NOTES AND CORRESPONDENCE

Comments on "A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness"

SCOTT D. RUTHERFORD

Department of Environmental Science, Roger Williams University, Bristol, Rhode Island

MICHAEL E. MANN

Department of Meteorology, and Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, Pennsylvania

CASPAR M. AMMANN

Climate Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado

EUGENE R. WAHL

National Climatic Data Center, National Oceanic and Atmospheric Administration, Boulder, Colorado

(Manuscript received 18 March 2009, in final form 15 December 2009)

ABSTRACT

In a recent paper, Christiansen et al. compared climate reconstruction methods using surrogate ensembles from a coupled general circulation model and pseudoproxies. Their results using the regularized expectation maximization method with truncated total least squares (RegEM-TTLS) appear inconsistent with previous studies. Results presented here show that the poor performance of RegEM-TTLS in Christiansen et al. is due to 1) their use of the nonhybrid method compared to the hybrid method; 2) a stagnation tolerance that is too large and does not permit the solution to stabilize, which is compounded in another paper by Christiansen et al. by the introduction of an inappropriate measure of stagnation; and 3) their use of a truncation parameter that is too large. Thus, the poor performance of RegEM-TTLS in both Christiansen et al. papers is due to poor implementation of the method rather than to shortcomings inherent to the method.

Christiansen et al. (2009) provide a comparison of climate reconstruction methods as applied to surrogate ensembles from the ECHAM4/Ocean Isopycnal Model (OPYC3) coupled general circulation model. Such comparisons are useful and necessary endeavors to understand the relative strengths and weaknesses of climate reconstruction techniques. We wish to address the authors' use of one technique used in their comparisons, the regularized expectation maximization (RegEM) method using truncated total least squares (TTLS). The RegEM-TTLS reconstructions shown in Christiansen et al. (2009) seem inconsistent with results presented in Mann et al. (2007, hereafter MRWA07) and with those of Lee et al. (2008) and Riedwyl et al. (2009), who both used a different implementation of RegEM than did MRWA07 but showed that RegEM generally performed as well as or better than the other methods they tested. This apparent discrepancy led us to investigate possible causes of these disparate results.

We see three main differences between RegEM-TTLS as applied by Christiansen et al. (2009) compared to MRWA07. First, Christiansen et al. (2009) do not employ the hybrid frequency-band approach favored by MRWA07, where both proxies and instrumental data were separated into two frequency bands and each frequency band was reconstructed separately. Second, Christiansen et al. (2009) use a stagnation tolerance (stopping criterion)

Corresponding author address: Scott Rutherford, Department of Environmental Science, Roger Williams University, Bristol, RI 02809.
E-mail: rutherford@fox.rwu.edu

