. B., Dunstone, N. J., Halloran, P. R., Andrews, T., and Bellouin, N.: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability, Nature, 484, 228–233, doi: 10.1038/nature10946, 2012.

Bradley, R. S.: Climate of the last millenium, 2003.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, 2001.

Browne, M. W.: Cross-Validation Methods, Astronomy, Astrophysics, 44, 108–132, 2000.

Bunn, A. G., Graumlich, L. J., and Urban, D. L.: Trends in twentieth-century tree growth at high elevations in the Sierra Nevada and White Mountains, USA, The Holocene, 15, 481–488, doi: 10.1191/0959683605hl827rp, 2005.

Büntgen, U., Franck, D. C., Nievergelt, D., and Esper, J.: Summer Temperature Variations in the European Alps, a.d. 755-ĂŞ2004, Journal of Climate, 19, 5606–5623, 2006.

Cahill, N., Kemp, A. C., Horton, B. P., and Parnell, A. C.: A Bayesian hierarchical model for reconstructing relative sea level: from raw data to rates of change, Climate of the Past, 12, 525–542, 2016.

Casado, M., Ortega, P., Masson-delmotte, V., Risi, C., Swingedouw, D., Daux, V., Genty, D., Maignan, F., Solomina, O., Vinther, B., Viovy, N., and Yiou, P.: Impact of precipitation intermittency on NAO-temperature signals in proxy records, Climate of the Past, 9, 871–886, doi: 10.5194/cp-9-871-2013, 2013.

Cook, E. R., D'Arrigo, R. D., and Mann, M. E.: A Well-Verified, Multiproxy Reconstruction of the Winter North Atlantic Oscillation Index since A.D. 1400, Journal of Climate, 15, 1754–1764, 2002.

Crowley, T. J.: Causes of climate change over the past 1000 years, Science, 289, 270–277, 2000.

Crowley, T. J. and Unterman, M. B.: Technical details concerning development of a 1200 yr proxy index for global volcanism, Earth System Sciences, 5, 187–197, 2013.

Cuffey, K. M., Clow, G. D., Alley, R. B., Stuiver, M., Waddington, E. D., and Saltus, R. W.: Large Arctic temperature change at the Wisconsin-Holocene glacial transition, Science, 270, 455–458, 1995.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, 39, 1–38, 1977.

Dickson, R., Lazier, J., Meincke, J., Rhines, P., and Swift, J.: Long-term coordinated changes in the convective activity of the North Atlantic, Progress in Oceanography, 38, 241–295, doi: 10.1016/S0079-6611(97)00002-5, 1996.

Drinkwater, K. F., Belgrano, A., Borja, A., Conversi, A., Edwards, M., Greene, C. H., Ottersen, A., Pershing, J., and Walker, H. A.: The North Atlantic Oscillation : Climate significance and meteorological impacts, in: The response of marine ecosystems to climate variability with the North Atlantic Oscillation, edited by Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M., chap. 10, American Geophysical Union, doi: 10.1029/134GM10, 2003.

Esper, J., Frank, D., Büntgen, U., Verstege, A., Luterbacher, J., and Xoplaki, E.: Long-term drought severity variations in Morocco, Geophysical research letters, 34, L17 702, doi: 10.1029/2007GL030844, 2007.

Etheridge, D. M., Steele, L. P., Langenfelds, R. L., and Francey, R. J.: Natural and anthropogenic changes in atmospheric CO 2 over the last 1000 years from air in Antarctic ice and firn, Journal of Geophysical Research, 101, 4115–4128, 1996.

Evan, A. T., Vimont, D. J., Heidinger, A. K., Kossin, J. P., and Bennartz, R.: The Role of Aerosols in the Evolution of Tropical North Atlantic Ocean Temperature Anomalies, Science, 324, 778–781, doi: 10.1126/science.1167404, 2009.

Evan, A. T., Foltz, G. R., Zhang, D., and Vimont, D. J.: Influence of African dust on oceanâĂŞatmosphere variability in the tropical Atlantic, Nature Geoscience, 4, 762–765, doi: 10.1038/NGEO1276, 2011.

Fisher, D. A., Koerner, R. M., and Reeh, N.: Holocene climatic records from Agassiz Ice Cap, Ellesmere Island, NWT, Canada, The Holocene, 5, 19–24, 1995.

