Reconstructing climatic modes of variability from proxy records using ~~CliMoRec~~ ClimIndRec version 1.0

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

21 août 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, 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 useful for improving our understanding of their behaviours. For doing so, several statistical methods exist, among which the Principal Component Regression is one of the most widely used in paleoclimatology. Here, we provide the software ClimIndRec to the climate community~~the computer device CliMoRec,~~, it is based on four regression methods (PCR, Partial Least Squares, Elastic Net and Random Forest) and cross validation algorithms, ~~that~~ and enables systematic reconstruction of a given climate ~~mode~~ index. A prerequisite is that there are proxy records in the database that overlap in time with its observed ~~timeseries~~variations. The relative efficiency of the methods can vary, according to the statistical properties of the mode and the proxy records used~~, thereby allowing to assess sensitivity related~~. Here, we assess sensitivity to the reconstruction technique. ~~CliMoRec~~ ClimIndRec is modular as it allows different inputs like the proxy database or the regression method. As an example, it is here applied to the reconstruction of the North Atlantic Oscillation by using ~~Pages 2K~~ PAGES 2k database. In order to identify the most reliable reconstruction among those given by the different methods, we use the modularity of ~~CliMoRec~~ ClimIndRec 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~~targeted reconstruction period. The best reconstruction of the NAO that we obtain is using the Random Forest approach. It shows significant correlation with former reconstructions, but exhibits ~~better~~ higher validation scores.

1   Introduction

The interdependent components of the climate system, such as the atmosphere and the ocean, vary at different timescales. The interactions between those components [Mitchell et al., 1966] lead the climate to vary from the hourly to the multidecadal timescales. Preindustrial control simulations of global coupled climate models 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]. External factors such as volcanic aerosols [Mignot et al., 2011; Swingedouw et al., 2015; Khodri et al., 2017], 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], also influence the variations and dynamics of the climate system by altering the Earth's radiation balance. By only considering the impact of external forcings which are not due to the human activity, we can characterise the so-called natural climate variability.

An unequivocal synchronous rise in both the greenhouse gas concentration in the atmosphere and the global mean temperature has been observed in instrumental measurements [Stocker et al., 2013]. However for temperatures, fluctuations around this trend from a decade to another [Kosaka and Xie, 2013; Santer et al., 2014; Swingedouw et al., 2017] highlight the modulating role of natural variability at decadal to multidecadal scales. Thereby, improving our knowledge about past natural climate variability and its sources is essential to better understand the potential coming changes in climate.

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. 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 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 then affect the poleward heat transport and the related 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 relatively short duration of the instrumental records, which prevents robust statistical evaluation of their properties (e.g. spectrum, stability of teleconnections, underlying mechanisms). To partly overcome this limitation, reconstructions of climate beyond the period of direct measurements have been performed in numerous studies that combine appropriate statistical methods and information from proxy records. Proxy records provide indirect

estimates of past local or regional climate, derived from natural archives coming for instance from sediment cores, speleothems, ice cores or tree rings. According to its nature, each proxy record has a specific temporal resolution, from years to millennia, and 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].

Based on the assumption that climate modes such as the NAO affect climate conditions in different locations, some studies have used 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~~ Luterbacher et al. (2001) firstly proposed a ~~complete methodology of nested~~ partly monthly and seasonal reconstruction of the NAO that extends back to 1500 using the Principal Component Regressions ~~(PCRs)~~ [Hotelling, 1957] (PCR) method. Another study have reconstructed the SLP fields in Europe further back to 1500 using a PCR approach as well [Luterbacher et al., 2002], and have found consistencies with the Luterbacher et al. (2001) NAO reconstruction. Cook et al. (2002) have also proposed a complete methodology of nested PCR using annually resolved proxy records bounding the North Atlantic to reconstruct the NAO variability further back to 1400. More recently, 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 historical NAO index on their common time window. Instead of nesting reconstructions of different sizes, which 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. Regression-based methods have also been used for reconstructing other climate modes indices than NAO, such as for instance El-Niño Southern Oscillation index [Li et al., 2013] and the Atlantic Multidecadal Variability index [Gray et al., 2004; Wang et al., 2017].

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 present the computer ~~device CliMoRec (Climate Mode~~ tool ClimIndRec (Climate Index Reconstruction) version 1.0, which includes multiple statistical approaches, for reconstructing climate modes indices. It is based on four regression methods: the PCR [Hotelling, 1957], the Partial Least Squares regression (PLS) [Wold et al., 1984], the Elastic-net regression (E-net) [Zou and Hastie, 2005] and the Random Forest (RF) [Breiman, 2001]. It communicates with a large proxy database, that contains various types of proxy records distributed worldwide and which are sensitive to different climate variables. ~~CliMoRec~~ ClimIndRec is thus designed to reconstruct the past variability of different climate modes (Fig ~~.~~ 1). It should be stressed that ~~CliMoRec~~ ClimIndRec 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~~ ClimIndRec can also be used to reconstruct other kinds of climate ~~time-series~~ timeseries such as temperatures or precipitations in a given location.

~~In section~~ Section 2 ~~, the database and some general statistical tools are first presented. The reconstruction methods, are then described in a mathematical formalism in section 3. Section 4 compares these~~ develops the rationale and the value-added of ClimIndRec for climate timeseries reconstruction. Section 3 compares the four regression methods by reconstructing the NAO index over the last millenium 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 training samples~~ or the proxy records selection for regression. Final

section ~~5~~ 4 presents a discussion including some outlooks for next version of ~~CliMoRec~~ ClimIndRec and the conclusions of this study.

2 Data, notations and methodologies

2.1 ~~Data~~ General methodology of ClimIndRec

~~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~~ In the present study, we aim at reconstructing a climatic timeseries assuming proxy records as predictors. As an example, building a PCR model consists in determining the Principal Component of the proxies and finding the best linear combination of them to reconstruct the climate index over the training period. Then, the reconstruction consists in projecting the first seen data on the built orthogonal basis, and applying the estimated regression coefficients to reconstruct the climate index over the whole reconstruction window.

We here compare four models that all consist in regression methods among which the PCR has been used in many paleoclimate studies [Luterbacher et al., 2001 ; Luterbacher et al., 2002 ; Cook et al., 2002 ; Gray et al., 2004 ; Ortega et al., 2015 ; Wang et al., 2017]. The methods we added (PLS, Enet and RF) aim at exploring alternative approaches than PCR and comparing different reconstructions using relevant metric. PLS is a similar approach to PCR where the difference is that the matrix of Empirical Orthogonal Functions (EOFs) is calculated by maximising the variance of the projected proxies on the EOFs and the targeted climate index instead of the variance of the projected proxies [Wold et al., 1984. Enet belongs to the regularized regression methods family that are not usually used in paleoclimate reconstructions [Zou and Hastie, 2005]. It is here investigated in order to find if this kind of regression approach is relevant for climate index reconstruction. Finally, the RF method is an aggregation of multiple predictors called "regression trees", which are non-linear regression approaches [Breiman, 2001]. The mathematical details for each method are developed in appendix 1. To put ClimIndRec and this study in a wider context of regression-based approaches, the following rationale to evaluate and thus compare reconstructions would be available for any alternative regression method and can thus be extrapolated to the other existing ones. Hence, updates of ClimIndRec will be dedicated to propose other regression methods such as lasso regression [Tibshirani, 1996] or adaptive lasso regression [Zou, 2006]. ~~This index is indeed relatively simple to calculate from SLP time series as it only requires two locations with instrumental records: one within the center of action of~~

In the case of the reconstruction of climate indices, regression methods seek to establish for each common time step the relationships between the proxies and the climate index to be reconstructed over the period of instrumental measurements. This set of relationships constitutes a statistical model of the considered climate index. The paleo-variations of proxy records are then translated into a climate index in the past using the relationships previously established by the statistical model. Since each regression method can be used in different ways, they must be optimized to make the reconstruction as robust as possible. In the case of PCR, for example, the number of principal components of the proxies used to regress the climate index directly affects the reconstruction since it modifies the set of predictors. The term "hyperparameter" is used to design this ensemble of parameters inherent to each method. They are identified for each method in appendix 1. Their tuning (or optimization) using cross-validation technics [Stone, 1974 ; Geisser, 1975] are developed later in this section.

Reconstruction of a same climate index obtained from different regression methods may significantly differ. Thus, if the same index is reconstructed using different regression methods that each suggest different interpretations of the past, it may be difficult to compare them directly. A common approach is to separate the observation years (called learning period) in two to evaluate a statistical model. The first period, called the training (or calibration) period, is used to build the model using hyperparemeter tuning, and thus to establish relationships between the climate index and proxies. The proxies of the second period, called the testing (or validation) sample, are then translated into a climate index over the years of observations of this period. The actual values of the ~~Azores anticyclone (typically Gibraltar) and one within the Icelandic low (typically Reykjavik)~~ climate index can then be compared with the reconstructed climate index over the testing period using a given metric which will be defined in section 2.3.2. It gives a score estimating the model ability to reconstruct the climate index using first-seen data of proxies. This procedure is called "hold-out approach" [Sammut et al., 2009]. ~~The reference NAO index 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 periodsince 1856.~~

The scores obtained for different regression methods for a given training/testing sample might be impacted by the specific sampling. This is overcame by repeating the hold-out approach several times where years of observations between the training and the testing samples are shuffled. Thereby an ensemble of scores is obtained, yielding an evaluation of the methods ability to reconstruct the climate index. The most robust regression model is the one that has the highest scores as it means this is the most accurate to reconstruct the climate index using first-seen data of proxies. This most robust regression method is then applied to the whole learning period to build a final model and infer the paleo-variations of the climate index from proxy records. In our study, and by default in ClimIndRec, the determination of the testing samples is performed using a block-style approach over time. This means that the first testing period of a given size encompasses the first time steps of the learning period. This testing period is then shifted by one time step which gives the second testing period of same size. And so on until each time step of the learning period has been used at least once for testing. The reason is that for climate timeseries, autocorrelations are often large, so that one get skills from persistence alone. Thus sampling is usually used with a block-style approach for climate timeseries.

~~In terms of proxies, we use the state-of-the-art Pages 2k database~~ The reconstruction might also largely differ for a same reconstruction method according to both the proxy records used and the years of observations used. Here, the sources of uncertainties associated with the proxy selection as well as the learning period used can be reduced using the same hold-out approach with evaluation and comparison of the most optimal sets using scores.

The number of proxy records and the reconstruction period are thus fixed for the different training/testing period sections and the final model, in contrast with some previous studies which used nested approaches [~~Pages 2K Consortium, 2017~~Cook et al., 2002 ; Wang et al., 2017]~~in its latest 2017 version (P2k2017). The proxy records which resolutions lower than annual have been removed. Even if these proxy records could be interpolated to a finer temporal scale and used for the reconstruction, their use is not recommended~~. 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, ClimIndRec can be used to perform reconstructions on different time windows which can then be aggregated to perform a nested reconstruction, with associated scores for each portion of time.