DOI: 10.1175/2009JCLI3146.1

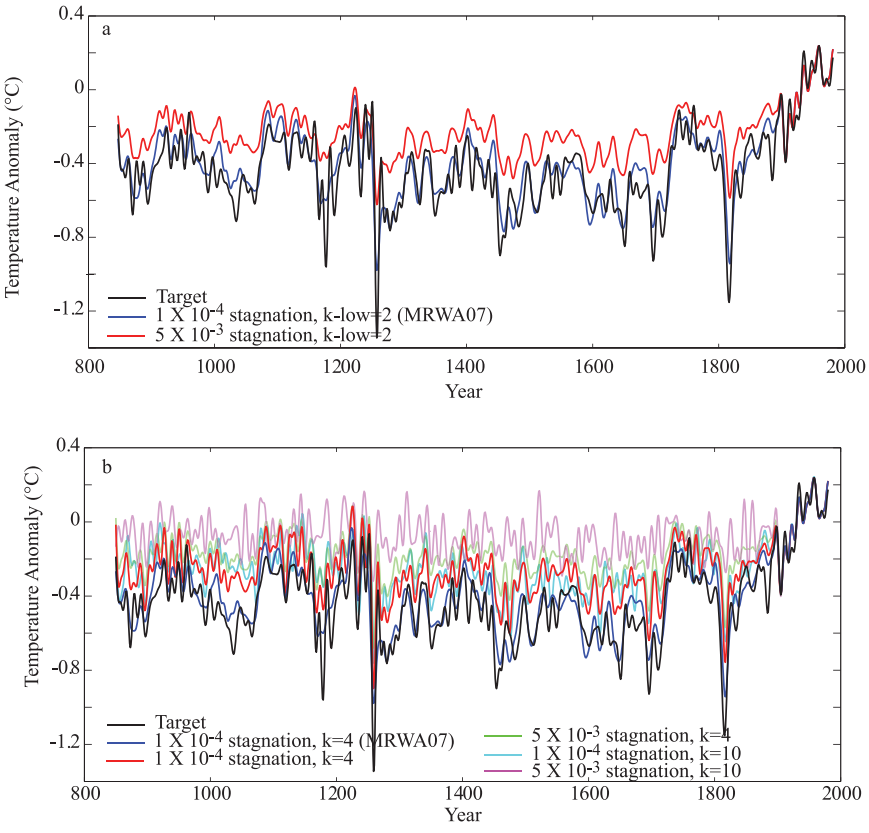

FIG. 1. Results of pseudoproxy experiments. (a) Results using the hybrid frequency-band approach. The stagnation tolerance used by Christiansen et al. results in an interim result that is not fully converged (red line). (b) Results of experiments using the nonhybrid RegEM-TTLS approach. Note that all nonhybrid results perform poorly relative to the hybrid, MRWA07 approach (blue line). The magenta line most closely approximates the implementation of Christiansen et al. whereas the red line reflects the MRWA07 nonhybrid approach.

that is large compared to that used in MRWA07, although they do test the impact of varying the stagnation tolerance (in addition, in Christiansen et al. 2010, a problematic measure of stagnation was introduced). Finally, their truncation parameter (i.e., the number of principal components retained in RegEM) is considerably larger than that used in MRWA07.

To examine the impact of these differences, we used the annual mean surface temperature field between 70°N and 70°S from a transient simulation of the Climate System Model (CSM; Ammann et al. 2007) and created white-noise pseudoproxies representing the 104 unique locations of the Mann et al. (1998) network between 70°N and 70°S. In MRWA07, this is referred to as pseudoproxy network ''A.'' We use a calibration period of 1900–80, a verification period of 850–1899, and a pseudoproxy signal-to-noise ratio of 0.4 (86% noise). In all cases presented in Fig. 1 and Table 1, we used the same pseudoproxy realization. Although we focus here on the Northern Hemisphere mean as in Christiansen et al. (2009), we include multivariate field scores to demonstrate that the RegEM settings that produce the best hemispheric mean scores also produce good multivariate scores.

We will begin with the results achieved using the MRWA07 implementation of RegEM-TTLS and move toward what appears to be the implementation of Christiansen et al. (2009). First, we will address variations in the stagnation tolerance for the low-frequency component of the hybrid frequency-band implementation. Next, we will move from the hybrid to the nonhybrid implementation. Finally, we will examine the effect of stagnation tolerance and truncation parameter on the nonhybrid approach. In these experiments we will focus on reconstructing the Northern Hemisphere mean series, as that is the focus of Christiansen et al. (2009). We use RegEM-TTLS to first reconstruct the field and then spatially average the field to reconstruct the hemispheric mean. If one is only interested in the hemispheric mean, note that it is possible to directly reconstruct the mean

TABLE 1. Verification scores for pseudoproxy experiments. Calibration period is 1900–80; verification period is 850–1899. Score shown in bold are not significant at the alpha = 0.05 level.

			NH multivariate		NH mean		
Hybrid/nonhybrid	Stagnation tolerance	Truncation (k)	RE	CE	RE	CE	r^2
Hybrid	1×10^{-4}	2 (low-f)	0.35	−0.04	0.96	0.70	0.72
Hybrid	5×10^{-3}	2 (low-f)	0.30	−0.13	0.78	−0.68	0.70
Hybrid	1×10^{-4}	10	0.05	−0.52	**0.21**	**−5.17**	**0.01**
Hybrid	5×10^{-3}	10	0.07	−0.49	**0.12**	**−5.86**	**0.00**
Nonhybrid	1×10^{-4}	4	0.28	−0.15	0.83	−0.31	0.78
Nonhybrid	5×10^{-3}	4	0.25	−0.20	0.65	−1.71	0.78
Nonhybrid	1×10^{-4}*	10	0.21	−0.26	0.80	−0.58	0.51
Nonhybrid	5×10^{-3}	10	0.08	−0.47	**0.30**	**−4.48**	0.12

* Stopping criterion not met after 1000 iterations.

series using RegEM in what amounts to an errors in variables (EIV) approach (Mann et al. 2008).