Gao, C., Robock, A., and Ammann, C.: , Volcanic forcing of climate over the past 1500 years: an improved ice core-based index for climate models, 113, D23 111, 2008.

Geisser, S.: The predictive sample reuse method with applications, Journal of the Royal Statistical Society, 70, 320–328, 1975.

George, S. S. and Nielsen, E.: Hydroclimatic Change in Southern Manitoba Since A.D. 1409 Inferred from Tree Rings, Quaternary Research, 58, 103–111, doi: 0033-5894/02, 2002.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, 2007.

Graumlich, L. J., Pisaric, M. F. J., Waggoner, L. A., Littell, J. S., and King, J. C.: Upper Yellowstone river flow and teleconnections with Pacific basin climate variability during the past three centuries, Climatic change, 59, 245–262, 2003.

Gray, S. T., Graumlich, L. J., Betancourt, J. L., and Pederson, G. T.: A tree-ring based reconstruction of the Atlantic Multidecadal Oscillation since 1567 A.D., Geophysical Research Letters, 31, 1–4, doi: 0.1029/2004GL019932, 2004.

Graybill, D. A.: International Tree-ring Data Bank NV516, URL `https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring`, 1994a.

Graybill, D. A.: International Tree-ring Data Bank NV517, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 1994b.

Graybill, D. A.: International Tree-ring Data Bank UT508, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 1994c.

Graybill, D. A.: International Tree-ring Data Bank UT509, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 1994d.

Guillot, D., Rajaratnam, B., and Emile-Geay, J.: Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Climate of the Past, 9, 324–352, 2015.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Tardif, R., Steiger, N., and Perkins, W. A.: The last millennium climate reanalysis project: Framework and first results, Journal of Geophysical Research: Atmospheres, 121, 6745–6764, 2016.

Hanhijarvi, M., Tingley, M. P., and Korhola, A.: Pairwise Comparisons to Reconstruct Mean Temperature in the Arctic Atlantic Region Over the Last 2000 Years, Climate dynamics, 41, 2039–2060, 2013.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, American Meteorological Society, August, 1095–1107, doi: 10.1175/2009BAMS2607.1, 2009.

Hawkins, E., Ortega, P., and Suckling, E.: Estimating Changes in Global Temperature since the Preindustrial Period, Journal of Climate, September, 1841–1856, doi: 10.1175/BAMS-D-16-0007.1, 2017.

Hegerl, G. C., Crowley, T. J., Allen, M., Hyde, W. T., Pollack, H. N., Smerdson, J., and Zorita, E.: Detection of Human Influence on a New, Validated 1500-Year Temperature Reconstruction, Journal of Climate, 20, 650–666, doi: 10.1175/JCLI4011.1, 2007.

Helama, S., Holopainen, J., Timonen, M., and Mielikäinen, K.: An 854-Year Tree-ring chronology of Scots Pine for South-West Finland, Studia Quaternaria, 31, 61–68, doi: 10.2478/squa-2014-0006, 2014.

Hoerl, A. E. and Kennard, R. W.: Ridge regression : Biased estimation of nonorthogonal problems, Technometrics, 12, 55–67, 1970.

Homrighausen, D. and McDonald, D. J.: Leave-one-out cross-validation is risk consistent for lasso, Machin Learning, 97, 65–78, doi: 10.1007/s10994-014-5438-z, 2014.

Hotelling, H.: Analysis of a complex of statistical variables into Principal Components, Journal of Education Psychology, 24, 498–520, 1933.

Hotelling, H.: The relations of the newer multivariate statistical methods to factor analysis, British Journal of Statistical Psychology, 10, 69–76, 1957.

Hurrell, J. W.: Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation, Science, 269, 676–679, 1995.

Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M.: An overview of the North Atlantic Oscillation, Geophysical Monograph, 134, 1–35, doi: 10.1029/134GM01, 2003.

Jones, P. D., Jonsson, T., and Wheeler, D.: Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland, International Journal of Climatology, 17, 1433–1450, doi: 0899-8418/97/131433-18\$17.50, 1997.

