# Review of
# Reconstructing climatic modes of variability from proxy records using CliMoRec version 1.0
# by Michel et al

June 3, 2019

**Recommendation**: *Minor Revisions*

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

## 1 Scientific Comments

### 1.1 Paleo Mumbo Jumbo

A common feature of paleoclimate statistics is the introduction of methodological twists that have little or no theoretical justification. Witness the tortured Principal Component regression method of *Mann et al.* (1998), which has created endless backlash from statisticians for little gain. Other examples abound. The lesson is that, unless there is a clear theoretical or heuristic justification for modifying a tried and true method, one had best stick to the tried and true method.

While the paper carefully describes the classic regression flavors or learning algorithms used here (PCR, elastic net, Random Forest, PLS), it also wraps them into an extremely unconventional form of bootstrapping (subselecting parts of the training period in an unspecified way) and averaging over this sample to obtain their "best" reconstructions (Section 2.2, point #6). I know of no justification for doing this, and it seems highly redundant with the cross-validation approach. I thought I had pointed this out in my original review, but I cannot find it there, as it got overshadowed by other considerations. It is now time to address this serious issue.

My recommendations are:

- Provide a theoretical justification
- Demonstrate using simulated data that this is a sensible (I suspect this won't work, but I'm open to surprises).
- Clearly explain the rationale in the text.
- Make sure users can easily turn off this feature, in case they want to stick to tried and true methods.

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

### 1.2 Statistical Models are Models too

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

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

### 1.3 Perfunctory Validation

Validation has improved in this version with the addition of more metrics, and an analysis of residuals, which will be very reassuring to informed readers. A more fundamental issue is the sampling: the authors currently do not specify how they perform cross-validation, with substantial implications for the estimation of generalization error.

Put simply, cross-validation is a way to estimate generalization error, that is, the error that one would make by estimating values of the target (here, the NAO index) that lie outside the range of the training interval (*Arlot and Celisse*, 2010). That is ostensibly the goal of reconstructing pre-instrumental values of a climate index, so one wants a way to estimate this generalization error using instrumental values. Cross-validation does this by selectively removing a subset of the training interval, and using it to compute validation metrics. If one does this in a sensible manner over suitable permutations, one can show that CV provides a good estimate of the actual generalization error. Here is the rub: a lot depends on the sampling mechanism. There are basically two choices: removing points at random, or removing blocks of consecutive points. The first looks like Venetian blinds in the data matrix, so it is sometimes called "blinds-style cross-validation"; the other is called "block-style cross-validation", for obvious reasons. This makes little difference when the data are independent; but in climate timeseries, autocorrelation is almost always incredibly large, so that one gets skill from persistence alone. Thus, if you remove the year 1911 from your training sample but have both 1910 and 1912, you can produce a skillful estimate of the NAO index in 1911 without having any proxies at all! You will get enough information from past and future values of the index to produce a reasonable estimate of the index at that withheld point. In climate timeseries it is essential that cross-validation be done in the block style. Is this what was done here?

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

### 1.4 Regression Methods vs Inferential Framework

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

### 1.5 Modes vs indices

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

### 1.6 Forcing attribution

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

### 1.7 Feature selection

The conclusion states: *We have shown that for Enet, PLS and particularly PCR which is frequently used in paleclimatology, selecting proxy records with a strong correlation with the index to be reconstructed over the training periods is a good way to improve the NSCE scores, and hence it allows more reliable reconstructions (section 4.1.1). Contrarily, RF gives more reliable reconstructions using the whole set of records (section 4.1.1).* This is entirely unsurprising for PCR and PLS, as they do are not designed to achieve *feature selection* (*Friednman, Hastie & Tibshirani*, 2008, chap 18). However, that is one of the purposes of RF, so it is entirely expected that it would not need additional screening prior to application. I'm a little more perplexed that Enet benefits from screening, as its L1 penalty encourages zero coefficients that effectively turn off features (here, proxies) that don't help prediction. I suspect things might be different if the LASSO is used first for feature selection, and then ridge regression applied to minimize prediction error (with the correct parameter choice), as opposed to apply both at once. The extreme variance suppression of the Enet estimate in Fig 10 suggests that the parameter choice is not optimal, in this case at least.

### 1.8 Uncertainties

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

in time. While it is true that this point remains depressingly under-appreciated in the paleoclimate community, some methods can deal with this, like BARCAST (see *Tingley and Huybers*, 2013, , Fig. 1) and LMR (*Hakim et al.*, 2016). In the regression context, this can be taken care of with frozen network analysis, or bootstrapping, as done in *PAGES 2k Consortium* (2017) (see their Figs. 7 and 8). At the very least, the authors should flag that this choice neglects changes in proxy availability over time, highlighting potential improvements for future versions of the toolbox.

## 2 Editorial Comments

Gallicisms have almost entirely disappeared; nice job! One remains: "contrarily" should be replaced by "on the contrary".

As pointed out before: *The description of methods is incredibly tedious. Sections 3.1.2, 3.2.1, 3.3.2 explain the obvious step of linear model prediction as a matrix multiplication. None of this is useful in any way as long as the code is shared.*

I maintain that the mathematical details of these regression methods is of limited use: people with a statistical background already know them, and people without a statistical background are unlikely to read them. This section should be moved to an appendix, so that the few readers who really need the details can find them, but it doesn't clutter the narrative. What *would* be interesting is to discuss the *modeling* assumptions underlying these methods (as requested above), but that is not what is done here.

Progress has also been made in that the authors are now using the up-to-date version of the PAGES 2k database. Yet, they insist on calling it Pages 2K. Not a deal-breaker, but it would be nice to use the correct spelling (PAGES 2k).

Re: code, the GitHub link works, but Zenodo registration is still a good idea to encourage code citation.

## References


Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40–79, doi:10.1214/09-SS054.

Hastie, T., R. Tibshirani, and J. Friedman (2008), The elements of statistical learning: data mining, inference and prediction, 2 ed., *Springer Verlag*.

Frost, C., and S. G. Thompson (2000), Correcting for regression dilution bias: comparison of methods for a single predictor variable, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*(2), 173–189, doi:10.1111/1467-985X.00164.

Hakim, G. J., J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger, and W. A. Perkins (2016), The last millennium climate reanalysis project: Framework and first results, *Journal of Geophysical Research: Atmospheres*, *121*, 6745 – 6764, doi:10.1002/2016JD024751.

Mann, M. E., R. S. Bradley, and M. K. Hughes (1998), Global-scale temperature patterns and climate forcing over the past six centuries, *Nature*, 392, 779–787, doi:10.1038/33859.

PAGES 2k Consortium (2017), A global multiproxy database for temperature reconstructions of the Common Era, *Scientific Data*, *4*, 170,088 EP, doi:10.1038/sdata.2017.88.

Tingley, M. P., and P. Huybers (2010a), A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and applications to paleoclimate reconstruction problems, *J. Clim.*, *23*, 2759–2781, doi:10.1175/2009JCLI3016.1.

Tingley, M. P., and P. Huybers (2010b), A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 2: Comparison with the Regularized Expectation-Maximization Algorithm, *J. Clim.*, *23*, 2782–2800, doi:2009JCLI3016.1.

Tingley, M. P., and P. Huybers (2013), Recent temperature extremes at high northern latitudes unprecedented in the past 600 years, *Nature*, *496*(7444), 201–205, doi:10.1038/nature11969.

Tingley, M. P., P. F. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam (2012), Piecing together the past: statistical insights into paleoclimatic reconstructions, *Quaternary Science Reviews*, *35*(0), 1 – 22, doi:10.1016/j.quascirev.2012.01.012.