For the hybrid approach, we focus on the low-frequency component of the hybrid method because it is the low-frequency component of the reconstruction that is at issue here. Christiansen et al. (2009) use the stagnation tolerance specified in Rutherford et al. (2003) for comparing RegEM-TTLS to other methods. The choice of a stagnation tolerance, however, is not necessarily (or generally) uniform from application to application. By default, RegEM starts iterating by infilling missing values with the mean of the available data. If the means of the calibration and reconstruction periods are the same, relatively few iterations might be needed to achieve convergence. If, however, there is a relatively large difference between the means, then several hundred iterations may be needed and interim solutions should be examined to determine if the solution has stabilized. If the solution has not stabilized with the given stagnation tolerance, a lower tolerance should be used to force more iterations and interim results again examined. RegEM-TTLS is not computationally intensive when used to reconstruct a few principal components (PCs) of the target field, and it is not detrimental to set the stagnation tolerance at a conservatively low value and examine interim solutions to ensure convergence. Christiansen et al. (2009) commit an important error by failing to do so. Our results in Fig. 1 (see also Table 1) clearly demonstrate that, had we used the same stagnation tolerance used in Christiansen et al. (2009), RegEM-TTLS would have performed poorly in MRWA07. Furthermore, the results of Christiansen et al. (2009, their Fig. 16) show that a smaller stagnation tolerance results in a much better mean reconstruction by RegEM-TTLS. Indeed, it appears that the RegEM-TTLS results would be on par with the best results achieved by the other methods (e.g., Fig. 2 in Christiansen et al. 2009) had a proper stagnation tolerance been used.

Furthermore, when assessing stagnation, it is essential to examine the solution itself and not skill measures derived from the solution. Christiansen et al. (2010) suggest that the value of the correlation coefficient during the validation interval can be used as a measure of stagnation. This is an inappropriate procedure for two fundamental and well-known reasons: 1) the choices made in a methodological optimization procedure *should not* depend on information from outside the calibration interval (i.e., on information from the validation interval). Otherwise, the assumption of statistical independence of calibration and validation is violated, and the ''validation'' procedure is compromised (in this case, the procedure instead represents a joint calibration–validation optimization). Second, if one *were* to choose this approach, the particular metric suggested by Christiansen et al. (2010)—the validation correlation coefficient—is an extremely poor measure of skill (see, e.g., Wahl and Ammann 2007; MRWA07), because it is insensitive to the accuracy of the estimated mean and variance in the reconstruction. As the iterations advance, much of the convergence in the solution arises from an increasingly more accurate estimate of the mean and variance of the data outside the calibration interval, particularly with regard to the lower frequencies for the case at hand (see below). Yet, this information is completely discarded by the validation metric suggested by Christiansen et al. (2010). Convergence in this case cannot be meaningfully measured, as the stabilization of the correlation coefficient values may gave the false impression that convergence been achieved, when it in fact it has not.

One of the key innovations in the use of RegEM for the problem of paleoclimate reconstruction has been the development of a hybrid frequency-band approach (MRWA07; Mann et al. 2008). In testing with a relatively long 125-yr (1856–1980) calibration period, MRWA07 found that the hybrid method modestly outperformed the nonhybrid method (MRWA07, supplementary Fig. 10).

