Clim. Past Discuss., 9, C2179–C2191, 2013
www.clim-past-discuss.net/9/C2179/2013/

Climate of the Past
Discussions
Open Access

# *Interactive comment on* "Evaluating climate field reconstruction techniques using improved emulations of real-world conditions" *by* J. Wang et al.

**J. Wang et al.**

jianghaw@usc.edu

Received and published: 30 September 2013

On behalf of my co-authors, I would like to thank both reviewers for their time and valuable comments. Please see following text for response to individual comments.

## 1 Response to reviewer #1's comments

***1) I think the arguments in section 2.1 leading to the conclusion that "proxies are not indicative of local temperatures .." are wrong. ... the max SNR experiments***

***may be useful as sensitivity studies (like a best case scenario) but they have nothing to do with reality...***

We agree that the max SNR experiment is idealized. As we pointed out on page 3023, line 1, the realistic pseudoproxy designs are only end-members of real-world conditions. We nevertheless note that many proxies, particularly those representative of precipitation, do reflect large-scale teleconnections, while potentially having weaker local temperature correlations. For instance, similar to Liu and Alexander (2007) and Christiansen et al. (2009), we note that some of M08 tree-ring proxies are precipitation sensitive, which are affected by the North American Monsoon that is in turn related to ENSO. This signal is reflected in tree-ring records as a combination of local temperature and teleconnected precipitation signals. This is a feature that we feel is important to represent in more realistic pseudoproxy experimental designs. These considerations lead us to incorporate the possibility that such proxies can still inform reconstructions of large-scale temperature patterns, and thus we design the max SNR network in our experiments to reflect potential teleconnections.

We emphasize again that the max SNR and local SNR designs are adopted as the best-possible and worst-possible scenarios of real-world conditions. In Sect 2.1 of the revised manuscript, clarifications have been made on Page 7 and 9. Reviewers are also referred to the supplementary information (SI) for some further details and alternative considerations of the pseudoproxy designs.

***...I am not quite sure how the max SNR procedure works ... is the noise added to the local temperature where the proxy is positioned or to the local temperature from the grid-point with maximum correlation?***

Noise is added to the temperature grid-point with maximum correlation. This was explained in the manuscript on Page 3022 line 12–13, and in the revised manuscript on Page 8, the second paragraph from the bottom.

***2) The manuscript lacks many details about the methodologies. Are annual means of temperatures and proxies used? What is the calibration period? Are the time-series centered to zero over the calibration interval. Is the trend in the calibration period removed before the correlations are calculated?***

Please read the revised manuscript, page 16, last paragraph for clarity. Relevant editorial changes have also been made in Sect 2.1 as footnotes 1 and 2.

***It is stated on page 3020 that many of the proxies are of decadal resolution. Is this reduction of degrees of freedom taken into account when the significance of the correlations are calculated? And how are the corresponding pseudo proxies treated? These should somehow reflect the resolutions of the real proxies.***

The reduction of degrees of freedom is taken into account by the significance test: corr_isospec.m available at https://code.google.com/p/common-climate/source/browse/corr_isospec.m, which is designed to estimate the significance of correlations between (non i.i.d.) serial-correlated time series.

The test based on Ebisuzaki (1997) is non-parametric, robust, spectrum-preserving and conservative. The test “resamples” data in the frequency domain, which retains the same autocorrelation as the original series, hence preserving the persistence properties of the original timeseries. In our tests on synthetic AR(1) series, we found this method to be less prone to false positives than a classic T-test with degrees of freedom adjusted for autocorrelation, as is commonly done. See Ebisuzaki (1997) for more details.

Pseudoproxies are generated on an annual basis without accounting for the actual resolutions of the real-world proxies in the M08 network. The manuscript has been revised to acknowledge this limitation, with relevant citations added on Page 10 -11 in Sect. 2.3.

***...The authors should also give some values for the different truncations used with the RegEM and CCA methods. These might depend on the specific ensemble (do they?)...***

For RegEM-TTLS, a fixed truncation parameter (trunc = 5) is selected for most ensemble members. The values of truncation parameters for CCA depend on the specific ensemble, and are selected as the set that gives the minimum RMSE in the leave-half-out cross-validation procedure. Since our pseudoproxies also reflect the temporal heterogeneity of proxy availability, the set of truncation parameters is also updated for each pattern of missing values. Hence, it is not possible to give a set of average values for the CCA truncation parameters – the values change for each pattern, each ensemble member and each SNR level. As a general rule, a small truncation parameter (heavy regularization) is needed when many values are missing. Further details can be found in the SI in the corresponding section (Sect. 1.3).

