# Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors

Eugene R. Wahl[1] and Jason E. Smerdon[2]

Received 24 January 2012; accepted 12 February 2012; published 20 March 2012.

[1] The performance of climate field reconstruction (CFR) and index reconstruction methods is evaluated using proxy and non-informative predictor experiments. The skill of both reconstruction methods is determined using proxy data targeting the western region of North America. The results are compared to those targeting the same region, but derived from non-informative predictors comprising red-noise time series reflecting the full temporal autoregressive structure of the proxy network. All experiments are performed as probabilistic ensembles, providing estimated Monte Carlo distributions of reconstruction skill. Results demonstrate that the CFR skill distributions from proxy data are statistically distinct from and outperform the corresponding skill distributions generated from non-informative predictors; similar relative performance is demonstrated for the index reconstructions. In comparison to the CFR results using proxy information, the index reconstructions exhibit similar skill in calibration, but somewhat less skill in validation and a tendency to underestimate the amplitude of the validation period mean. **Citation:** Wahl, E. R., and J. E. Smerdon (2012), Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors, *Geophys. Res. Lett.*, *39*, L06703, doi:10.1029/2012GL051086.

## 1. Introduction

[2] The increasing number of high-resolution climatic proxies spanning all or part of the Common Era has driven ongoing efforts to derive seasonal and annual estimates of global to regional climate reconstructions from multi-proxy networks [e.g., *Jones et al.*, 2009]. In the context of temperature reconstructions specifically, vigorous debates have developed about the applied reconstruction methodologies, the nature of the climate-proxy connection, and the estimated uncertainties in derived reconstructions [e.g., *von Storch et al.*, 2004; *Mann et al.*, 2007; *Wahl and Ammann*, 2007; *Christiansen et al.*, 2009; *Tingley and Huybers*, 2010; *Smerdon et al.*, 2011a, 2011b]. Among these, a recent study by *McShane and Wyner* [2011, hereinafter MW11] has argued that “proxies are severely limited in their ability to predict average temperatures and temperature gradients.” This conclusion is based in part on cross-validation statistics of Northern Hemisphere mean temperature (NHMT) reconstructions derived from the Lasso regression method and the unscreened *Mann et al.* [2008] multi-proxy network. The MW11 conclusions have been debated by over a dozen discussants including the present authors (see introduction by *Stein* [2011]), but here we consider a new argument that is based on the difference between reconstructions that target full spatiotemporal climatic fields and those targeting climatic indices only. This is in contrast to the MW11 paper and discussions, which focused exclusively on index reconstructions or only NHMTs derived from field reconstruction methods.

[3] It is often overlooked that some of the NHMT reconstructions for the Common Era are derived from methods that target spatial patterns in global and hemispheric temperature fields. These so-called climate field reconstruction (CFR) methods are in contrast to index approaches that only target NHMT time series. While CFR methods have been widely applied on regional scales [e.g., *Cook et al.*, 1994; *Luterbacher et al.*, 2004; *Neukom et al.*, 2010], they comprise a small subset of the global or hemispheric temperature reconstructions produced to date. For example, only two CFRs were represented in the collection of twelve NHMT reconstructions highlighted by the Intergovernmental Panel on Climate Change in Assessment Report Four [*Jansen et al.*, 2007], and only one additional global CFR has been published since [*Mann et al.*, 2009]. Despite this modest representation in the group of large-scale NHMT reconstructions, the utility of CFR data products is widely recognized [cf. *Ammann and Wahl* 2007; *Hegerl et al.*, 2006]; they already have been used for important dynamical insights [e.g., *Mann et al.*, 2009], and the motivation to further develop and apply CFR products will only increase. Continued assessments of CFRs, their underlying methodologies, their skill relative to index methods, and their comparison to other regional sources of paleo-information are therefore highly warranted.