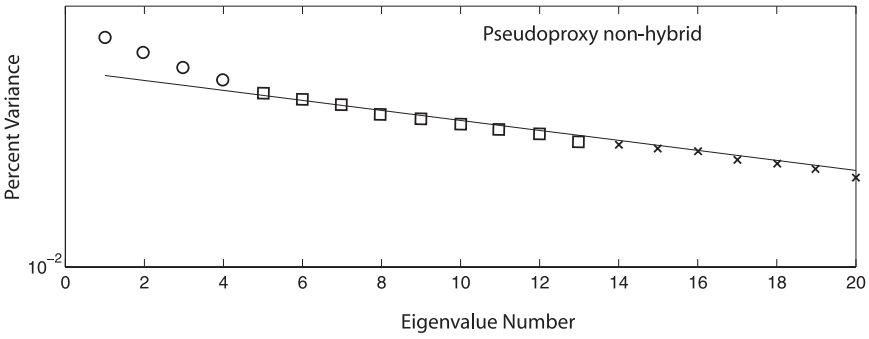

FIG. 2. LEV plot for the pseudoproxy network used here for an 1856–1980 calibration period as in MRWA07 and in Mann et al. (2008). Note the relatively few eigenvalues (circles) that lie above the background trend and are selected by our method compared to the truncation parameter of 13 (squares) chosen by Christiansen et al. (2009). In this case, we fit the first $N/2$ eigenvalues and retain $k + 1$ eigenvalues that lie above the fit by a tolerance equal to the standard deviation of the regression residuals.

In addition, there are good reasons, beyond the inclusion of decadal-resolution proxies (e.g., the potentially quite distinct patterns characterizing interannual versus interdecadal climate variability), to favor a hybrid approach for climate reconstructions as discussed in Rutherford et al. (2005).

It appears that Christiansen et al. (2009) only tested the nonhybrid method on their 100-yr (1900–2000) calibration period. To further assess the impact of these two different approaches, we applied them to the short, 80-yr, calibration period (1900–80) examined in MRWA07 using a stagnation tolerance and truncation parameter as in MRWA07. This direct comparison with the shorter calibration period clearly shows the nonhybrid approach used by Christiansen et al. (2009) underestimates the long-term variance while the hybrid approach does not (Fig. 1b, Table 1). Furthermore, as we change the stagnation tolerance to more closely approximate Christiansen et al. (2009), we find that the nonhybrid performance continues to deteriorate because the solution has not stabilized (Fig. 1b, Table 1). The difference between the performance of the hybrid and nonhybrid methods shown (the blue and red lines in Fig. 1b) is consistent with the mean reconstructions for 30 pseudoproxy noise realizations (not shown).

Last, we examine the choice of truncation parameter in the nonhybrid case. In MRWA07 we elaborated on the method for selecting the number of instrumental PCs to reconstruct and for setting the truncation parameter and provided computer code where appropriate (see eigenselect.m in the MRWA07 supplementary online material at http://www.meteo.psu.edu/~mann/PseudoproxyJGR06). The selection methods discussed by MRWA07 do not require any information outside of the calibration period and are easily applied to real-world situations.

For both the nonhybrid and high-frequency component of the hybrid, we use the log-eigenvalue (LEV) method (Wilks 2006) to choose a truncation parameter. In our experience with both pseudoproxy networks and real proxy networks, the choice of truncation parameter (± 1) is visually apparent (Fig. 2) and is much lower than that found by Christiansen et al. (2009). To make the choice of "k" more objective, we use a regression-based approach were we retain the eigenvalues that lie above the background trend (http://www.meteo.psu.edu/~mann/PseudoproxyJGR06/code/eigenselect.m). Although variations on our scheme might result in a slightly different truncation parameter, it is difficult to justify a truncation parameter as large as that used by Christiansen et al. (2009). As with the stagnation tolerance, choosing a truncation parameter similar to that of Christiansen et al. (2009) in the nonhybrid method degrades RegEM-TTLS performance (Fig. 1b, Table 1). For the low-frequency component of the hybrid method, we originally advocated the retention of enough eigenvalues to explain 50% of the low-frequency variance (MRWA07) and have subsequently found that 33% provides for even better results (Mann et al. 2009). In our pseudoproxy experiments (e.g., MRWA07) this typically corresponds to the retention of two low-frequency eigenvalues.