***3) A major conclusion is that the reconstructions are best in periods with forced variability (section 4.3.1). It is argued that this is because the temperature anomaly is more spatially coherent in these periods. However, I don’t think this argument is very strong. The patterns in Figs. 7 an 8 could be a consequence of the bias being large/small in periods with large/small temperature anomalies. It would be nice to see the CE split into bias and variance. Also the authors should give some kind of measure of the spatial coherence in the 100 year slices***

The spatial coherence in each 100-year slice is shown in Figs. S3 - S12. An ENSO-like pattern is evident in all periods (regardless of CFR methods and SNR levels). Overall, the CE scores are indeed positive in more regions during the AD 1059 – 1149, 1150 – 1249, 1750 – 1849 intervals than during the other periods. These features are consistent with patterns shown in Figs 7 and 8. Nonetheless, a point that we have emphasized in the manuscript is that the pattern is model-dependent and might change when using a different GCM for validation.
CE is not easily slit into variance and bias unlike the MSE. The relative contributions of bias and variance to MSE have been added to the SI (Sect 2.1). The MSE decomposition shown in Figs. S13 and S14 helps understand the CE patterns shown in Figs. 7 and 8 in the manuscript: for instance, for CCA in the local SNR case (Fig. S13), the squared bias on average contributes to more than 60% of the MSE.

***p3017: What is meant by "analytical uncertainties of proxies" and "each methods inherent risk properties"?***

“Analytical uncertainties of proxies” refers to measurement errors; “each method’s inherent risk properties” refers to a given method’s sensitivity to input data (SNR, in other words, noise level) and parameters. Text has been rephrased on page 3 (last paragraph) in the revised manuscript for clarity.

***p3023 and Fig. 4: In the text it is the “sum of the SNR" in the figure caption it is the "average SNR". I can understand that the number of proxies becomes smaller going back in time and therefore the sum of the SNR falls accordingly. But does this also happen for the average SNR? Is it the worst proxies that reach the longest back in time? Some clarification is needed here.***

“...sum of SNRs...” should also have been “average SNR” in the text. The point of

looking at “effective SNR” is to eliminate the influence of proxy availability — having thousands of proxies with SNR = 0.05 available for the past millennium does not necessarily guarantee skillful reconstructions. As shown in the staircase pattern in Fig. 4, it is indeed the “worst” (in terms of resolution and SNR) proxies that reach the furthest back in time. In the revised manuscript, editorial changes have been made for clarity on page 10 in Sect. 2.2.

***Section 2.3. Regarding the persistence of the noise there is a discussion of this issue in Christiansen and Ljungqvist, J. Climate, 25, 7998-8003. Here the uncertainties are compared when the noise is white, ar1, or "realistic".***

This article has now been cited in the relevant area of discussion in the revised manuscript (Page 11, first paragraph of Sect 2.3).

***...A few authors, e.g. Tingley, use Bayesian inference. But even here the underlying model is based on linear regression...***

It is true that Bayesian methods that have heretofore been applied to CFR problems (Li, B. et al., 2010; Tingley and Huybers, 2010) involve a linear data model. An important distinction, however, is that Bayesian hierarchical models (BHMs) have different assumptions for process and data levels (e.g. process structure, types of forward proxy models). Thus, when the data-level model becomes non-linear, BHMs are no longer comparable to linear regression (Tingley et al., 2012; Tolwinski-Ward et al., 2013), and may indeed provide a more general alternative than the methods investigated in our manuscript.

***p2025 Strictly speaking the matrix P in Eqs. 3 and 4 must be a pruned version that match the length of the calibration interval. Also "suboptimal" is a***

***weak word to use here where $P^tP$ is non-invertible***

This is a good point. We now use subscripts "c" to denote that the relationships in Eq 4 is defined over the calibration interval and explain the subscripts in the subsequent text on Page 12.