[4] Here we focus on an assessment of CFR and index reconstructions using paleoclimatic proxy and non-informative predictor experiments. We test the skill of both a CFR and an index method using proxy data targeting the western region of North America. These experiments are compared against results targeting the same region, but derived from non-informative predictor data comprising only red-noise time series. All of our experiments are performed as ensembles, providing estimated Monte Carlo (MC) distributions of reconstruction skill. Our results demonstrate that derived CFRs are more skillful in validation than equivalent index reconstructions, but both can be clearly separated from skill distributions generated from non-informative red-noise

[1] Paleoclimate Branch, NCDC, NOAA, Boulder, Colorado, USA.
[2] Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, USA.

0094-8276/12/2012GL051086

predictors that reflect the full temporal autoregressive (AR) structure of the underlying proxy data.

## 2. Data and Reconstruction Methods

[5] We employ the HadCRUT3v 5° × 5° gridded surface temperature product [*Brohan et al.*, 2006] as the target data, and specifically focus on a regional grid (95–130° W, 30–55° N) spanning the subtropical and temperate regions of western North America and the immediately adjacent Pacific Ocean. This area was originally chosen to study regional temperature responses to large tropical volcanic events, similar to a European study [*Fischer et al.*, 2007], with a focus on enabling comparisons of earth system models for fidelity in terms of their regional forced response. The regional target is also supported by synthetic reconstruction experiments that demonstrate the tendency for spatial reconstruction skill to concentrate in areas of greatest proxy richness [*Smerdon et al.*, 2011a]; the selected region has one of the highest densities of annually-resolved proxy data in the world.

[6] The proxy data comprised all ring-width and maximum latewood density tree-ring chronologies publically available in the International Tree Ring Data Base (ITRDB) covering the period 1400–1980 C.E. in a region slightly larger than the target area (extending south to 25° N; see Figure S2 in the auxiliary material).[1] Only dendrochronological data were used to maximize the homogeneity of the proxy information. All data are available via NOAA's National Climatic Data Center, Paleoclimatology Branch/ World Data Center for Paleoclimatology (http://www.ndc.noaa.gov/paleo). All reconstructions extend to 1500 C.E., thus avoiding the use of sparsely replicated information from chronologies with few trees in the early years of their stacks. Calibrations were done from 1904–80 (1980 is the latest year of common coverage for the proxy data) and validations from 1875–1903. These periods were selected to provide the longest span of instrumental data coverage: 1) over the full grid; and 2) with thirty percent or more spatial coverage of the grid prior to the calibration interval.

[7] For CFRs, we apply ordinary least squares (OLS) regression to a truncated empirical orthogonal function (TOEF) representation of the target field and a collection of leading principal components (PCs) of the proxy data. This form of PC spatial regression [cf. *Cook et al.*, 1994] has been used for skillful CFRs in other regional contexts that include Europe [e.g., *Luterbacher et al.*, 2004; cf. *Fischer et al.*, 2007] and South America [*Neukom et al.*, 2010]. In implementing the CFR procedure, the *n* number of reconstructed instrumental PCs generated by the OLS regression ($\mathbf{U_n}$) are substituted back into the singular value decomposition of the instrumental field, $\mathbf{T}_{\text{field/fitted}} = \mathbf{U_n D_n V'_n}$, to derive the reconstructed field (where the subscript *n* denotes the rank-reduced matrices). In this formula, $\mathbf{U_n}$ is the matrix of reconstructed PC time series, $\mathbf{D_n}$ is the diagonal matrix of singular values, and $\mathbf{V'_n}$ is the transposed matrix of EOFs. Reconstructions were done for both annual and February-March average temperatures; other sub-annual time periods were evaluated but did not produce well-validated reconstructions. The annual reconstructions are used in this paper. Further information on the reconstruction details and calibration/validation statistics is available in the auxiliary material. Also available are the reconstructed annual time series of the regional mean and spatial maps of decadal averages.