Karspeck, A. R., Stammer, D., Kohl, A., ..., and Rosati, A.: Comparison of the Atlantic meridional overturning circulation between 1960 and 2007 in six ocean reanalysis products, Journal of Climate, 26, 7392–7413, 2015.

Khodri, M., Izumo, T., Vialard, J., Janicot, S., Cassou, C., Lengaigne, M., Mignot, J., Gastineau, G., Guilyardi, E., Lebas, N., Robock, A., and McPhaden, M. J.: Tropical explosive volcanic eruptions can trigger El Niño by cooling tropical Africa, Nature Communications, 8, No. 778, doi: 10.1038/s41467-017-00755-6, 2017.

Kohavi, R.: A study of Cross-Validation and Boostrap for Accuracy Estimation and Model Selection, 1995.

Kosaka, Y. and Xie, S.-p.: Recent global-warming hiatus tied to equatorial Pacific surface cooling, Nature, 501, 403–407, doi: 10.1038/nature12534, 2013.

Lamb, H. H.: The early medieval warm epoch and its sequel, Paleogeography, Paleoclimatology, Paleoecology, 1, 13–37, 1965.

Lehner, F., Raible, C. C., and Stocker, T. F.: Testing the robustness of a precipitation proxy-based North Atlantic Oscillation reconstruction, Quaternary Science Reviews, 45, 85–94, 2012.

Li, J., Xie, S., Cook, E. R., Morales, M. S., Christie, N. C. J., Chen, F., D'Arrigo, R., Fowler, A. M., and Gou, X.: El Niño modulations over the past seven centuries, Nature climate change, 3, 822–826, 2013.

Lindholm, M. and Jalkanen, R.: Subcentury scale variability in height-increment and tree-ring width chronologies of Scots pine since AD 745 in northern Finland, The Holocene, 22, 571–577, doi: 10.1177/0959683611427332, 2011.

Luterbacher, J., Werner, J. P., Smerdon, J. E., Fernández-Donado, L., González-Rouco, F. J., Barriopedro, D., Ljungqvist, F. C., Bertolin, C., Bothe, O., Brázdil, R., Camuffo, D., DobrovolnyÌĄ, P., Gagen, M., Garciá-Bustamante, E., Ge, Q., Gómez-Navarro, J., Guiot, J., Hao, Z., Hegerl, G. C., Holmgren, K., Klimenko, V. V., Martin-Chivelet, J., Pfister, C., Roberts, N., Schindler, A., Schurer, A., Solomon, O., von Gunten, L., Wahl, E., Wanner, H., Welter, O., Xoplaki, E., Yuan, N., Zanchettin, D., Zhang, H., and Zerefos, C.: European summer temperatures since Roman times, Environnemental Research Letters, 11, 1–12, doi: 10.1088/1748-9326/11/2/024001, 2016.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, PNAS, 35, 13 252–13 257, 2008.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Amman, C., Faluvegi, G., and Ni, F.: Global signature and dynamical orignis of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, doi: 10.1126/science.1177303, 2009.

Maxwell, R. S., Hessl, A. E., Cook, E. R., and Pederson, N.: A multispecies tree ring reconstruction of Potomac River streamflow (950-2001), Water resources research, 47, W05 512, doi: 10.1029/2010WR010019, 2011.

McCabe-Glynn, S., Johnson, K. R., Strong, C., Berkelhammer, M., Sinhan, A., Cheng, H., and Edwards, R. L.: Variable North Pacific influence on drought in southwestern North America, Nature Geoscience, 6, 617–621, doi: 10.1038/ngeo1862, 2013.

McCarroll, D., Loader, N. J., Jalkanen, R., Gagen, M. H., Hakan Grudd, H., and Gunnarson, B. E.: Fennoscandia 1200 Year Tree Growth Data and Summer Temperature Reconstruction, The Holocene, 23, 471–484, 2013.

Meeker, L. D. and Mayewski, P. A.: A 1400-year high-resolution record of atmospheric circulation over the North Atlantic and Asia, The Holocene, 12, 257–266, 2002.

Mignot, J., Khodri, M., Frankignoul, C., and Servonnat, J.: Volcanic impact on the Atlantic ocean over the last millenium, Clim. Past. Discuss., 7, 2511–2554, doi: 10.5194/cpd-7-2511-2011, 2011.