It should be stressed that the approach of ClimIndRec is implicitly assuming that the climate index to reconstruct is a linear combination of the proxy records. It means assuming that the climate reacts to proxies while the correct etiological relationship is the other way around. Hence, it has to be specified that since climate variations affect proxies variations, we can attempt to estimate past climate fluctuations using statistical methods. Another caveat to highlight is that the proxy records used know their own limitations and uncertainties that can come from various sources such as the measurement methodologies, the dating uncertainties or the transfer function used to infer bio/geochemical data into paleoclimate variations. This inevitably leads to an underestimation of the true link between the climate index and the climate variable associated to the proxy record and therefore leads to a biased reconstruction with loss of variance (Isobe et al., 1990). To overcome this issue, previous climate index reconstruction studies [~~Hanhijarvi et al., 2013~~Ortega et al., 2015; Wang et al., 2017] ~~, 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 44 annually-resolved proxy recordsused in Ortega et al. (2015)and not present in P2k2017 (see supplementary informations) . We end up with a database of 554 well-verified and worldwide distributed annually-resolved proxy records.~~ rescaled the variance of the reconstruction according to the observed climate index variance. However it implies that the variance of the climate index is stationary which might not be true. In this study we thus present the raw reconstructions and the loss of variance will be quantified and specified.

ClimIndRec is developed using both bash and R scripts. It uses different R packages (presented supplementary table 5) that could be used independently to blindly perform reconstructions of any climate index. The added-value of ClimIndRec is to integrate synchronous hold-out approach and cross-validation according to the user inputs (proxy records, regression method, reconstruction period targeted, proxy records pre-selection). It therefore allows several inputs to be tested and provides relevant metrics that can be used to determine the optimal regression model.

2.2 ~~Methodology~~Step-by-step rationale for reconstruction and model evaluation

The general reconstruction ~~procedure follows 11~~ and model evaluation procedure follows 12 steps, among which the first three are inputs selection and the others are ~~already implemented in CliMoRec~~implemented in ClimIndRec. These are applied sequentially as follows (Fig ~~.~~1):

1. An observational time series ~~of the~~ representing modulations of the targeted 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 T between the observed climate index and the selected proxy records is identified and extracted for ~~calibrating the reconstruction~~ evaluating the reconstruction method
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~~ following a block-style approach to perform splits, R depending on the size of the learning period and the size of testing periods determined by the user.
7. 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

7

8. Each of the R sets of periods and proxies is calibrated over the training window for all the different ~~statistical parameters of~~ sets of hyperparameters of the given method selected in 3, and the best performing ~~set~~ one is identified
9. The corresponding optimal setup is then applied to extend the reconstruction over the ~~target period $\mathcal{T}$ for each ensemble~~ testing period for each member
10. ~~A validation score is computed for each member by comparing~~ Validation scores are computed by comparing each of the observation-based testing series and each training sample-based individual reconstruction over the corresponding testing period
11. The ~~final reconstruction is calculated as the average of all the individual R reconstructions~~ corresponding hyperparameters are tuned over the whole learning period T and the final model is built
12. The final reconstruction is obtained by applying the final model to the proxies over the reconstruction period $\mathcal{T}$

~~Thus CliMoRec provides the mean reconstruction of the chosen mode~~

Thus ClimIndRec provides the final reconstruction with associated uncertainties (appendix 3) and a vector with ~~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. 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~~

2.3 Mathematical evaluation

This section aims to clarify the technical details of the methodology presented in sections 2.1 and 2.2. It will thus call on the elements mentioned above

2.3.1 Data notation

To 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~~ In this study, 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 $\mathcal{T}$ and it is excluded otherwise (section 2.2, step 2). ~~Following these steps, the proxy record~~, Hence, the proxy records matrix does not contain missing values.

Fig ~~.~~2 illustrates how the proxy data are organised in the input matrix X. We denote $X^1 = (X_t^1)_{t \in \mathcal{T}}, \ldots, X^p = (X_t^p)_{t \in \mathcal{T}}$, 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 grouping the individual records: $X = [X^1, \ldots, X^p]$. $Y = (Y_t)_{t \in T}$ is the target climate index, defined on the historical time window T 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 $N = n + m$ 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 learning set is denoted $\{X_{(T)}, Y\}$, and the reconstruction set is denoted $\{X_{(\mathcal{T})}\}$.

2.4 ~~Terms, notations 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 $X^1, \ldots, X^p$. 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 $\mathcal{T}$.~~

2.3.1 Terms and validation metrics

We denote the chosen reconstruction method by $\mathcal{M}$. Each method is defined by a specific number of ~~parameters~~ hyperparameters 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~~ an hyperparameters vector (θ):

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

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

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 reconstruction $\hat{Y}$ 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~~ are used for different tasks. 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]:

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

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

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

S_{NSCE} ~~will be~~ is used to validate the reconstruction methods over the testing period, and S_{RMSE} ~~will allow~~ allows to determine the optimal ~~parameters~~ hyperarameters (θ) 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]~~.~~, additional informations about this metric are presented in the appendix 2. 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.

2.4   ~~Final reconstruction and parameter tuning~~

~~We split~~

2.3.1   Hyperparameter tuning by cross-validation and final reconstruction

As mentioned above, the initial learning ~~period T in~~ R partitions of two subsets: $\left\{ T^{(r)}_{(train)}, T^{(r)}_{(test)} \right\}, \forall 1 \leq r \leq R$ (section 2.2, step 6). 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~~ hyperparameters θ_{opt} on a given training set $\{X_{train}, Y_{train}\}$, we use the K-fold cross validation approach (KFCV; section 2.2, step 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]. Here, it is used as an optimization method to empirically determine an optimal set of ~~parameters~~ hyperparameters for θ. 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 kth drawn sample. Then, $\{X_{(-k)}, Y_{(-k)}\}$ is the set containing all the K-1 other sets. For all possible values of θ contained in Θ, we scan the K models based on the sets s $\{X_{(-k)}, Y_{(-k)}\}_{1 \leq k \leq K}$. The empirical optimal set of ~~parameters~~ hyperparameters is obtained by minimizing the averaged S_{RMSE} functions on the K splits by considering all possible combinations of θ [Stone, 1974]. Mathematically, the optimal KFCV set of ~~parameters~~ hyperparameters θ_{KF} is determined by:

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

It should be noted that if $dim\theta > 1$, then the different hyperparameters need to be optimized simultaneously, with nested KFCVs.

Using this approach, we retain the ~~empirical estimation of the optimal set of parameters~~ hyperparameters vector $\hat{\theta}_{opt} = \theta_{KF}$ for the given method $\mathcal{M}$ and a given learning set $\{X, Y\}$. ~~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.~~ 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 $Y^{(r)}_{(test)}$ are compared to the individual reconstructions using the S_{NSCE} function. This way, R NSCE scores are obtained for $\mathcal{M}$. In section ~~4~~3, the distributions of the NSCE scores will be used as a metric to compare different reconstructions. Fig.4 shows the the ~~calculation to get~~ calculations that gives the NSCE scores for a given method $\mathcal{M}$.

3 Regression methods

We present each method in two steps: model fitting (for training) and reconstruction (for testing). For each method the proxy matrix is denoted as $X \in \mathbb{R}^{n \times p}$ the proxy predictor set and the climate index as $Y \in \mathbb{R}^n$. $X_{(rec)} \in \mathbb{R}^{m \times p}$ is the testing dataset from which a $\mathbb{R}^m$ reconstruction is build using the regression method

2.1 Principal Component Regression (PCR)

2.0.1 Modeling

The Principal Component Regression Hotelling, 1957method 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 It should be stressed that the K-fold cross validation sampling is also implemented following a block-style approach in ClimIndRec for the same reasons than for the hold-out approach. This means that de K groups of observations are made along time instead of being randomly split. Also, the covariance matrix of X:

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

We calculate the orthogonal basis formed by the vectors $v_1, \ldots, v_p$ by diagonalizing S:

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

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

$$\ldots$$

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

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

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

11

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

$$\lambda_k = Var(Xv_k) = v_k^T S v_k \qquad \forall 1 \leq k \leq p$$

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

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

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

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

Where ϵ is a white noise vector of size n. The best estimator for $\beta = (\beta_1, \ldots, \beta_q)$, is given by the Ordinary Least Squares (OLS) estimator which minimizes $\|\hat{\epsilon}\| = \|Y - \hat{Y}\|$:

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

2.0.1 Reconstruction

We project the testing matrix $X_{(rec)}$ on the pre-calculated orthogonal basis V:

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

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

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

2.0.1 Parameters

Here, q is the unique tuning parameter. The choice of that parameter clearly affects the reconstruction and then the NSCE scores. Here the parameter vector θ is unidimensional and takes its values in the discrete set $(i)_{1 \leq i \leq p}$ choice of K can have implications for the estimation of optimal hyperparameters. Large K leads to more diverse training samples, thereby bringing more variable estimates of RMSE. On the other hand, small K leads to a low number of samples used, thereby increasing the bias due to the particular way splits have been made. Additional works have shown that this choice is poorly influencing the final reconstruction obtained (not shown) so that we decided to set it to K=5 for this study. It is set at K=5 by default in ClimIndRec but it can certainly be changed in order to produce alternative reconstructions.

2.1 Partial Least Squares Regression

The PCA keeps most of the initial variance in X in a lower number of vectors. But EOFs $v_1, \ldots, v_p$ 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 Least Squares regression (PLSr) method Wold et al., 1984. The first latent variable (LV), denoted $\xi_1 = \sum_{j=1}^{p} v_{1,j} X^j = X v_1$, where $X \in \mathbb{R}^{n \times p}$ and $v_1 \in \mathbb{R}^p$ is the linear combination of the initial variables $X^1, \ldots, X^p$ such as:

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

In a similar approach to the PCR, the second LV is $\xi_2 = \sum_{j=1}^{p} v_{2,j} X^j = X v_2$, orthogonal to ξ_1, such as:

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

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

Finding the loadings is not as trivial as for PCR. Indeed the empirical covariance matrix is not necessary definite positive and thus cannot be 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. $X_0 \leftarrow X$ for $h = 1, \ldots, r$ $v_h \leftarrow \frac{X_{h-1}^T Y}{\|X_{h-1}^T Y\|^2}$ $\xi_h \leftarrow X_{h-1} v_h$ $X_h = X_{h-1} - \frac{\xi_h \xi_h^T}{\|\xi_h\|^2} X_{h-1}$ *(deflation phase)* 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:

$$Y = \Psi \beta + \epsilon$$

Where ϵ is a white noise vector of size n. The best estimator for $\beta = (\beta_1, \ldots, \beta_q)$, is given by the Ordinary Least Squares (OLS) estimator which minimizes $\|\hat{\epsilon}\| = \|Y - \hat{Y}_{q_{KF}}\|$:

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

2.0.1 Reconstruction

The 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:

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

Once the model has been evaluated, it is launched over the whole learning set $\{X_{(T)}, Y\}$ with a K-fold cross-validation to optimize the hyperparameters such as done previously for training samples. The reconstruction is obtained by applying the estimated coefficient vector on the sub-matrix $\Psi_{(rec)} = (\xi_{(rec)}^1, \ldots, \xi_{(rec)}^r) \in \mathbb{R}^{m \times r}$:

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

2.0.1 ~~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.~~ thus obtained by inferring proxy records variations in $\{X_{(\mathcal{T})}\}$ into a reconstructed climate index over the reconstruction period $\mathcal{T}$.