[8] A multiple OLS regression method is used to compute index reconstructions for the spatial mean of the target region. This method employs the proxy PC network used in the CFR reconstruction to reconstruct the regional mean time series directly, rather than $\mathbf{U_n}$ as part of the singular value decomposition of the climate field. Using the same proxy information to derive both the field and index reconstructions allows direct comparison of these results.

## 3. Ensemble Generation

[9] Uncertainty ensembles for the CFRs are derived from a simplified version of the method outlined by *Li et al.* [2007] (D. Nychka, personal communication, 2009). In this process, the regression residuals were computed and then modeled as full AR time series. One thousand realizations of these residuals were then generated using the "hosking-sim" algorithm in the R programming language, all of which are modeled to have the same AR characteristics as the residuals in the original regression. Each set of random-draw residuals was added back to the original reconstructed PCs to create one thousand random draws from the assumed underlying distribution of the instrumental PCs. The fitting process was then redone for each group of the random-draw instrumental PCs, and the newly-generated reconstructed PC sets were substituted into the CFR equation above. This process generates a thousand-member ensemble of CFRs, conditional on the proxy data, rather than confidence intervals estimated from only one reconstruction realization [*Li et al.*, 2007].

[10] To generate a parallel CFR ensemble based on non-informative-proxy data, one thousand sets of full-AR spectrum, red noise-simulated proxy data (based on empirical estimates of the AR spectra from the proxy network) were input into the CFR algorithm to generate a MC set of reconstructions using non-informative predictors. The uncertainty ensemble methodology outlined above was then used for each of these CFR members (with 2000 replications), yielding two million CFRs based on non-informative predictors. This nested approach is required because of the stochasticity of the noise draws generating each set of non-informative predictors.

[11] Generation of uncertainty ensembles for the index reconstructions followed the methods used for the CFRs. In this case, the draws of the AR-modeled residuals were applied directly to the fitted regional mean time series to generate an ensemble of random draws from the assumed underlying distribution of instrumental data, and the fitting process was redone. For reconstructions based on non-informative predictors, the process outlined for the CFRs was used.

## 4. Results and Analysis

[12] The ensemble skill of the reconstructions is shown in Figure 1 (grid-level performance of the CFR method) and Figures 2 and 3 (spatial mean performance of both the CFR and index methods). Standard measures of skill/merit in

[1] Auxiliary materials are available in the HTML. doi:10.1029/2012GL051086.

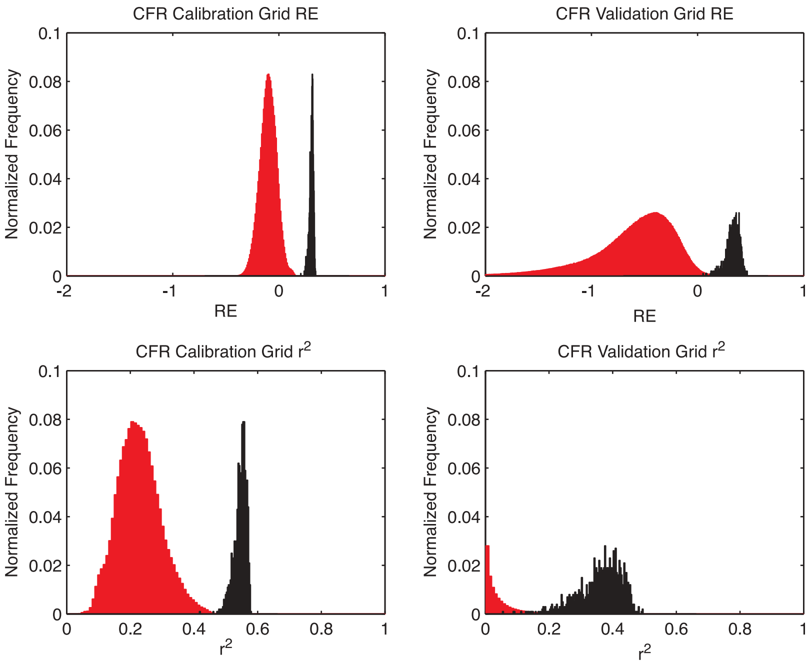

**Figure 1.** Calibration and validation RE and r$^2$ (distributions as indicated in the plot) for the spatial performance of the CFR method. Black distributions show reconstruction skill derived from proxy predictors, while those obtained from the AR non-informative predictors are shown in red. For comparison, the red histograms have been scaled to have the same maximum values as the black histograms.