Mitchell, J. M. J., Dzerdzeevskii, B., Flohn, H., Hofmeyr, W. L., Lamb, H. H., Rao, K. N., and Wallén, C. C.: Climatic change: Technical note No. 79, report of a working group for the commission of climatology, World Meteorologicl Organization, Geneva, Switzerland, 1966.

Mysterud, A., Stenseth, N. C., Yoccoz, N. G., Langvatn, R., and Steinheim, G.: Nonlinear effects of large-scale climatic variability on wild and domestic herbivores, Nature, 410, 1096–1099, doi: 10.1038/35074099, 2001.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I? A discussion of principles, Journal of climatology, 10, 282–290, 1970.

Naurzbaev, M. M., Vaganov, E. A., Sidorova, O. V., and Schweingruber, F. H.: Summer temperatures in eastern Taimyr inferred from a 2427-year late-Holocene tree-ring chronology and earlier floating series, The Holocene, 12, 727–736, doi: 10.1191/0959683602hl586rp, 2002.

Neelin, J. D., Anthony, S. B., Hirst, A. C., Jin, F.-f., Wakata, Y., Yamagata, T., and Zebiak, S. E.: ENSO theory, Journal of Geophysical Research, 103, 14 261–14 290, doi: 0148-0227/98/97JC-03424509.00, 1998.

Ortega, P., Lehner, F., Swingedouw, D., Masson-Delmotte, V., Raible, C. C., Casado, M., and Yiou, P.: A model-tested North Atlantic Oscillation reconstruction for the past millennium, Nature, 523, 71–74, doi: 10.1038/nature14518, 2015.

Pages 2K Consortium: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, doi: 10.1038/NGEO1797, 2013.

Pages 2K Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Scientific Data, 4, doi: 10.1038/sdata.2017.88, 2017.

Pearson, K.: On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2, 559–572, 1901.

Poole, M. A. and O'Farrell, P.: The assumption of the linear regression model, in: Transactions of the Institute of the British Geographers, p. 145, doi: 10.2307/621706, 1971.

Preisendorfer, R. W.: Rule N, in: Principal Components Analysis in Meteorology and Oceanography, chap. 3.d., pp. 199–204, 1988.

Reynolds, D. J., Scourse, J. D., Halloran, P. R., Nederbragt, A. J., Wanamaker, A. D., Butler, P. G., Richardson, C. A., Heinemeier, J., Eiriksson, J., Knudsen, K. L., and Hall, I. R.: Annually resolved North Atlantic marine climate over the last millennium, Nature Communications, 7, doi: 10.1038/ncomms13502, 2016.

Salzer, M. W. and Kipfmueller, K. F.: Reconstructed Temperature and Precipitation on a Millennial Timescale from Tree-Rings in the Southern Colorado Plateau, U.S.A, Climatic change, 70, 465–487, 2005.

Santer, B. D., Bonfils, C., Painter, J. F., Zelinka, M. D., Mears, C., Solomon, S., Schmidt, G. A., Fyfe, J. C., Cole, J. N. S., Nazarenko, L., Taylor, K. E., and Wentz, F. J.: Volcanic contribution to decadal changes in tropospheric temperatures, Nature Geoscience, 7, 185–189, doi: 10.1038/ngeo2098, 2014.

Schneider, L., Smerdson, J. E., Büntgen, U., Wilson, R. J. S., Myglan, V., Kirdyanov, A. V., and Esper, J.: Revising midlatitude summer temperatures back to A.D. 600 based on a wood density network, Geophysical Research Letters, 42, doi: 10.1002/2015GL063956, 2015.

Schneider, T.: Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values, Journal of Climate, 14, 853–871, 2001.

Schweingruber, F. H.: International Tree-ring Data Bank SWIT177, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 2007.

Seidenglanz, A., Prange, M., Varma, V., and Schulz, M.: Ocean temperature response to idealized Gleissberg and de Vries solar cycles in a comprehensive climate model, Geophysical Research Letters, 39, 1–6, doi: 10.1029/2012GL053624, 2012.

Shindell, D. T., Schmidt, G. A., Mann, M. E., and Faluvegi, G.: Dynamic winter climate response to large tropical volcanic eruptions since 1600, Journal of Geophysical Research, 109, D05 104, doi: 10.1029/2003JD004151, 2004.