2.1 ~~Elastic Net regression~~Data

2.1.1 ~~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:~~

$$Y = X^1\beta_1 + \cdots + X^p\beta_p + \epsilon$$

~~Where $\epsilon \sim \mathcal{N}(0, \sigma^2)$~~ The assessment of the proposed 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. This index is indeed relatively simple to calculate from SLP timeseries as it only requires two locations with instrumental records: one within the center of action of the Azores anticyclone (typically Gibraltar) and ~~$Cov(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$. The reconstruction of Y, given p proxy records $X^1, \ldots, X^p$ is obtained by the equation:~~

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

~~$\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ are the regression coefficients, which are obtained by the OLS predictor. However, this usual regression model is known to 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 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: $\|\beta\|_k^k = \sum_{j=1}^{k} |\beta_j|^k$. With $k = 1$ in Lasso regression, and $k = 2$ in Ridge regression. The penalized loss~~

functions are given by:

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

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

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

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 :

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

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

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

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

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

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

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

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

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

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

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

2.1.1 Reconstruction

The reconstruction is obtained by applying the estimated regression coefficients $\hat{\beta}^{enet}$ on the validation variables $X_{val}^1, \ldots, X_{val}^p$:

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

15

2.1.1 Parameters

For Enet method, the tuning parameters are λ and α. The latter controls the relative balance between the Lasso and Ridge regularization, while the former controls the overall intensity of regularization as λ_1 (resp. λ_2) in Lasso (resp. 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 one within the Icelandic low (typically Reykjavik). The reference NAO index is then calculated as the parameter estimation. The finer these discretizations are, the more reliable the parameters will be, at the expanse of the computational time. normalized SLP difference between these two locations. Jones et al. (1997) have for example proposed an index spanning the whole period since 1856.

2.2 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 reconstruction uncertainty given by regression trees. Random forests encompass a large variety of regression methods

In terms of proxies, we use the state-of-the-art PAGES 2k database [Breiman, 2001Pages 2K Consortium, 2017] . Here, we present the most classical kind of random forests known as random-input random forests Breiman, 2001.

2.1.1 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 ith 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, $\{Y_i, X_i\}_{1 \leq i \leq n}$ in its latest 2017 version (hereafter P2k2017). Proxy records which resolutions are lower than annual were removed. Even if these proxy records could be interpolated to a finer temporal scale and used for the reconstruction, their use is not recommended as the interpolated timeseries will present high auto-correlation coefficients, which could inflate the correlations with the NAO and thus their weight in the final reconstruction, $\forall 1 \leq i \leq p$, 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 :

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

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 jth variable lower than d are placed in the left child node. Hence, all observations with a jth 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 :

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

16

where $\bar{Y}_t$ is the averaged Y_t 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.

2.1.1 Reconstruction

The reconstruction is obtained by splitting each testing series in the different trees. 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 reconstruction.

2.1.1 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 tuning, as long as the number of trees is sufficiently high given p, which guarantees convergent results for any value of m potentially leading to spurious results [Breiman, 2001 Hanhijarvi et al., 2013]. 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. We added 44 annually-resolved proxy records used in Ortega et al. (2015) and not present in P2k2017 (see supplementary informations). We end up with a database of 554 well-verified and worldwide distributed annually-resolved proxy records.

3 Results

3.1 Methodological sources of uncertainty in the reconstruction

We apply CliMoRec ClimIndRec with the four methods presented above to the reconstruction of the NAO. In the following, each reconstruction is obtained by averaging $R = 50$ R individual reconstructions performed for R training/testing random draws. splits. R depends on the size of the testing samples relative to the size of the learning period as we perform block-style splits of the data to produce training and testing splits (section 2.1 and 2.2). Here, we set the relative length of the training splits as 80% of the learning period. NSCE scores are also produced , thus produced and stored in a vector of R 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 identified in section 2.1, 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 two steps are described below, before assessing the reconstruction itself. In this section the training periods length is set to 80% of the length of the training sample and the testing periods length is set to 20% of the length of the learning period. As mentioned in section 2.3.3, K-fold cross-validations is used with K=5 using a block-style partition of data.

17

3.1.1 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 reconstructions of $R = 50$ individual members for each method. 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 $\mathcal{T} = [\![1000, 1970]\!]$ and the learning period $\mathcal{T} = [\![1856, 1970]\!]$ encompassing 110 available proxy records with $n = 115$. ~~In this section the training periods length is set to 92 and testing periods length to 23.~~

Fig ~~.~~ 6 shows that RF method, particularly useful for ~~large~~ _larger_ datasets, is more efficient using the ~~whole set of proxy records with $med(S_{NSCE}) = 0.18$~~ _proxy records correlated at the 80% confidence level_ _with $med(S_{NSCE}) = 0.15$_ (_med_ is the 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 3 other regression methods are more adapted when the finest proxy selection (95%) is applied, as highlighted by Ortega et al. (2015) for the PCR. Fig ~~.~~ 6 is also evidencing that the widely used PCR ~~methods~~ _method_ 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~~ _2.3.2_) by using ~~the most constraining~~ _any_ 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 (see section ~~2.4~~ _2.3.2_) for every choice of proxy selection.

Overall, RF gives the best NSCE scores and also provides the best reconstruction. Nevertheless, it should be stressed that these results have been obtained for a particular ~~length in the training/testing windows of (92/23~~ _learning period (1856-1970)_). The sensitivity to this is assessed in the next section.

3.1.2 Sensitivity to the ~~length of training and testing periods~~ _learning period_

~~To estimate the sensitivity of the reconstruction performance to the length of the training and the testing periods, we set again the reconstruction period to $\mathcal{T} = \{1000, \ldots, 1970\}$, and the learning period to $\mathcal{T} = \{1856, \ldots, 1970\}$, with $n = 115$. Based on the findings of the previous subsection, we only keep the proxy records which are significantly correlated with the NAO index at the 95% confidence level 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 $R = 50$ reconstructions with different window sampling for each method by gradually increasing the length of the training periods: from 30% to 90% of the initial size of the learning period, with a step of 5%. Training periods lengths out of this interval gives extreme negative scores and have thus not been considered.~~

~~According to the NSCE metric we find a large range of splitting periods for which validation scores are relatively similar and significantly positive for RF and Enet (from around 35% to 85%) (Fig. ??). For PLS the only acceptable setup is obtained by using the split 80% of the total size of the learning period for the training ($n_{train} = 92; n_{test} = 23$). The only optimal window split is 70% of the total for the training for PCR ($n_{train} = 80; n_{test} = 35$) (Fig. ??). 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. ??).~~

3.1.3 ~~Sensitivity to the reconstruction period~~

In this section, we keep for each method the optimal selection of proxy records over the training periods (see section ~~4.1~~3.1.1)~~and the optimal training/testing windows length (see section 4.1.2)~~. We explore the impact of the reconstruction period. This affects the final reconstruction in two different ways, both related to the final proxy selection~~. Firstly it modifies the initial set of proxy records considered (as they need to cover the whole reconstruction period). Secondly, it changes the period of overlap with the observations and then the randomly drawn training/testing splits.~~

, as explained in section 2.1.

We run the reconstruction for 31 periods $\mathcal{T}$: from 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 (see Fig ~~.~~7). Fig ~~.~~7.a) shows the NSCE scores obtained for the different reconstruction/learning periods. Using the NSCE metric, we find that the best reconstruction time window is ~~1000-1973 for the RF method and 1000-1970 for the 3 others~~1000-1972 for PLS and RF methods and 1000-1971 for Enet and PCR methods.

Following the optimal setup for each method from section ~~4.1~~3.1.1~~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~~47 records and the three others uses 21 records. (Fig 7b). 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 . 7).~~

$med(S_{NSCE}) \simeq 0.16$ and $S_{NSCE} \in [-0.4, 0.4]$ (Fig 7a).

~~In contrast with the previous setups investigated in this study, the~~Results show that the four methods are strongly affected by the choice of the reconstruction period. Thus, we recommend to determine this period carefully with ~~numerous~~different simulations on different time windows, following the approach we presented here, easily performable using ClimIndRec. 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~~3.1.1 ~~, 4.1.2 and 4.1.3~~and this section).

3.2 ~~Reconstructions assessment~~

~~We now compare and assess the best reconstructions obtained for each of the methods~~Scientific results We compare and investigates the reconstruction with highest scores for each method following section 3.1. The four optimized reconstructions are obtained by ~~maximizing the NSCE scores on the training/testing period (see section 4.1.2) and the total reconstruction period (see section 4.1.3),~~ using the full set of proxy records for RF and only using the proxy records significantly correlated at the 95% confidence level with the NAO index over ~~each training periods~~ the learning period for the other methods (see section ~~4.1~~3.1.1). RF and Enet reconstructions are performed for the period 1000-1972 while PCR and PLS reconstructions are performed for the period 1000-1970 (section 3.1.2).

19

3.2.1 Comparison with previous ~~workFig .~~ works

Fig 8 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. The regression uncertainties (see supplementary informations) are also shown for the four reconstructions of this study on Fig .8. The normality of the residuals for the fours methods has been verified for both the models built over the training samples and the final model as demonstrated in Fig .10. Tab. 1 and Fig .8 shows that the NAO reconstruction based on RF is distinguishable from the four others including Ortega et al. (2015). Indeed its correlation with the other indices ~~is between 0.69 and 0.79~~ ranges between 0.49 and 0.67 (Tab. 1) while the paired correlations obtained between the others are greater than ~~0.95~~0.88. Additionally Fig .9 shows that the RF reconstruction has a higher correlation with the Jones et al. (1997) NAO index than the other indices: r=~~0.96~~0.98 (p<0.01), while Ortega et al. (2015) reconstruction has a correlation of 0.45 (p<0.01). The RF reconstruction that uses ~~108~~46 proxy records (22 common proxies with Ortega et al. 2015) presented in Fig 11, has the best NSCE scores ($med(S_{NSCE}) = 0.18; S_{NSCE} \in [-0.33, 0.39]$) $med(S_{NSCE}) = 0.16; S_{NSCE} \in [-0.24, 0.33]$) (section 3.1.1) and its correlation scores ($med(S_{cor}) \simeq 0.47; S_{cor} \in [0.09, 0.81]$$med(S_{cor}$ 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 from Ortega et al. (2015). This improvement in performance may arise from the inclusion of new relevant proxy records into the reconstruction, but also from the use of a new statistical regression methods for climate index reconstruction: the RF. Finally, it has to be stressed that the 5 reconstructions presented in Fig .8, including Ortega et al. (2015), do not show a predominant positive NAO phase during the ~~MCA~~Medieval Climate Anomaly, contrary to the hypothesis formulated by Trouet et al. (2009).

3.2.2 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)(~~$r \simeq -0.09; p \simeq 0.23$~~$r \simeq -0.11; p > 0.18$). The same is true for the best reconstruction of the other methods (not shown) and Ortega et al. (2015). None of the reconstructions (including Ortega et al. (2015) 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 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]. We use the 10 large volcanic eruptions selected in Ortega et al. (2015) and a second selection (see supplementary informations) of the 11 largest volcanic eruptions from the well-verified reconstruction of Sigl et al. (2015). By using a superposed epoch analysis and Rao et al. (2019) Monte-Carlo approach ~~as in Ortega et al. (2015)~~to calculate significance (see appendix 4), 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~~its p-value just ~~above~~ under 0.1 (Fig .12). 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, PLSand PCR~~PLS, but one year after the eruption with a p-value under 0.05 (Fig 12). For RF, Enet and PCR, 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~~ Individual response analysis shows that for the RF reconstruction, this result is particularly significant for the 2 largest 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

4   Discussion and conclusion

4.1   Discussion, caveats and outlooks

The results presented above regarding the NAO have all been obtained using ~~CliMoRec~~ClimIndRec. 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~~ ClimIndRec 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~~ ClimIndRec could be further improved in different ways.