paleoclimate reconstruction are used for comparison of proxy versus non-informative-predictor reconstructions [*Cook et al.*, 1994]: (1) RE measures explained variation in the calibration interval and can be interpreted as a measure of normalized squared error, as well as explained variation in the validation interval relative to the *calibration* period mean, thus rewarding successful reconstruction of a change in mean across the two periods [*Wahl and Ammann*, 2007]; (2) Pearson's r$^2$ measures coherence of variability between reconstructions and their instrumental targets; (3) CE

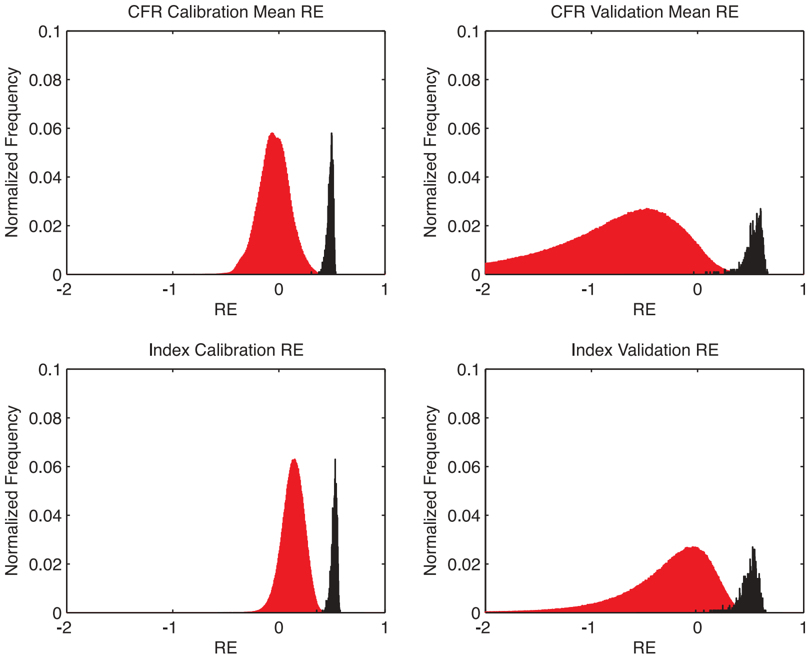

**Figure 2.** Calibration and validation RE distributions for the CFR and index regional mean reconstructions. For the CFR, the mean was computed from an unweighted arithmetic average of the grid-point time series in the field reconstruction. Proxy (non-informative predictor) histograms are shown in black (red). For comparison, the red histograms have been scaled to have the same maximum values as the black histograms.