Sigl, M., Winstrup, M., McConnell, J. R., ..., and Woodruff, T. E.: Timing and climate forcing of volcanic eruptions for the past 2,500 years, Nature, 523, 543–549, 2015.

Singh, H. K. A., Hakim, G. J., Tardif, R., Emile-Geay, J., and Noone, D. C.: Insights into Atlantic multidecadal variability using the Last Millennium Reanalysis framework, Journal of Geophysical Research: Atmospheres, 14, 157–174, 2018.

Stahle, D. K., Burnette, D. J., and Stahle, D. W.: A Moisture Balance Reconstruction for the Drainage Basin of Albemarle Sound, North Carolina, Estuaries and Coasts, 36, 1340–1353, doi: 10.1007/s12237-013-9643-y, 2013.

Stahle, D. W.: International Tree-ring Data Bank AR050, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 1996a.

Stahle, D. W.: International Tree-ring Data Bank LA001, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 1996b.

Stahle, D. W. and Cleaveland, M. K.: International Tree-ring Data Bank AR052, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 2005a.

Stahle, D. W. and Cleaveland, M. K.: International Tree-ring Data Bank FL001, URL https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring, 2005b.

Stahle, D. W., Villanueva Diaz, J., Brunette, D. J., Cerano Paredes, J., Heim Jr., R. R., Fye, F. K., Acuna Soto, R., Therell, M. D., Cleaveland, M. K., and Stahle, D. K.: Major Mesoamerican droughts of the past millennium, Geophysical research letters, 38, L05 703, doi: 10.1029/2010GL046472, 2011.

Stein, M. L.: Equivalent of Gaussian Measures and Prediction, in: Interpolation of Spatial Data: Some Theory for Kriging, chap. 4, p. 179, 1999.

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Climate Change 2013, The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2013.

Stone, M.: Cross-Validatory choice and assesment of statistical predictions, Journal of the Royal Statistical Society, 36, 111–147, 1974.

Swingedouw, D., Terray, L., Cassou, C., Voldoire, A., Salas-Mélia, D., and Servonnat, J.: Natural forcing of climate during the last millennium: fingerprint of solar variability, Climate Dynamics, 36, 1349–1364, doi: 10.1007/s00382-010-0803-5, 2011.

Swingedouw, D., Ortega, P., Mignot, J., Guilyardi, E., Masson-delmotte, V., Butler, P. G., Khodri, M., and Séférian, R.: Bidecadal North Atlantic ocean circulation variability controlled by timing of volcanic eruptions, Nature Communications, 6, No. 6545, doi: 10.1038/ncomms7545, 2015.

Swingedouw, D., Mignot, J., Ortega, P., Khodri, M., Menegoz, M., Cassou, C., and Hanquiez, V.: Impact of explosive volcanic eruptions on the main climate variability modes, Global and Planetary Change, 150, 24–45, doi: 10.1016/j.gloplacha.2017.01.006, 2017.

Tibshirani, R.: Regression shrinkage and selection via Lasso, Journal of the Royal Statistical Society, 58, 267–288, doi: 0035-9246/96/58267, 1996.

Tingley, M. P.: A Bayesian ANOVA Scheme for Calculating Climate Anomalies, with Applications to the Instrumental Temperature Record, Journal of Climate, 25, 777–791, 2012.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems, Journal of Climate, 23, 2759–2781, 2010a.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation ?Maximization Algorithm, Journal of Climate, 23, 2782–2800, 2010b.

Tingley, M. P. and Huybers, P.: Recent temperature extremes at high northern latitudes unprecedented in the past 600 years, Nature, 496, 201–5, 2013.

Tosh, R.: International Tree-ring Data Bank CA051, URL `https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring`, 1994.

Touchan, R., Garfin, G. M., Meko, D. M., Funkhouser, G., Erkan, N., Hughes, M. K., and Wallin, B. S.: Preliminary reconstructions of spring precipitation in southwestern Turkey from tree-ring width, International journal of climatology, 23, 157–171, doi: 10.1002/joc.850, 2003.