***Section 3.1.1: It should be specified which field $\mu$ and $\sigma$ are calculated from. There is a discussion of the relative merits of RegEM Ridge and RegEM TTLS in the Comment/Reply (J. Climate, 23, 2832-2838 and 2839-2844) that could be cited here.***

$\mu \in \Re^{1\times p}$ is the mean of the $p = p_p + p_t$ variables to be estimated, while $\sigma \in \Re^{p\times p}$ is the corresponding covariance matrix of these $p$ variables. The subscripts $p, t$, as described in P3025 line 12, denote proxies and instrumental data, respectively. Clarifications and references have been added in the revised manuscript in Sect. 3.1.1.

***Section 3.1.4: It is probably difficult to describe GraphEM in a few lines, but I did not become much wiser from reading this section.***

GraphEM is described in more detail in Sect. 1.4 of the SI. Guillot et al. (submitted) is the manuscript that introduces the method. The manuscript is still under revision, with the submitted version available online at http://arxiv.org/abs/1309.6702.

***Section 4.1: The authors state that the CE and RE statistics are related to the MSE. This is probably correct but it is not easy to see just how. In Fig. 6 the authors also present the MSE, the bias, and the variance. I find these statistics much more useful than the CE and CR and will urge the authors to present the bias and the variance throughout the paper. This will also help interpreting the***

***conclusion that the CE statistics depends on the type of variability (beginning of section 4.3.1).***

While MSE is a very common loss function, its numerical value can be hard to interpret, as it is context-dependent: in one context, an MSE of 0.01 is excellent, but once variables are rescaled, it may no longer be. In contrast, the major advantage of $R^2$, RE and CE, and the reason why they are ubiquitous in the paleoclimate and weather forecasting literature, is that they are scale-invariant. We prefer CE to MSE because it is a scaled version of MSE, and is more stringent than either $R^2$ or RE: CE>0 means that the reconstruction has higher skill than the climatology., with a maximum value of 1. A more detailed description of these statistics and their merits can be found in Bürger (2007), as pointed out in the manuscript (Sect 4.1). In certain cases it is informative to explore how each method achieves its error reduction by separating bias and variance. We show some examples of MSE decomposition in the SI (Figs. S13 - S14), and relevant discussions have been added in the SI in Sect. 2.1.

***Figure 5: What is the calibration period and what is the baseline? It is my experience that reconstructions usually are biased towards zero (relative to the mean over the calibration period) as also mentioned by the authors in the end of the discussion. But here the bias is positive also when the target is positive.***

The calibration period is 850 - 1995 AD, and the baseline is the calibration mean. Details can be found in the revised manuscript on Page 16 (last paragraph of Sect. 3.2).

What's reflected in the result is a consequence of regression dilution — bias is positive as a consequence of a positive mean value in the calibration period (in this particular simulation, NCAR CSM1.4). In other words, if the calibration mean is negative, the bias would also be negative.

***I notice that the target in the reconstruction period is well inside the range of the target in the calibration period. This might lead to conservative (low) estimates of the bias (see the Comment/Reply mentioned above) and should be discussed. I think this is a more serious limitation than the mere low internal variability mentioned in the beginning of the discussion. It is not easy to see the tick-marks on the x-axes.***

The mean and variability of the reconstruction interval is well outside of the calibration period. Since we don't know what true climate was like in the past, it is also difficult to say if the models give us conservative targets or not. Fig. 5 is edited for clearer tick-marks.

***p3035: "This is encouraging .... ". Yes, but as I argue in the major comment above I think these high correlations in the max SNR network are spurious. The discussion should be changed to reflect this (perhaps the max SNR experiments should be removed from the paper).***

Comments for p3035 are addressed together with the max SNR discussion at the beginning of the response.

***p3039, end of discussion: Much of the bias in the reconstruction methods comes from expressing the temperatures as a sum of the proxies plus noise (as in Eq. 3). This is unphysical and will lead to biased estimators. The solution is to go to forward models which express proxies as sums of temperatures plus noise. This will remove the bias but may increase the variance. The variance can then be reduced by temporal and or spatial smoothing. This is explained in Christiansen, J. Clim. 674-692, 2011 which should be cited here. Forward***

***models are also used in the Bayesian settings of Tingley et al. Ammann et al. deals with the different but closely related problem of errors-invariables models where noise is both on the dependent and independent variable.***

These are excellent points, which we integrated into the discussion of the revised manuscript (Page 26, last paragraph of Sect. 5).