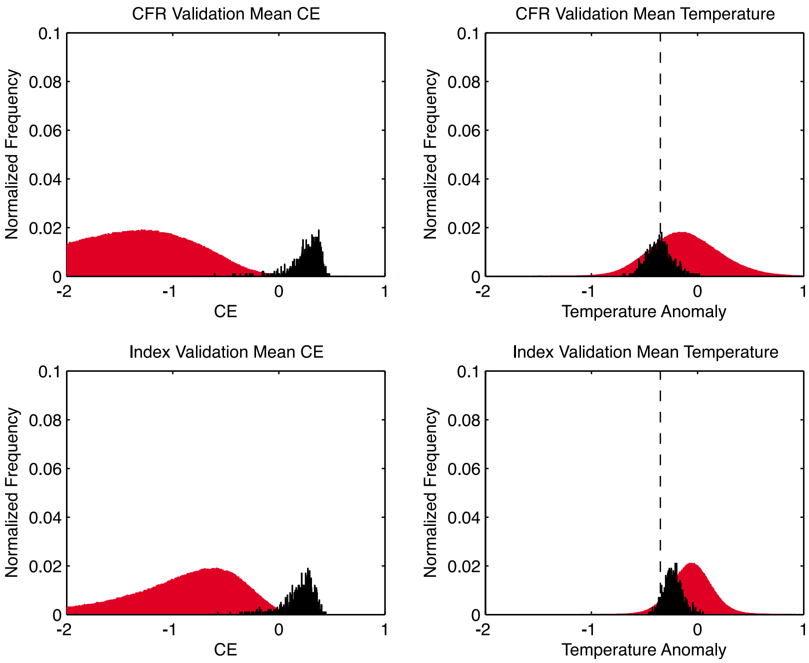

**Figure 3.** Validation CE distributions for the regional mean reconstructions (left) and mean temperature results (right), for the (top) CFR and (bottom) index reconstructions. True validation interval mean (−0.35°C) is shown by a dashed line. For comparison, the red histograms have been scaled to have the same maximum values as the black histograms.

measures explained variation in validation relative to the *validation* period mean, and thus does not reward successful reconstruction of a change in mean across the two periods [*Wahl and Ammann*, 2007].

[13] CFR performance clearly separates across the proxy vs. non-informative predictors. The histograms for all CFR measures of merit at the grid level (computed across all individual grid cells, Figure 1) have, at most, extremely small amounts of overlap between the proxy and non-informative cases; the calibration RE histograms separate completely. The same result occurs when the CFR output is used to calculate the spatial mean (Figures 2 and 3). These separations also highlight the expected lack of skill in the non-informative case, most clearly seen by the largely ≤0 scores for the RE and CE skill measures (RE/CE ≤0 indicates no skill in relation to calibration (RE) and validation (CE) climatology [*Cook et al.*, 1994]). The non-informative CFR $r^2$ results are consistent with this expectation (Figure 1); in particular, the grid-level validation $r^2$ histogram is strongly clustered near the lower bound of zero for this measure, indicating essentially no skill in the validation-period reconstructions derived with non-informative predictors. That some apparent skill is obtained in the non-informative case, notably in the CFR case for calibration grid-level $r^2$ (Figure 1) and RE of the spatial mean (Figure 2), is also an expected result of a randomly-generated process; the MC ensemble analysis allows this apparent skill to be identified as systematically lower than that obtained with real proxies.

[14] The index reconstructions similarly indicate that proxy information clearly outperforms non-informative predictors (Figures 2 and 3). The most notable difference between the CFR and index reconstruction cases is that while the index method performs as well as or better than the CFR method in calibration, it performs somewhat less well than the CFR in validation. This difference can be seen in the validation RE and CE measures, and is partly due to the loss of amplitude for reconstruction of the regional mean in the validation period exhibited by the index case (Figure 3). The median value for the CFR proxy reconstructions of the validation mean (−0.36°C) is very close to the instrumental value (−0.35°C), whereas the median value for the index proxy reconstructions (−0.23°C) is higher by 0.12°C. Note also the reduced performance of a composite-plus-scale method, relative to the index regression method (cf. auxiliary material, sections VII and VIII).

[15] The non-informative predictors perform very poorly in reconstructing the validation mean in both the CFR and index reconstruction cases (Figure 3), which can provide an important test of a reconstruction's ability to detect changes in mean state over time (in this case in relation to the calibration period mean). The lack of climate information in the non-informative predictors clearly shows up in the spread of the validation mean reconstructions across a wide range of both negative and positive values, whereas the proxy data generate validation mean reconstructions that are correctly <0 and clustered much closer to the instrumental value, with the loss of amplitude noted for the index reconstructions.