~~Firstly, CliMoRec~~ ClimIndRec 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~~ ClimIndRec as their use in an interpolated version would artificially increase their weight in the reconstruction and thus possibly induce spectral artefacts in the reconstruction [Hanhijarvi et al., 2013]. ~~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~~ The most optimal way to develop a statistical model over the instrumental period is to use as many proxies as possible and as many years of observations as possible. This leads to a paradox since periods that are well covered by observation data are the most recent ones, which are generally less well covered by proxies. However, future versions of ClimIndRec will be dedicated to develop other probabilistic-based reconstruction approaches to deal with missing data such as Bayesian Hierarchical Models [~~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~~ or regularized Expectation-Maximization algortihms [Schneider, 2001 ; Mann et al., 2008 ; Guillot et al., 2015]. Another point that is limiting the capacities of ~~CliMoRec~~ ClimIndRec 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~~ ClimIndRec does not account for dating uncertainties in proxy records. Future developments of ~~CliMoRec~~ ClimIndRec 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~~ ClimIndRec 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 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) [Wang et al., 2014]. 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~~ ClimIndRec appears as an interesting avenue to explore.

4.2 Conclusions

We have proposed and described here four statistical methods for reconstructing modes of climate variability and have compared them for a particular example: the reconstruction of the NAO. By identifying and minimizing the sources of reconstruction uncertainty due to the method used (~~sections 3, 4.1~~section 3.1.1, ~~4.1~~3.1.2~~and 4.1.3~~), the time frame considered (section ~~4.1.3~~3.1.2) and the proxy selection (sections ~~4.1~~3.1.1~~and 4.1.3~~), we found the optimal NAO reconstructions. It was obtained for the RF method over the time frame ~~1000-1973 using the 108~~ 1000-1972 using the 46 proxy records available on this time frame (section ~~4.2~~3.2.1)~~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 while it clearly outperforms the other methods (section 3.1) 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~~0.49; p<0.01 over the period 1073-1855).

We have shown that for Enet, PLS and particularly PCR which is frequently used in ~~paleclimatology~~paleoclimatology, selecting proxy records with a strong correlation with the index to be reconstructed over the training periods is a good way to improve the NSCE scores, and hence it allows more reliable reconstructions (section ~~4.1~~3.1.1). Contrarily, RF gives more reliable reconstructions using the ~~whole set of records (section 4.1~~proxy records significantly correlated at the 80% confidence level with the NAO (section 3.1.1). This may be due to the fact that it has been mainly developed for large datasets [Breiman, 2001]. 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. ~~CliMoRec~~ ClimIndRec should allow to easily perform such new reconstructions.

In order to extract the most robust reconstruction, numerous simulations are needed. To simplify it, ~~CliMoRec~~ ClimIndRec performs a reconstruction by considering ~~several~~ different 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 splits, and a threshold confidence level for the correlation between the proxy records and the target index. ~~CliMorRec~~ ClimIndRec is an opportunity to reconstruct quickly and with quantified reliability ~~several climate modes~~different climate timeseries. This may allow ~~us to improve~~ improving our understanding of the last millennium large-scale climate variations, such as the ~~MCA and the LIA~~Medieval Climate Anomaly and the Little Ice Age, as well as the interactions between ~~the modes~~ climate modes and their role in driving climate, which will be analysed in future studies.

23

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~~ ClimIndRec 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.~~ appendix 1.