***Table 1, caption: What is the calibration period and the verification period? Is the mean spatial or temporal?***

The calibration period is 1850–1995 AD; the verification period refers to whichever period is being verified. In this study, for instance, the verification period would be each 100-year period. The mean is the temporal mean for each spatial point. These details have now been incorporated into the revised manuscript in the last paragraph on Page 16. Table 1 serves as a general description for these skill metrics, and hence no specific values of calibration/verification periods are provided.

## 2 Response to reviewer #2's comments

***The authors have explored the correlation patterns of the Mann et al (2008) proxy network and found that a substantial amount is correlated not with the observed local temperature but with the temperature afar. ... To be honest, I think the reader is here being asked to believe beyond what can be considered reasonable. I guess that the authors include this "teleconnected network" in their analysis to meet possible claims that, although a particular proxy may not be recording the local temperature, it may still be useful. ... but I would include in the text a disclaimer of the author's view about how realistic this really is. ...***

Sect. 2 of the manuscript has been revised to address this point. Important corrections have been made on page 7 by adding a paragraph for clarification. On page 9, the second paragraph is added as a disclaimer. Reviewers and other readers are referred to the SI (Sect. 3) for further details of our considerations on a more realistic pseudoproxy network design. Please also see our more detailed response to reviewer #1's first comment, which reflects concerns similar to those raised here.

***I guess that this statement is time-scale dependent, and that longer time scales require a smaller network for the same skill. The authors indicate somewhere else that they are focusing on decadal skill, but it is not clear whether this is still required for their conclusion to be valid.***

Results presented in this manuscript are based on 20-year low-passed reconstruction results. The main reason for this choice is because we are ultimately interested in reconstructing low-frequency variability. Filtered time series are presented also for visual clarity (especially for the global mean time series).

Unfiltered results are similar to those based on 20-year low-passed reconstructions, and they do not violate the conclusions we make in the manuscript. A simple example is provided in the table below, in which the verification statistics are summarized for the global mean time series, using the two realistic SNR staircase networks. RegEM-TTLS in the local SNR case displays the largest difference between unfiltered (total) series and low-passed (low) series. The poor performance of RegEM-TTLS, as reflected in results based on the unfiltered time series, actually further supports our conclusion in the manuscript, that RegEM-TTLS displays poor risk properties and shows strong sensitivity to noise.

## References

Bürger, G.: On the verification of climate reconstructions, Climate of the Past, 3, 397–409, 10.5194/cp-3-397-2007, 2007.

Christiansen, B., Schmith, T., and Thejll, P.: A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness, J. Clim., 22, 10.1175/2008JCLI2301.1, 2009.

Ebisuzaki, W.: A Method to Estimate the Statistical Significance of a Correlation When the Data Are Serially Correlated, Journal of Climate, 10, 2147–2153, 1997.

Guillot, D., Rajaratnam, B., and Emile-Geay, J.: Statistical Paleoclimate Reconstructions via Markov Random Fields, Ann. Appl. Stat., http://arxiv.org/abs/1309.6702, submitted.

Li, B., Nychka, D. W., and Ammann, C. M. : The value of multi-proxy reconstruction of past climate (with discussions and rejoinder) , Journal of the American Statistical Association, 105, 883–911, 10.1198/jasa.2010.ap09379, 2010.

Liu, Z. and Alexander, M. A.: Atmospheric bridge, oceanic tunnel, and global climatic teleconnections, Rev. Geophys., 45, 10.1029/2005RG000172, 2007.

Tingley, M. P. and Huybers, P.: 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, 10.1175/2009JCLI3015.1, 2010.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quatern. Sc. Rev., 35, 1–22, 10.1016/j.quascirev.2012.01.012, 2012.

Tolwinski-Ward, S., Anchukaitis, K., and Evans, M.: Bayesian parameter estimation and interpretation for an intermediate model of tree-ring width, Climate of the Past Discussions, 9, 615–645, 10.5194/cp-9-1481-2013, 2013.