## 5. Conclusions

[16] Our results indicate that proxy information outperforms non-informative predictors (with equivalent AR structures as the proxy information) for both field and spatial mean reconstructions in our study area. This performance success is extremely strong when a CFR method is used for both purposes; it is equivalently strong in calibration for index reconstructions employing the multiple regression method we used, but somewhat less strong in validation, especially in terms of correctly reconstructing the spatial

mean. The very strong performance of the CFR method is perhaps expected given that multiple PCs associated with spatial patterns must be skillfully reconstructed. Infusing random noise into this process, even if its temporal AR structure matches that of the proxy data, will reduce retention of the spatial information and in turn the capacity of a spatiotemporal regression to generate artificial skill relative to proxies with true signal content. The inclusion of spatial pattern information in the CFR process may also help explain the absence of amplitude loss for the spatial mean reconstruction outside the calibration period, an effect sometimes noted for use of OLS in inverse model [climate = f(proxies)] applications such as used here [cf. *Ammann et al.*, 2010; *Smerdon et al.* 2011b]. This lack of amplitude loss was anticipated by pseudoproxy experiments performed in preparation for development of the western North America field reconstructions reported here (cf. auxiliary material, section I). Further examination of the extent to which TEOF CFRs are subject to this issue is an important area for further research.

[17] While there are differences between the target domain, the proxy network used, the methods employed, and the method of ensemble generation applied in our study and in MW11, our results clearly differ from MW11's negative findings about proxy efficacy for paleo-temperature reconstruction. Contrary to the conclusions by MW11, our findings make a strong case for the capacity of proxy information to inform skillful climate reconstructions, when measured against a non-informative AR null hypothesis (which in fact is a stronger noise model than the "empirical AR1" case considered by MW11 to be an appropriately sophisticated null model). This is true for both the index and CFR methods, while the latter outperforms the former in validation and such field approaches were not a subject of the MW11 examination. While many large-scale temperature reconstructions have been derived using index methods [*Jansen et al.*, 2007], multiple studies have used CFR techniques [*Jones et al.*, 2009]. Further work to understand the similarities and differences between CFR and index reconstruction methods is thus highly warranted, and should be considered in future work to benchmark reconstruction skill against null hypotheses.

[18] **Acknowledgments.** Supported in part by NSF grant ATM0902436 (JES), by NOAA grants NA07OAR4310060 (JES) and NA10OAR4320137 (JES), and by the Weather and Climate Impacts Assessment Science Program, NCAR (ERW). LDEO contribution 7527. Supplementary materials for this manuscript can be accessed at http://www.ldeo.columbia.edu/~jsmerdon/2012_grl_wahlandsmerdon_supplement.html.

## References

Ammann, C., and E. Wahl (2007), The importance of the geophysical context in statistical evaluations of climate reconstruction procedures, *Clim. Change*, *85*, 71–88, doi:10.1007/s10584-007-9276-x.

Ammann, C. M., M. G. Genton, and B. Li (2010), Technical note: Correcting for signal attenuation from noisy proxy data in climate reconstructions, *Clim. Past*, *6*, 273–279, doi:10.5194/cp-6-273-2010.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548.

Christiansen, B. T. S., and P. Thejll (2009), A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness, *J. Clim.*, *22*(4), 951–976, doi:10.1175/2008JCLI2301.1.

Cook, E. R., K. R. Briffa, and P. D. Jones (1994), Spatial regression methods in dendroclimatology: A review and comparison of two techniques, *Int. J. Climatol.*, *14*, 379–402, doi:10.1002/joc.3370140404.

Fischer, E. M., J. Luterbacher, E. Zorita, S. F. B. Tett, C. Casty, and H. Wanner (2007), European climate response to tropical volcanic eruptions over the last half millennium, *Geophys. Res. Lett.*, *34*, L05707, doi:10.1029/2006GL027992.