Acknowledgements
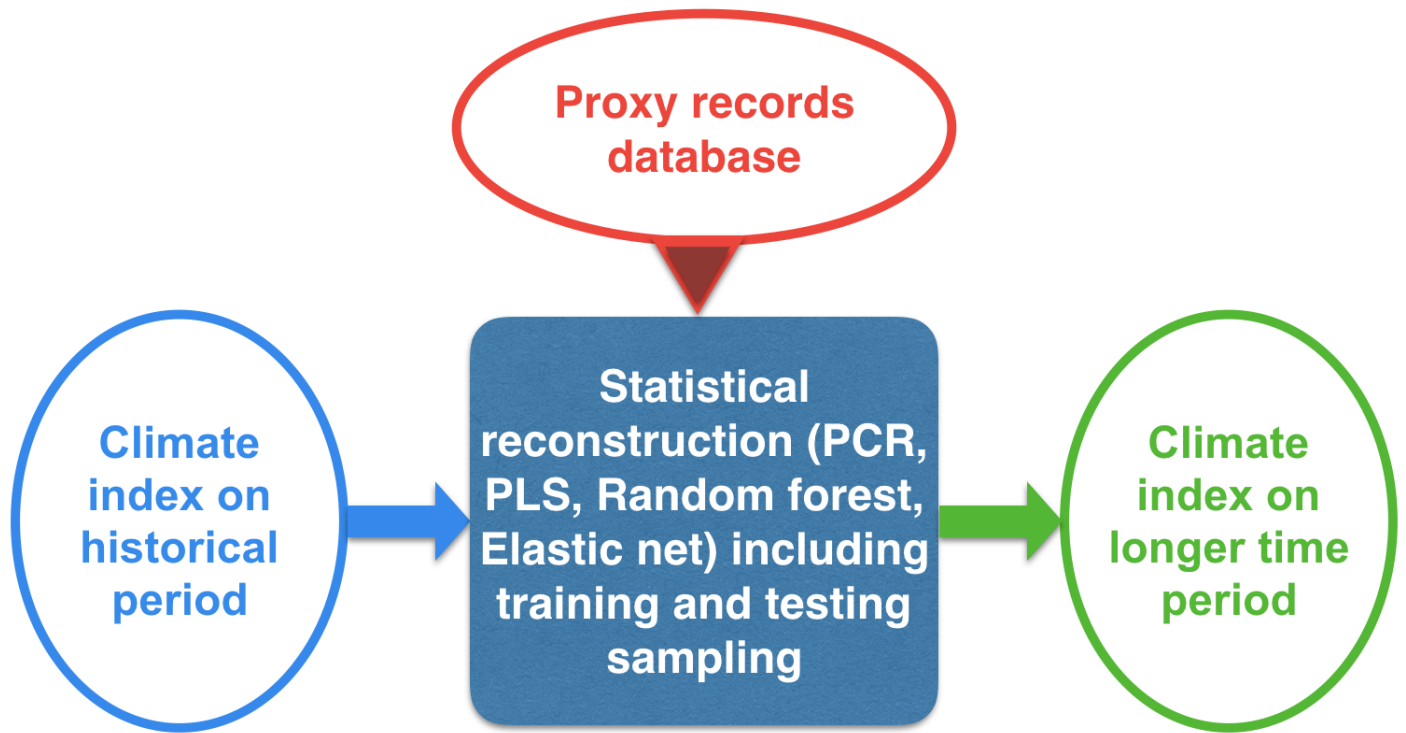
FIGURE 1 – *Scheme summarising the main features of* ~~CliMoRec~~ _ClimIndRec_.
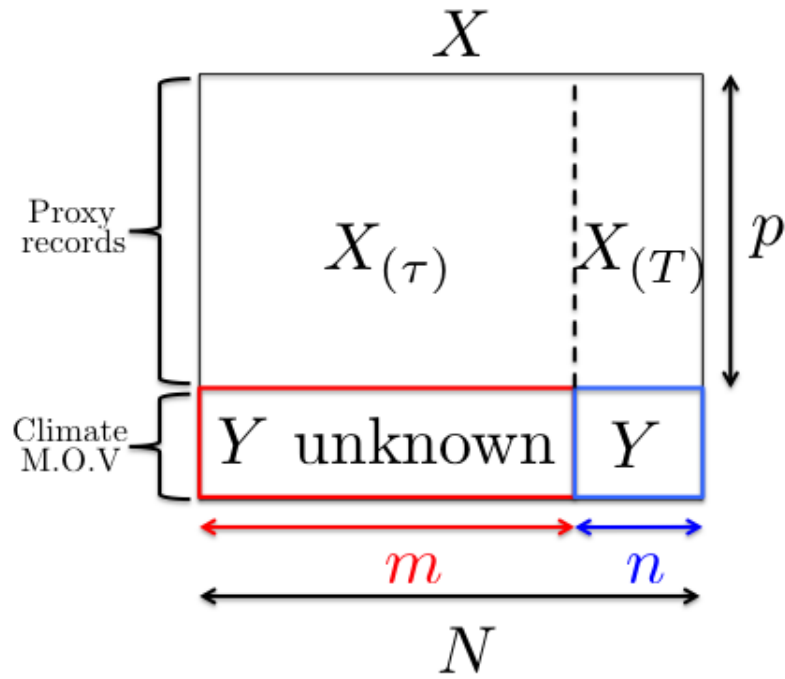
FIGURE 2 – *Scheme of the initial data. X and Y are respectively the proxy records matrix and the index of the considered mode of variability. N is the size of the common period of all proxy records. n is the size of the common period of all proxy records and the index of the mode of variability. m is 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.*
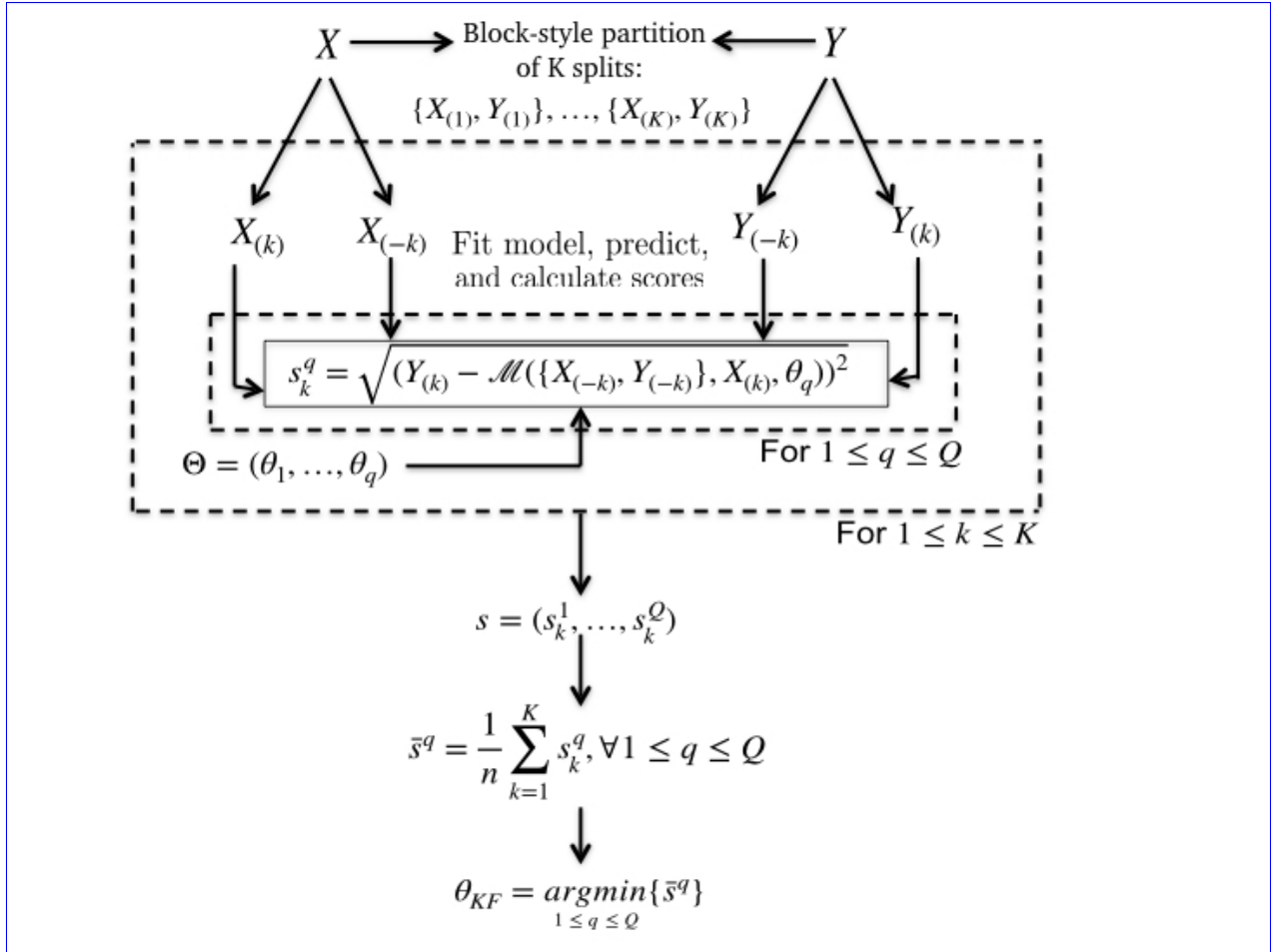
FIGURE 3 – *Scheme of a K-Fold cross validation procedure to select the optimal* ~~parameter~~ *hyperparameter of a specific learning method* $\mathcal{M}$*. X is the input set of predictors and Y the corresponding variability mode index.* $\forall 1 \leq i \leq n$, $\{X_{(k)}, Y_{(k)}\}$ *is the kth* ~~randomly drawn~~ *block-style based group of observation and* $\{X_{(-k)}, Y_{(-k)}\}$ *contains all observations except the ith.* $\Theta = (\theta_1, \ldots, \theta_Q)$ *is the ensemble of possibles values of the s hyperparameters $\theta \in \mathbb{R}^s$.*
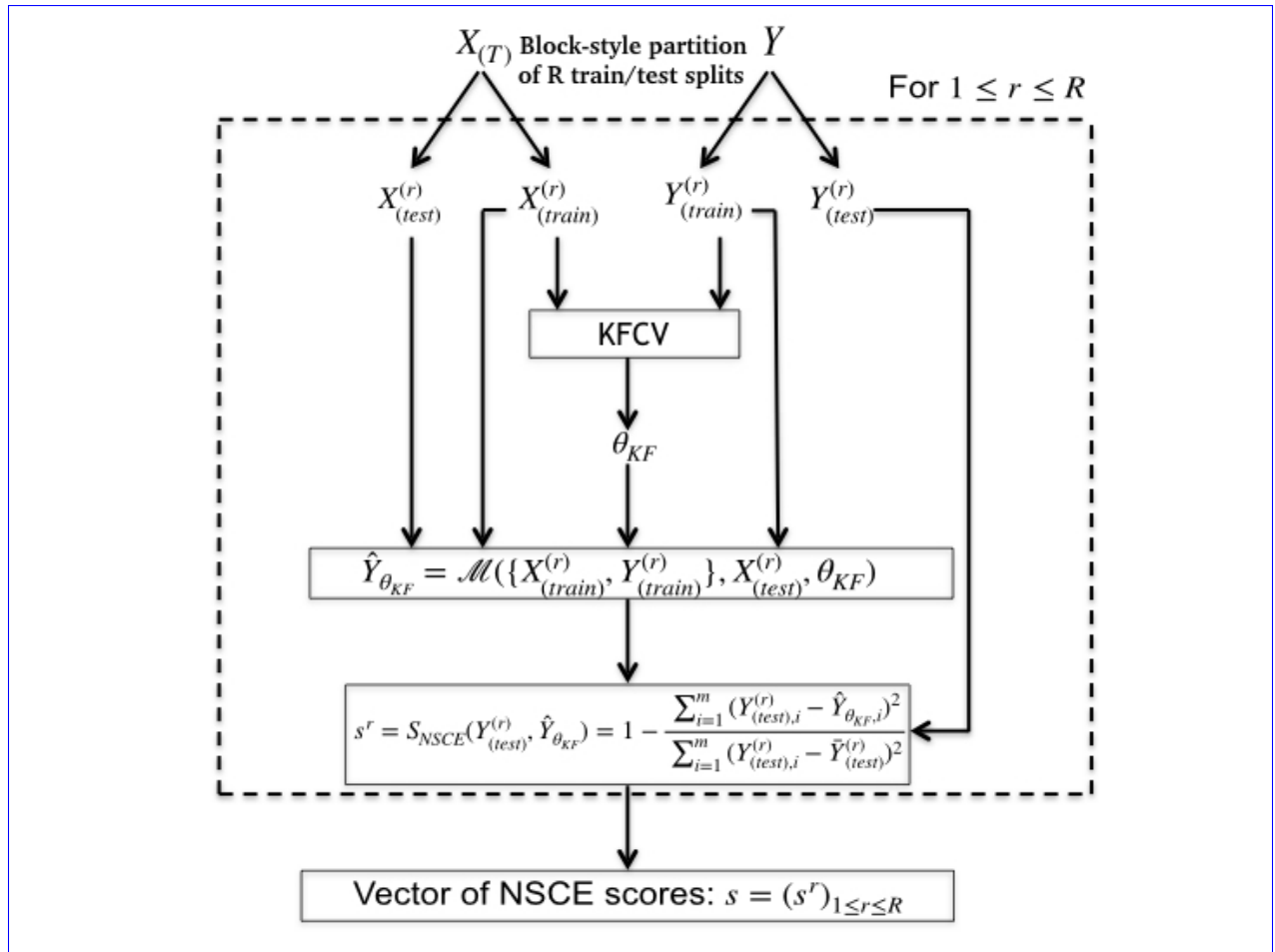
FIGURE 4 – Scheme of the whole procedure for scores calculation for a given method M. Y is the index of the chosen mode of variability. $X_{(T)}$ is the proxy dataset restricted to the period where Y known. $\{X_{(train)}^{(r)}, Y_{(train)}^{(r)}\}$ is the rth training sample and $\{X_{(test)}^{(r)}, Y_{(test)}^{(r)}\}$ is the rth testing sample. θ_{KF} is the empirically optimal set of parameters obtained by applying the KFCV (Fig .3 ; section ~~2.5.1~~2.3.3)