In addition, we have attempted to directly choose the regression parameter "h" in RegEM-Ridge, as suggested in Christiansen et al. (2010) with generally inferior results compared to our implementation of RegEM using TTLS. Attempts have also been made elsewhere to implement generalized cross validation (GCV) into RegEM-TTLS with results, so far, generally inferior to our selection criteria (Emile-Geay et al. 2008).

To summarize, the RegEM-TTLS performance observed in MRWA07 can be degraded to that shown in

Christiansen et al. (2009) by 1) replacing the hybrid approach advocated in MRWA07 with a nonhybrid approach, 2) choosing a stagnation tolerance that is too large and causes RegEM to stop iterating before the *solution* has converged, and 3) choosing a truncation parameter that is too large. Item 2 was in fact investigated by Christiansen et al. (2009, section 9) and they recognized that the stagnation tolerance used in the method comparisons was inappropriate. Curiously, however, they did not compare RegEM-TTLS with a proper stagnation tolerance to the other methods.

In their reply, Christiansen et al. (2010) attempt to defend their previous conclusions with new experiments. We can compare their hybrid results using $k = 2$ (which we expect would be consistent with our low-frequency selection method as discussed above) and the smaller stagnation tolerance with the results in the original publication. It appears that our implementation of RegEM-TTLS would most closely compare to the leftmost light-blue box plot in Figs. 1 and 2 of Christiansen et al. 2010. Comparing the relative bias and relative amplitude (Fig. 1 in Christiansen et al. 2010; we do not consider their correlation comparisons for reasons discussed earlier) with corresponding results in Christiansen et al. 2009 (their Fig. 4) indicates that RegEM-TTLS performs as well or better than the other methods.

Similarly, if one uses an early calibration period we can compare the relative trend (Fig. 2 in Christiansen et al. 2010) with those in the original publication (Christiansen et al. 2009, their Fig. 10) and find that RegEM-TTLS performs as well as or better than the other methods for the same experiment ($a = 1$), when applied as we have advocated. RegEM-TTLS underestimates the trend by an ensemble mean of approximately 0.25 (Christiansen et al. 2010) compared to approximately 0.45 (ensemble mean) for the best method shown in Christiansen et al. (2009, their Fig. 10). Although the ensemble spread has increased, the lower quartile of the RegEM-TTLS results is approximately equal to the best mean of the other methods. Indeed this early calibration approach is a most challenging test. In the context of climate field reconstruction (CFR), the method is being asked to reconstruct a pattern, globally synchronous warming associated with anthropogenic greenhouse gases, that does not exist during the calibration period.

In their reply, Christiansen et al. (2010) have added a new category of experiments distinct from those of the original publication and to which only RegEM is applied. In these experiments Christiansen et al. introduce an artificial warm period in their surrogates and show that RegEM underestimates this warming (their Fig. 3). The artificial warm period is created by varying the scaling of the principal components over time. This is a most challenging test, because it means that the covariance in the calibration period and most of the reconstruction period differs from that of the 50-yr period with the inflated principal components.

In MRWA07 we conducted experiments using the ECHAM and the global Hamburg Ocean Primitive Equation (ECHO-G) ''Erik'' simulation to test the ability of RegEM to reconstruct a period of past warmth approximately equal to that of the twentieth-century mean (model years 1000 to 1200 in the ECHO-G simulation). As implemented in MRWA07, RegEM-TTLS has no difficulty reconstructing the warm period.

We conducted an additional test of RegEM similar to reordering the reconstruction target as previously suggested (Ammann and Wahl 2007; Jones et al. 2009). We used the last 500 yr of the CSM run and reflected it about the y axis to create a 1000-yr-long series. The result is a hemispheric mean that is as warm at the beginning as it is at the end with a relatively stable, relatively cool period between (Fig. 3). Again, the hybrid method as implemented in MRWA07 has no difficulty reconstructing the ''ancient'' warm period. Furthermore, we allow the low-frequency truncation parameter to vary from 2 to 4 and show that the result is insensitive to this variation (Fig. 3). If we apply the nonhybrid method, however, RegEM fails to completely capture the cool period in the middle of series, even with a much lower stagnation tolerance (10^{-5}) and verification that the solution has stabilized. The magnitude of the difference between the mean hybrid and nonhybrid reconstructions from 30 pseudoproxy noise realizations (not shown) is consistent with that shown for the single realization (Fig. 3).