Hegerl, G. C., et al. (2006), Climate change detection and attribution: Beyond mean temperature signals, *J. Clim.*, *19*, 5058–5077, doi:10.1175/JCLI3900.1.

Jansen, E., et al. (2007), Palaeoclimate, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Inter governmental Panel on Climate Change*, edited by S. Solomon et al., pp. 433–497, Cambridge Univ. Press, Cambridge, U. K.

Jones, P. D., et al. (2009), High-resolution paleoclimatology of the last millennium: A review of current status and future prospects, *Holocene*, *19*, 3–49, doi:10.1177/0959683608098952.

Li, B., D. W. Nychka, and C. M. Ammann (2007), The "Hockey Stick" and the 1990s: A statistical perspective on reconstructing hemispheric temperatures, *Tellus, Ser. A*, *59*, 591–598, doi:10.1111/j.1600-0870.2007.00270.x.

Luterbacher, J., D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner (2004), European seasonal and annual temperature variability, trends, and extremes since 1500, *Science*, *303*, 1499–1503, doi:10.1126/science.1093877.

Mann, M. E., S. Rutherford, E. Wahl, and C. Ammann (2007), Robustness of proxy-based climate field reconstruction methods, *J. Geophys. Res.*, *112*, D12109, doi:10.1029/2006JD008272.

Mann, M. E., et al. (2008), Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, *Proc. Natl. Acad. Sci. U. S. A.*, *105*(36), 13,252–13,257, doi:10.1073/pnas.0805721105.

Mann, M. E., et al. (2009), Global signatures and dynamical origins of the little ice age and medieval climate anomaly, *Science*, *326*(5957), 1256–1260, doi:10.1126/science.1177303.

McShane, B. B., and A. J. Wyner (2011), A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?, *Ann. Appl. Stat.*, *5*, 5–44, doi:10.1214/10-AOAS398.

Neukom, R., J. Luterbacher, R. Villalba, M. Küttel, D. Frank, P. D. Jones, M. Grosjean, J. Esper, L. Lopez, and H. Wanner (2010), Multi-centennial summer and winter precipitation variability in southern South America, *Geophys. Res. Lett.*, *37*, L14708, doi:10.1029/2010GL043680.

Smerdon, J. E., A. Kaplan, D. Chang, and M. N. Evans (2011a), A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium, *J. Clim.*, *24*, 1284–1309, doi:10.1175/2010JCLI4110.1.

Smerdon, J. E., A. Kaplan, E. Zorita, J. F. González-Rouco, and M. N. Evans (2011b), Spatial performance of four climate field reconstruction methods targeting the Common Era, *Geophys. Res. Lett.*, *38*, L11705, doi:10.1029/2011GL047372.

Stein, M. L. (2011), Editorial on "A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?" by B. McShane and A. Wyner, *Ann. Appl. Stat.*, *5*(1), 1–4, doi:10.1214/10-AOAS449.

Tingley, M. P., and P. Huybers (2010), A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems, *J. Clim.*, *23*, 2759–2781, doi:10.1175/2009JCLI3015.1.

von Storch, H., E. Zorita, J. M. Jones, Y. Dimitriev, F. González-Rouco, and S. F. B. Tett (2004), Reconstructing past climate from noisy data, *Science*, *306*, 679–682, doi:10.1126/science.1096109.

Wahl, E. R., and C. M. Ammann (2007), Robustness of the Mann, Bradley, Hughes reconstruction of surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence, *Clim. Change*, *85*, 33–69, doi:10.1007/s10584-006-9105-7.

J. E. Smerdon, Lamont-Doherty Earth Observatory, Columbia University, PO Box 100, 61 Rte. 9W, Palisades, NY 10964, USA.

E. R. Wahl, Paleoclimate Branch, NCDC, NOAA, 325 Broadway, Boulder, CO 80305, USA. (eugene.r.wahl@noaa.gov)