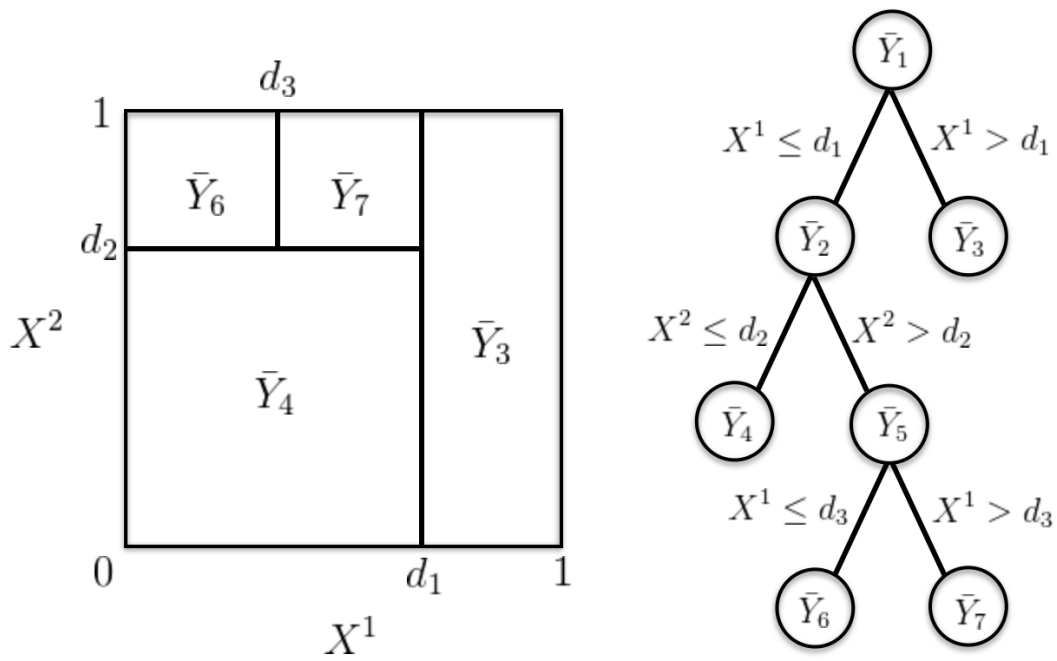
FIGURE 5 – *Dyadic partition of the unit square (left) and its corresponding regression tree (right). Y is the predictand and X^1, X^2, X^3 are the predictors. d_1, d_2 and d_3 are the optimal thresholds of the three steps respectively.*
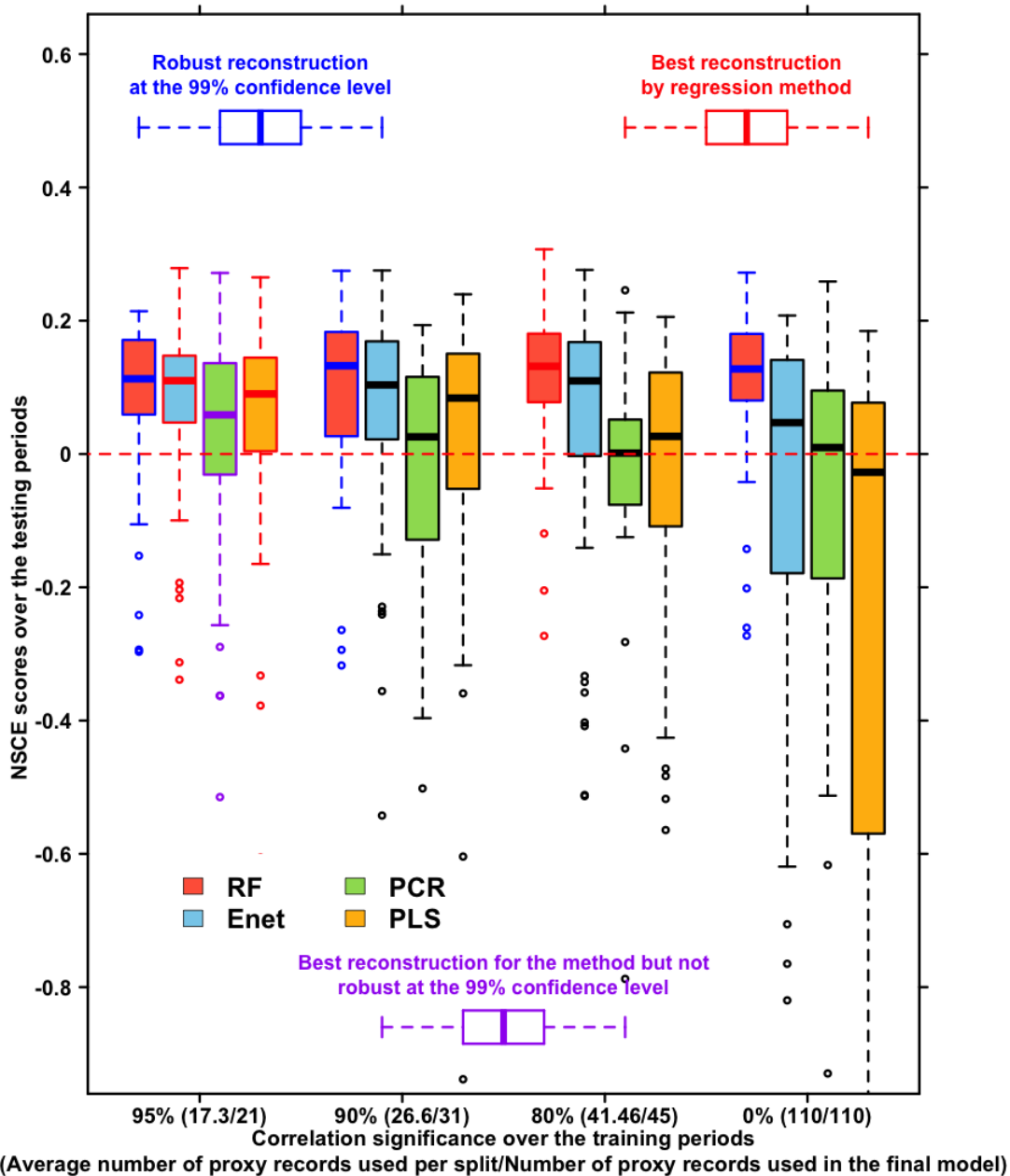
FIGURE 6 – Boxplot of NSCE scores obtained for the four methods and different groups of proxy records by reconstructing the NAO index on the period 1000-1970 with R = 50 training/testing ~~randomly drawn~~ samples. ~~Training samples size is n_train = 92, and testing samples size is n_test = 23.~~ Green boxplots are the NSCE scores obtained for the PCR method. Yellow boxplots are the NSCE scores obtained for the PLS method. Red boxplots are the NSCE scores obtained for the RF method. Blue boxplots are the NSCE scores obtained for the Enet method. The first cluster of boxplots is the NSCE scores obtained by using all the available proxy records over the period (110 proxy records). The second cluster of boxplots is the NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 80% confidence level over the training periods. The third cluster of boxplots is the NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 90% confidence level over the training periods. The fourth cluster of boxplots is the NSCE scores obtained by using only proxy records significantly correlated with the NAO index at the 95% confidence level 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.
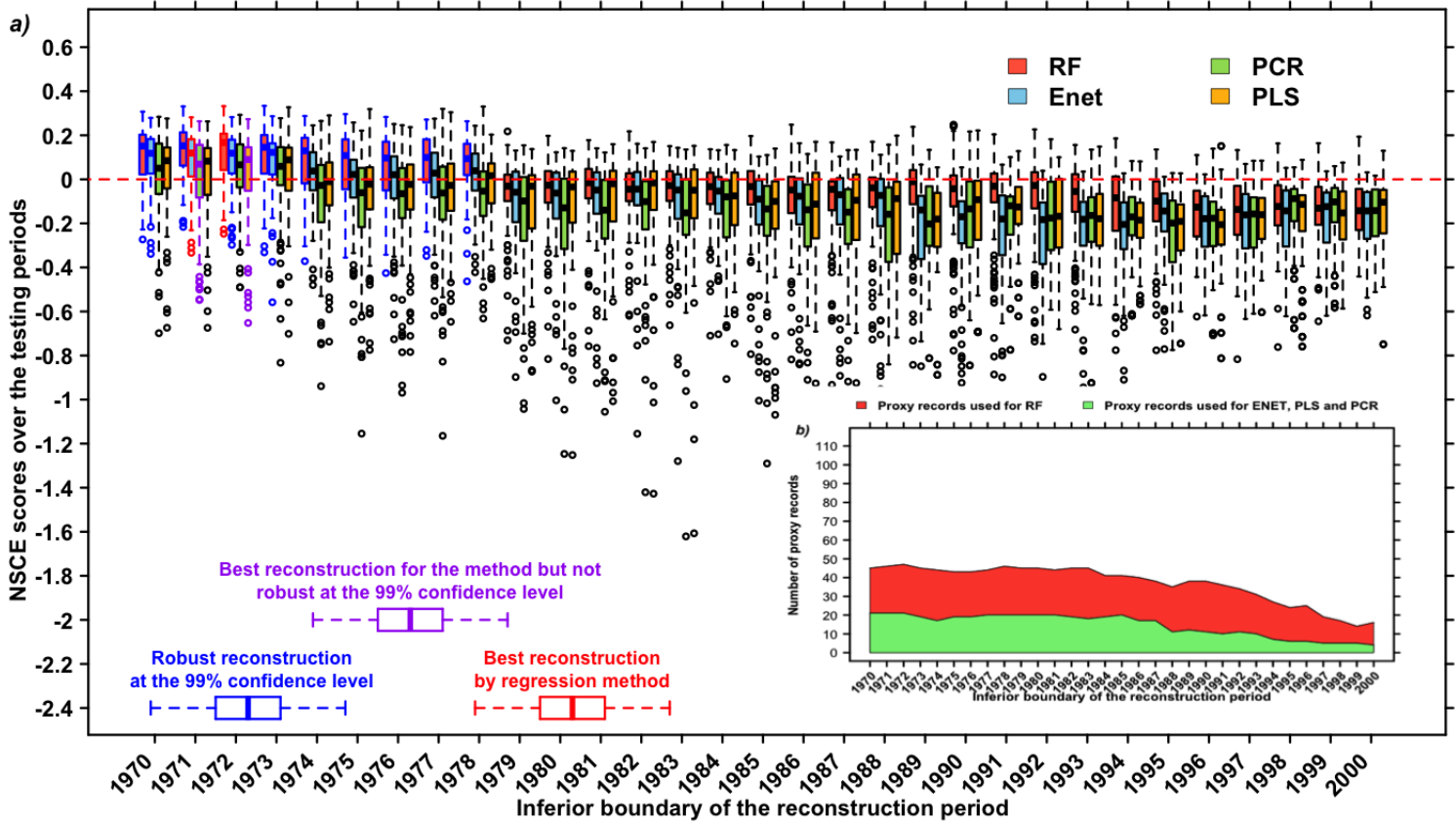
30