There are additional differences between the approaches of Christiansen et al. and MRWA07 (including this comment) that may produce differing results (including the magnitude of the hybrid/nonhybrid difference). These include the pseudoproxy network size and pseudoproxy locations, the calibration interval, the use of field noise realizations and pseudoproxy noise realizations, and the spatial resolution of the field. Continued experimentation may shed some light on the impact of these differences.

Finally, we wish to reiterate that the use of correlation coefficients as a measure of reconstructive skill, as done by Christiansen et al. (2010) is problematic, leading to unacceptable type I and type II errors in hypothesis testing of reconstruction skill (MRWA07; Wahl and Ammann 2007). Correlation coefficients do not penalize a reconstruction that poorly reproduces the mean and variance of the target series (MRWA07; Wahl and Ammann 2007—see Fig. 1S therein). For example, compare our hybrid verification results (top two lines of

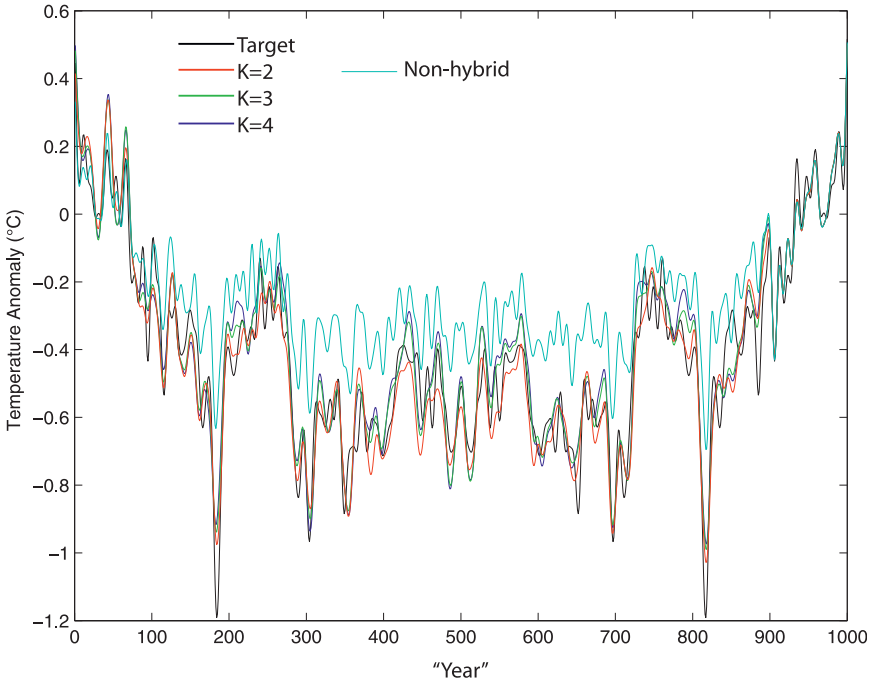

FIG. 3. Additional experiment wherein the last 500 yr of the CSM data is reflected about the *y* axis to create a 1000-yr-long series with warming at both ends. RegEM-TTLS as implemented in MRWA07 successfully reproduces the early warm period. Furthermore, the result is insensitive to the choice of low-frequency truncation parameter. When the nonhybrid method is used, RegEM fails to fully capture the low-frequency variation. (The calibration period is the last 100 yr.)

Table 1). While the validation r^2 values for the two cases are statistically indistinguishable, the former reconstruction (which uses an appropriately lower stagnation tolerance) is clearly better than the latter, as judged by appropriate validation metrics [reduction of error (RE); coefficient of efficiency (CE)] and by a simple visual comparison (Fig. 1a).