Touchan, R., Woodhouse, C. A., Meko, D. M., and Allen, C.: Millennial precipitation reconstruction for the Jemez Mountains, New Mexico, reveals changing drought signal, International journal of climatology, 31, 896–906, 2011.

Trenberth, K. E. and Fasullo, J. T.: Atlantic meridional heat transports computed from balancing EarthâĂŹs energy locally, Geophysical Research Letters, 44, 1919–1927, doi: 10.1002/2016GL072475, 2017.

Trenberth, K. E. and Shea, D. J.: Atlantic hurricanes and natural variability in 2005, Geophysical Research Letters, 33, 1–4, doi: 10.1029/2006GL026894, 2006.

Trouet, V., Esper, J., Graham, N., Baker, A., Scourse, J., and Frank, D.: Persistent positive North Atlantic oscillation mode dominated the Medieval Climate Anomaly, Science, 324, 78–80, 2009.

Vapnik, V. N.: The Nature of Statistical Learning Theory, Statistics for Engineering and Information Science, 2000.

Vieira, L. E. A., Solanki, S. K., Krivova, N. A., and Usoskin: Evolution of the solar irradiance during the Holocene, Astronomy, Astrophysics, 531, A6, 2011.

Vinther, B. M., Andersen, K. K., and Hansen, A. W.: Improving the Gibraltar/Reykjavik NAO index, Geophysical Research Letters, 30, 1–4, doi: 10.1029/2003GL018220, 2003.

Vinther, B. M., Clausen, B., Fisher, D. A., Koerner, R. M., Johnsen, S. J., Andersen, K. K., D, D.-J., Rasmussen, S. O., Steffensen, J. P., and Svensson, A. M.: Synchronizing ice cores from the Renland and Agassiz ice caps to the Greenland Ice Core Chronology, Journal of Geophysical Research, 113, D08 115, 2008.

Vinther, B. M., Jones, P. D., Briffa, K. R., Clausen, H. B., Andersen, K. K., D, D.-J., and Johnsen, S. J.: Climatic signals in multiple highly resolved stable isotope records from Greenland, Quaternary Science Reviews, 29, 26,455–26,470, 2010.

Visbeck, M., Chassignet, E. P., Curry, R. G., Delworth, T. L., Dickson, R. R., and Krahmann, G.: The Ocean's Response to North Atlantic Oscillation Variability, in: The North Atlantic : Climatic Significance and Environmental impacts, edited by Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M., doi: 10.1029/134GM06, 2003.

Wang, J., Emile-Geay, J., Guillot, D., Smerdson, J. E., and Rajaratnam, B.: Statistical paleoclimate reconstructions via Markov random fields, PNAS, 10, 1–19, 2014.

Wang, J., Yang, B., Ljungqvist, F. C., Luterbacher, J., Osborn, T. J., Briffa, K. R., and Zorita, E.: Internal and external forcing of multidecadal Atlantic climate variability over the past 1,200 years, Nature Geoscience, 2017.

Wilson, R., Miles, D., Loader, N. J., Cooper, R., and Briffa, K.: A millennial long March-July precipitation reconstruction for southern-central England, Climate Dynamics, doi: 10.1007/s00382-012-1318-z, 2013.

Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J.: The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses, J. Sci. Stat. Comput., 5, 735–743, 1984.

Woodhouse, C. A. and Brown, P. M.: Internation Tree-ring Data Bank CO572, URL `https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring`, 2006.

Young, G. H. F., McCarroll, D., Loader, N. J., Gagen, M., Kirchhefer, A. J., and Demmler, J. C.: Changes in atmospheric circulation and the Arctic Oscillation preserved within a millennial length reconstruction of summer cloud cover from northern Fennoscandia, Climate Dynamics, 39, 495–507, doi: 10.1177/0959683609351902, 2012.

Zhang, P., Linderholm, H. W., Gunnarson, B. E., Björklund, J. A., and Chen, D.: 1200 years of warm-season temperature variability in central Scandinavia inferred from tree-ring density, Clim. Past., 12, 1297–1312, doi: 10.5194/cp-12-1297-2016, 2016.

Zhang, Y. and Yang, Y.: Cross-validation for selecting a model selection procedure, Journal of Econometrics, 187, 95–112, doi: 10.1016/j.jeconom.2015.02.006, 2015.

Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, 67, 301–320, 2005.