FIGURE 7 – ~~*All of the reconstructions are performed for R =* 50 *randomly drawn training/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.*~~ *a) Reconstructions are performed using 31 reconstruction period for the four methods: from 1000-1970 to 1000-2000 by moving the superior born by* ~~*1.*~~ *1 with R = 50 training/testing samples. RF reconstructions are performed using the* ~~*whole set of available*~~ *proxy records* ~~*(110, section 4.1.1) with training samples length of*~~ *significantly correlated at the* 80% ~~*of*~~ *confidence level with the* ~~*length of*~~ *NAO over the* ~~*learning period*~~ *training periods (section* ~~*4.1.2*~~*3.1.1). PCR reconstructions are performed using by selecting the proxy records significantly correlated at the 95% confidence level with the NAO over the training periods (section* ~~*4.1*~~*3.1.1)* ~~*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 over the training periods (section* ~~*4.1*~~*3.1.1)* ~~*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 99% confidence level. Boxplots with red edges correspond to the scores associated with the best reconstruction for each method. b) Proxy records available/used by reconstruction period. Red area gives* ~~*the number*~~ *: Number of* ~~*available proxy*~~ *records* ~~*which is typically the number of records*~~ *used for* ~~*the RF reconstructions*~~*. Green area:* ~~*total*~~ *Number of records used for Enet, PCR and PLS for each reconstruction period.* ~~*Blue area: number of proxy records used per training/testing splits on average for Enet, PCR and PLS methods.*~~
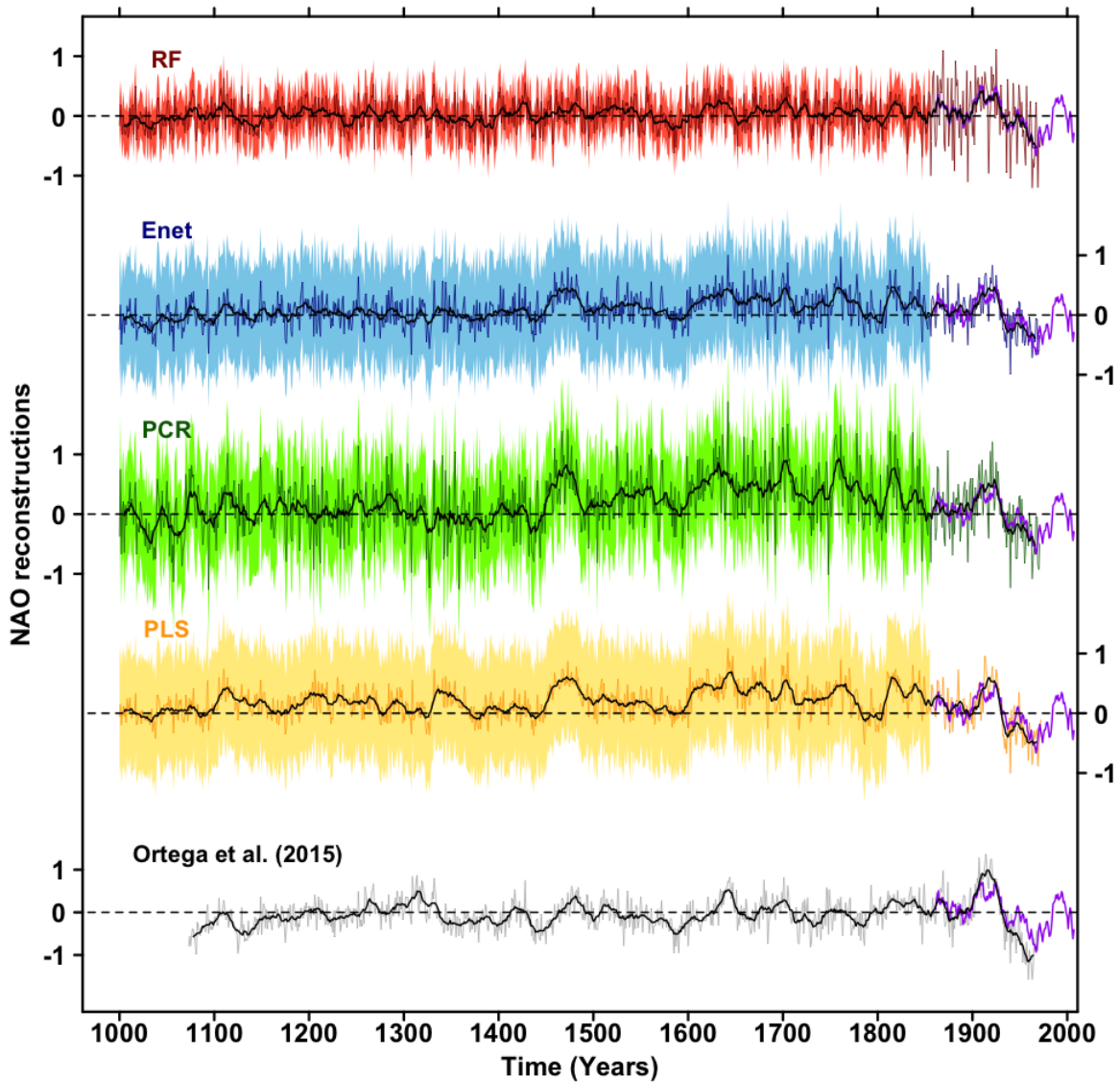
FIGURE 8 – Red line: RF reconstruction on the period ~~1000-1973~~ 1000-1972 (section ~~4.1.3~~3.1.2), using ~~the whole set of available proxy records (110, section 4.1.1) with training samples length of~~ significantly correlated at the 80% ~~of~~ confidence level with the ~~length of NAO~~ over the ~~learning period~~ training periods(section ~~4.1.2~~3.1.1). Blue line: Enet reconstruction on the period ~~1000-1970~~ 1000-1971 (section ~~4.1.3~~3.1.2) by selecting the proxy records significantly correlated with the NAO index at the 95% confidence level over the training periods (section ~~4.1.3~~3.1.1)~~with training samples length of 80% of the length of the learning period (section 4.1.2)~~. Green line: PCR reconstruction on the period ~~1000-1970~~ 1000-1971 (section ~~4.1.3~~3.1.2) by selecting the proxy records significantly correlated with the NAO index at the 95% confidence level over the training periods (section ~~4.1.3~~3.1.1)~~with training samples length of 70% of the length of the learning period (section 4.1.2)~~. Orange line: PLS reconstruction on the period ~~1000-1970~~ 1000-1972 (section ~~4.1.3~~3.1.2) by selecting the proxy records significantly correlated with the NAO index at the 95% confidence level over the training periods (section ~~4.1.3~~3.1.1)~~with training samples length of 80% of the length of the learning period (section 4.1.2)~~. Black line (tiny): Ortega et al. calibration constrained reconstruction [Ortega et al., 2015]. Red area: Regression uncertainties (see ~~supplementary~~appendix 3) 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 lines are the corresponding 11-year filtered reconstructions for each method. Purple lines: superposed 11-years filtered Jones et al. (1997) NAO index.
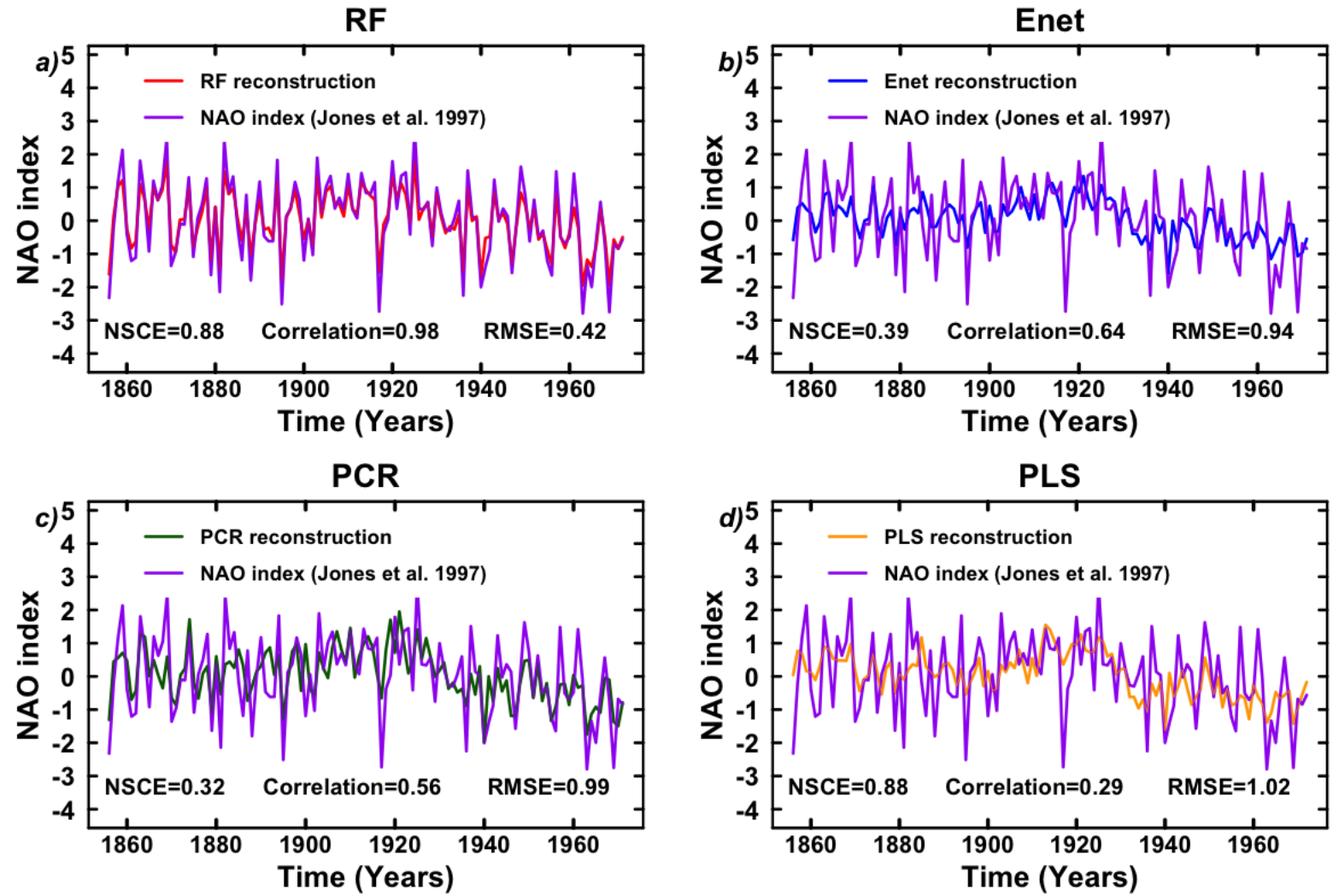
FIGURE 9 – *Comparison of reconstructions from this study with the original Jones et al. (1997) NAO index (Purple line) over their common period. a) RF reconstruction. b) Enet reconstruction. c) PCR reconstruction. d) PLS reconstruction. Purple NSCE, RMSE and correlation statistics are provided.*
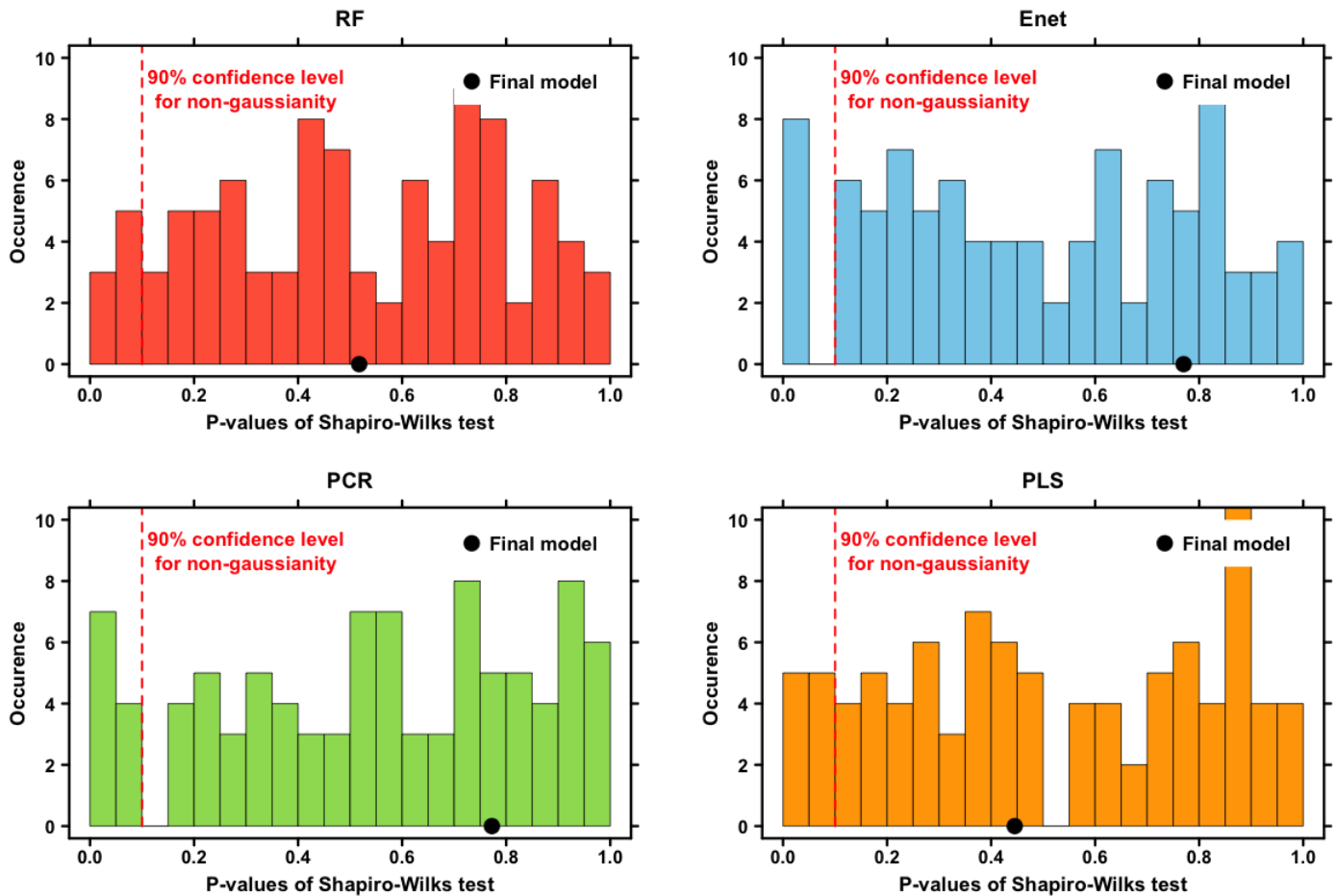
FIGURE 10 – P-values obtained from Shapiro-Wilk normality tests on the residuals from each reconstruction of Fig 8. For a), b) c) and d), the repartition of the 50 p-values obtained for each training/testing split are presented. Red dashed lines indicates the 90% confidence level for non-normality. For $0 \leq \alpha \leq 1$, if p-value<=α, it means that the residuals distributions is significantly not gaussian at the $1-\alpha$ confidence level (see shapiro.test R documentation). Black dots indicates the p-values of the residuals obtained for the final models.
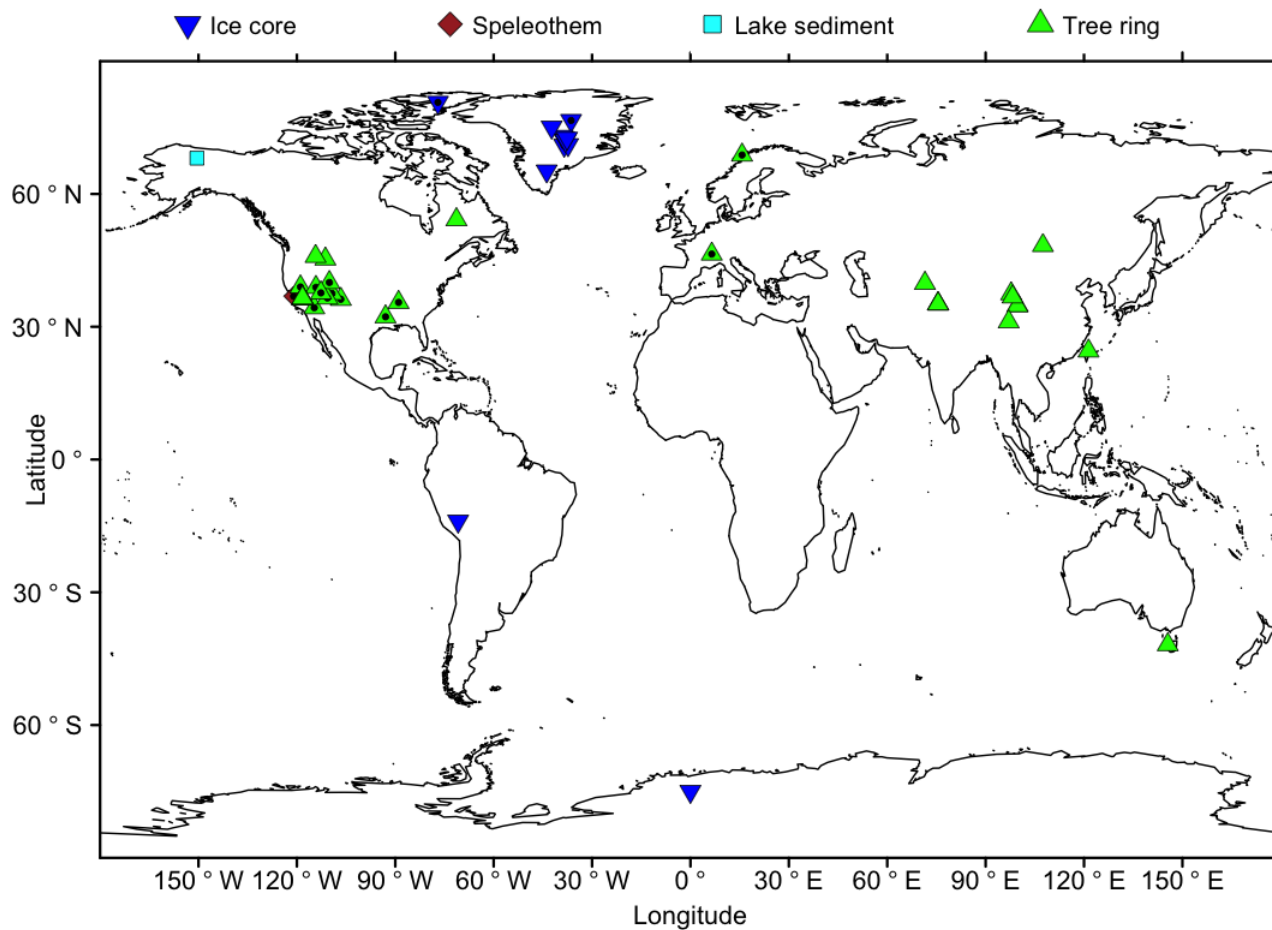
FIGURE 11 – Map of the ~~108~~ 46 proxy records used for the reconstruction of the NAO index form Jones et al. (1997) on the time window ~~1000-1973~~ 1000-1972 using the RF method. Points with a black ~~point~~ dot are the proxy records also used in Ortega et al. (2015)
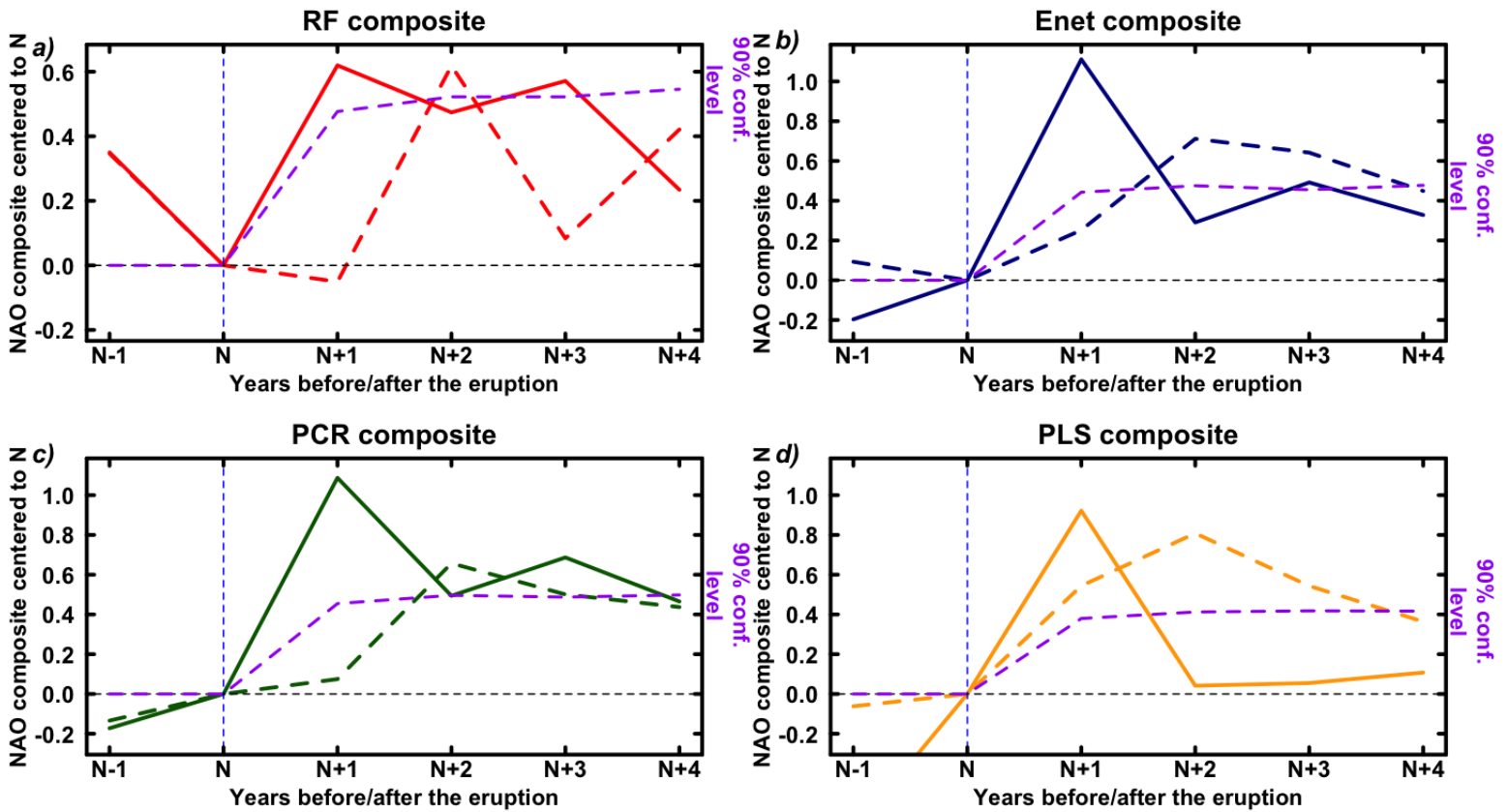
FIGURE 12 – *Superposed epoch analysis of the NAO response from two years (N-1) before to five years after (N+4) to the largest volcanic eruptions used by Ortega et al. (2015) (10 eruptions) and the 11 largest from Sigl et al. (2015). All of the composites are centred to their values at the year of the volcanic eruption* ~~occurrences~~occurrence. ~~For each method a~~90% confidence level ~~has been calculated by Monte-Carlo simulations using 1000 composites of eleven sampled 6 years long sub-series. The confidence born is calculated as the 99th percentile of the 1000 differences between the 4th and the 2rd values of the sample composite series (i.e between N+2 and N) . All of the composite series have been 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 (Rao et al., 2019, appendix 4). 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 (Rao et al., 2019, appendix 4). 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 (Rao et al., 2019, appendix 4). 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 (Rao et al., 2019, appendix 4).*