In addition to the technical issues described above, we also wish to comment on the selective characterization by Christiansen et al. (2009) of the published literature on this topic. First, two papers, Wahl and Ammann (2007) and Ammann and Wahl (2007), that were important parts of the ''sometimes uproarious debate that arose after Mann et al. (1998)'' and the ''fierce debate'' regarding amplitude loss, and which indicate the relative robustness of the Mann et al. (1998) reconstructions, are not mentioned. Second, MRWA07 not only did conduct tests regarding the ''the efficacy of validating the reconstructions on independent data'' but indeed highlighted this issue (see, e.g., Table 1 in MRWA07). Most puzzling of all is the blanket refutation of the conclusions of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) toward the end of Christiansen et al. (2010): ''statements like 'it is very likely that average Northern Hemisphere temperatures during the second half of the 20th century were higher than for any other 50-year period in the last 500 years' (Jansen et al. 2007) can not be justified from statistical reconstructions alone.'' Such a statement would require Christiansen et al. to have invalidated the reconstructions of at least a dozen different groups, which use many different methods (some not examined by Christiansen et al.) and a wide variety of paleoclimate data. Yet, as we have demonstrated, the authors have not invalidated even the RegEM-based estimates focused upon here.

Methodological comparisons, such as that undertaken by Christiansen et al. are a worthwhile and important endeavor, but they need to be implemented properly to allow comparative performance across methods to be meaningfully evaluated (e.g., Lee et al. 2008; Jones et al. 2009). MRWA07 focused on the various issues raised by Christiansen et al., that is, the impact on reconstruction quality of signal-to-noise ratios, spatial distributions of sites, noise characteristics, and the impact of noise realization (what Christiansen et al. call ''stochasticity''). The results reported by Christiansen et al. (2009) obscure meaningful diagnosis of the impact of these issues, however, by introducing into the RegEM algorithm a sequence of erroneous procedures. These procedures have the net effect of producing an altered ''RegEM'' algorithm

that does not resemble the RegEM algorithm advocated in MRWA07. The results presented in MRWA07 and here (and actually confirmed by Christiansen et al. 2010) show that, properly implemented, RegEM-TTLS, when tested with networks of data possessing characteristics consistent with those of actual proxy data, yields long-term climate reconstructions of considerable fidelity.

REFERENCES

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

——, F. Joos, D. Schimel, B. L. Otto-Bliesner, and R. Tomas, 2007: Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model. *Proc. Natl. Acad. Sci. USA,* **104,** 3713–3718.

Christiansen, B., T. Schmith, and P. Thejll, 2009: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *J. Climate,* **22,** 951–976.

——, ——, and ——, 2010: Reply. *J. Climate,* **23,** 2839–2844.

Emile-Geay, J., K. M. Cobb, M. E. Mann, and S. Rutherford, 2008: Low-frequency tropical Pacific sea-surface temperature reconstruction and error estimates. *Eos, Trans. Amer. Geophys. Union,* **89** (Fall Meeting Suppl.), Abstract PP21D-01.

Jones, P. D., and Coauthors, 2009: High-resolution paleoclimatology of the last millennium: A review of current status and future prospects. *Holocene,* **19,** 3–49.

Lee, T. C. K., F. Zwiers, and M. Tsao, 2008: Evaluation of proxy-based millennial reconstruction methods. *Climate Dyn.,* **31,** 263–281, doi:10.1007/s00382-007-0351-9.

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.

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

——, Z. Zhang, R. S. Bradley, M. K. Hughes, S. Miller, S. D. Rutherford, and F. Ni, 2008: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl. Acad. Sci. USA,* **105,** 13 252–13 257.

——, and Coauthors, 2009: Global signatures of the ''Little Ice Age'' and ''Medieval Climate Anomaly'' and plausible dynamical origins. *Science,* **326,** 1256–1260.

Riedwyl, N., M. Küttel, J. Luterbacher, and H. Wanner, 2009: Comparison of climate field reconstruction techniques: Application to Europe. *Climate Dyn.,* **32,** 381–395, doi:10.1007/s00382-008-0395-5.

Rutherford, S., M. E. Mann, T. L. Delworth, and R. J. Stouffer, 2003: Climate field reconstruction under stationary and nonstationary forcing. *J. Climate,* **16,** 462–479.

——, ——, T. J. Osborn, R. S. Bradley, K. R. Briffa, M. K. Hughes, and P. D. Jones, 2005: Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season and target domain. *J. Climate,* **18,** 2308–2329.

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 627 pp.