Interactive comment on Clim. Past Discuss., 9, 3015, 2013.

| Method | SNR | Frequency | CE | RE | $R^2$ | bias |
|---|---|---|---|---|---|---|
| RegEM-TTLS | local | total | $-3.57_{(1.77)}$ | $-0.05_{(0.41)}$ | $+0.15_{(0.04)}$ | $+0.06_{(0.03)}$ |
| | | high | $-8.54_{(3.90)}$ | $-8.54_{(3.90)}$ | $+0.07_{(0.03)}$ | $+0.00_{(0.00)}$ |
| | | low | $-0.12_{(0.39)}$ | $+0.85_{(0.05)}$ | $+0.52_{(0.09)}$ | $+0.06_{(0.03)}$ |
| | max | total | $+0.48_{(0.06)}$ | $+0.88_{(0.01)}$ | $+0.70_{(0.02)}$ | $+0.07_{(0.01)}$ |
| | | high | $+0.32_{(0.07)}$ | $+0.32_{(0.07)}$ | $+0.56_{(0.06)}$ | $-0.00_{(0.00)}$ |
| | | low | $+0.55_{(0.13)}$ | $+0.94_{(0.02)}$ | $+0.87_{(0.04)}$ | $+0.07_{(0.01)}$ |
| M09 | local | total | $-0.16_{(0.14)}$ | $+0.73_{(0.03)}$ | $+0.32_{(0.04)}$ | $+0.11_{(0.02)}$ |
| | | high | $-0.10_{(0.08)}$ | $-0.10_{(0.08)}$ | $+0.04_{(0.03)}$ | $+0.00_{(0.00)}$ |
| | | low | $-0.24_{(0.26)}$ | $+0.83_{(0.04)}$ | $+0.57_{(0.06)}$ | $+0.11_{(0.02)}$ |
| | max | total | $+0.68_{(0.03)}$ | $+0.93_{(0.01)}$ | $+0.76_{(0.02)}$ | $+0.05_{(0.01)}$ |
| | | high | $+0.58_{(0.03)}$ | $+0.58_{(0.03)}$ | $+0.58_{(0.03)}$ | $-0.00_{(0.00)}$ |
| | | low | $+0.74_{(0.06)}$ | $+0.97_{(0.01)}$ | $+0.90_{(0.03)}$ | $+0.05_{(0.01)}$ |
| CCA | local | total | $-1.73_{(0.13)}$ | $+0.37_{(0.03)}$ | $+0.21_{(0.03)}$ | $+0.25_{(0.01)}$ |
| | | high | $-0.11_{(0.06)}$ | $-0.11_{(0.06)}$ | $+0.10_{(0.02)}$ | $+0.00_{(0.00)}$ |
| | | low | $-3.28_{(0.25)}$ | $+0.42_{(0.03)}$ | $+0.54_{(0.06)}$ | $+0.25_{(0.01)}$ |
| | max | total | $+0.49_{(0.04)}$ | $+0.88_{(0.01)}$ | $+0.76_{(0.01)}$ | $+0.09_{(0.01)}$ |
| | | high | $+0.57_{(0.02)}$ | $+0.57_{(0.02)}$ | $+0.62_{(0.02)}$ | $+0.00_{(0.00)}$ |
| | | low | $+0.37_{(0.07)}$ | $+0.92_{(0.01)}$ | $+0.91_{(0.01)}$ | $+0.09_{(0.01)}$ |
| GraphEM | local | total | $-1.06_{(0.15)}$ | $+0.53_{(0.04)}$ | $+0.25_{(0.02)}$ | $+0.21_{(0.01)}$ |
| | | high | $-0.11_{(0.05)}$ | $-0.11_{(0.05)}$ | $+0.11_{(0.02)}$ | $+0.00_{(0.00)}$ |
| | | low | $-2.00_{(0.30)}$ | $+0.60_{(0.04)}$ | $+0.64_{(0.05)}$ | $+0.21_{(0.01)}$ |
| | max | total | $+0.59_{(0.05)}$ | $+0.91_{(0.01)}$ | $+0.78_{(0.01)}$ | $+0.08_{(0.01)}$ |
| | | high | $+0.61_{(0.02)}$ | $+0.61_{(0.02)}$ | $+0.64_{(0.02)}$ | $+0.00_{(0.00)}$ |
| | | low | $+0.53_{(0.09)}$ | $+0.94_{(0.01)}$ | $+0.94_{(0.01)}$ | $+0.08_{(0.01)}$ |

**Table 1.** Verification statistics summary for the global mean temperature time series, using the staircase networks. CE, RE, $R^2$ and bias are computed for the low-frequency component (20-year lowpass) of the reconstructed global mean. In the *Frequency* column, *total* refers to the unfiltered data, *high* refers to the interannually to decadally filtered data and *low* refers to the 20-year low-passed data. All numbers given outside of parentheses are the mean of the 100-member ensemble; numbers in parentheses are the corresponding standard deviation. The table here can be compared to Table 2 in the manuscript.