	RF	Enet	PCR	PLS	Ortega
RF	1.00	0.7	0.65	0.54	0.55
Enet	0.7	1.00	0.92	0.88	0.65
PCR	0.65	0.92	1.00	0.8	0.48
PLS	0.54	0.88	0.8	1.00	0.68
Ortega	0.55	0.65	0.48	0.68	1.00

TABLE 1 – Table of correlations between five reconstructions: Ortega et al. (2015) reconstruction; RF reconstruction on the period 1000-1972 using the proxy records significantly correlated with the NAO at the 80% confidence level; Enet reconstruction on the period 1000-1972 only using the proxy records significantly correlated with the 95% confidence level; PCR reconstruction on the period 1000-1970 only using the proxy records significantly correlated with the NAO at the 95% confidence level; PLS reconstruction on the period 1000-1970 only using the proxy records significantly correlated with the NAO at the 95% confidence level.

Code and data availability: ~~CliMoRec~~ ClimIndRec's code and the proxy records database are available at the link: ~~https://github.com/SimMiche/CLIMOREC~~ `https://github.com/SimMiche/ClimIndRec,` and the following Zenodo link: `https://zenodo.org/record/3372760#.XVxQGy2B288.`

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.

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.

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., Büntgen, U., Frank, D., Verstege, A., Nievergelt, D., and Liebhold, A.: 1200 years of regular outbreaks in alpine insects, Proc. Biol. Sci., 274 (1610), 671–679, 2006.

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.

Friedman, J., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, Journal of Statistical Software, 33, 1–22, 2010.

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.

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.

Isobe, T., Feigelson, E. D., Akritas, M. G., and Babu, G. J.: Linear regression in astronomy, Astrophysical Journal, Part 1, 364, 104–113, 1990.

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.

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.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, 2002.

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., Xoplaki, E., Dietrich, D., Jones, P. D., Davies, T. D., Portis, D., Gonzalez-Rouco, J. F., von Storch, H., Gyalistras, D., Casty, C., and Wanner, H.: Extending North Atlantic Oscillation Reconstructions Back to 1500, Atmos. Sci. Lett., 2, 114–124, 2001.

Luterbacher, J., Xoplaki, E., Dietrich, D., Rickli, R., Jacobeit, J., Beck, G., Gyalistras, D., Schmutz, C., and Wanner, H.: Reconstruction of Sea Level Pressure fields over the Eastern North Atlantic and Europe back to 1500, Clim. Dyn., 18, 545–561, 2002.

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.

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.

Mevik, B., Wehrens, R., and Liland, K. H.: The pls Package: Principal Component and Partial Least Squares Regression in R, Journal of Statistical Software, 18, 1–23, 2007.

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.

Pierce, D.: ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files, URL `https://CRAN.R-project.org/package=ncdf4`, r package version 1.16, 2017.

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.

Rao, M. P., Cook, E. R., Cook, B I an Anchukaitis, K. J., D'Arrigo, R. D., Krusic, P. J., and LeGrande, A. N.: A double bootstrap approach to Superposed Epoch Analysis to evaluate response uncertainty, Dendrochronologia, 55, 119–124, 2019.

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.

Sammut, C., Webb, G. I., Bhattacharya, I., Branke, J., and et al.: Holdout Set, in: Encyclopedia of Machine Learning, edited by Sammut, C. and Webb, G. I., p. 507, 2009.

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, 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`, 1998.

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 multi-decadal 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 Earthâ̆Źs 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.

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.

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.

Wickham, H.: stringr: Simple, Consistent Wrappers for Common String Operations, URL `https://CRAN.R-project.org/package=stringr`, r package version 1.2.0, 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.: The Adaptive Lasso and its Oracle Properties, Journal of the American Statistical Association, 101, 1418–1429, 2006.

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