

DISCUSSION OF: A STATISTICAL ANALYSIS OF MULTIPLE TEMPERATURE PROXIES: ARE RECONSTRUCTIONS OF SURFACE TEMPERATURES OVER THE LAST 1000 YEARS RELIABLE?*

BY JASON E. SMERDON^{†,‡}

Lamont-Doherty Earth Observatory of Columbia University[‡]

McShane and Wyner (2010; hereinafter MW10) reiterate a well-known and central challenge of paleoclimatology: it is fraught with uncertainties and based on noisy observations. Decades of research have aimed at characterizing these uncertainties and interpreting proxies through laboratory experiments, field observations, theory, process-based modeling, cross-record comparisons, and indeed through statistical modeling and hypothesis testing. It is against this larger backdrop that the problem addressed by MW10 must be considered. Attempts to reconstruct global or hemispheric temperature indices and fields using multi-proxy networks are an outgrowth of many efforts in paleoclimatology, but represent relatively recent pursuits in the field. They provide neither the principal scientific evidence supporting climate-proxy connections, nor the most compelling, and the inference by MW10 that their own findings demonstrate a widespread failure in the predictive capacity of climate proxies is at odds with most other independent lines of proxy research.

The above considerations notwithstanding, I focus on one principal argument by MW10 that uses cross-validation experiments to conclude that “*proxies are severely limited in their ability to predict average temperatures and temperature gradients.*” I demonstrate that this claim is based on a hypothesis test subject to Type II errors and therefore an inconclusive evaluation of the temperature sensitivity of proxy archives.

I perform additional cross-validation experiments using 283 time series that are randomly selected from the global CRU temperature field as infilled and subselected (1,732 total grid cells) by Mann et al. (2008). I choose 283 samples based on the total number of unique $5^\circ \times 5^\circ$ grid cells that

*Lamont-Doherty Earth Observatory contribution number XXXX

[†]This work was supported in part by NSF grant ATM-0902436 and NOAA grant NA07OAR4310060.

Keywords and phrases: Climate Change, Paleoclimate, Statistical Climate Reconstructions, Late-Holocene, Common Era, Multi-proxy Reconstructions, Model Validation

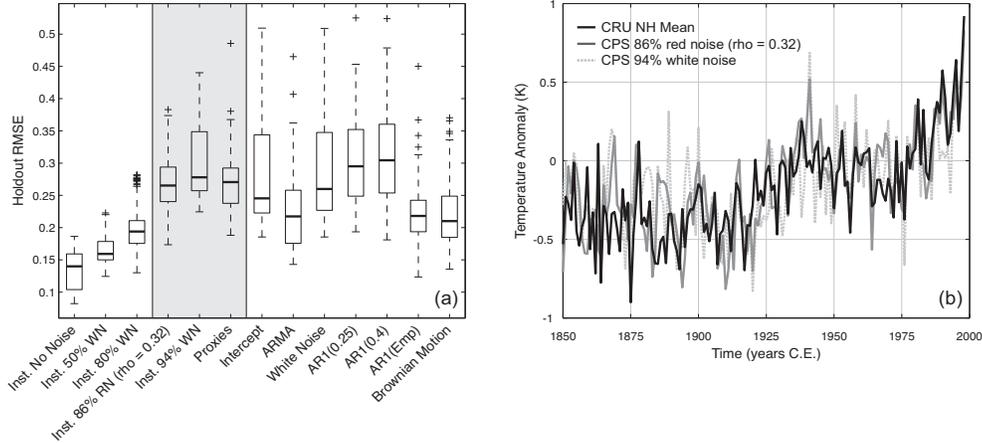


FIG 1. (a) Cross-validated RMSE on thirty-year hold out blocks for a subset of the original MW10 experiments (recomputed by the current author) and for the newly completed experiments using the instrumental data perturbed with various levels of noise. The 86% red-noise experiment is considered to be representative of the average proxy predictor (Mann et al. 2007) and compares well to the proxy result. (b) Area-weighted CPS reconstructions using the 86% red-noise and 94% white-noise predictor sets.

contain the 1,209 proxies in the Mann et al. (2008) proxy network. Time series from these cells are used to create five predictor datasets spanning the instrumental period by adding 0, 50, 80 and 94% white noise by variance, and 86% red noise by variance ($\rho = 0.32$), the latter of which has been argued to be an average representation of the noise in proxy records (Mann et al. 2007). These datasets are used to repeat the MW10 analysis using the Lasso to target the CRU NH mean index across 120 cross-validation experiments. My results are summarized in Figure 1a and compared to my own reproduction of some MW10 experiments.

I first note that even the no-noise experiment is subject to errors. This is an important baseline for the MW10 experiments, demonstrating that even ‘perfect proxies’ are subject to errors due to incomplete field sampling. Most importantly, however, the skill of the instrumental predictors diminishes with noise such that the 86% red-noise and 94% white-noise predictors perform *comparable to or worse than* the proxy network, and in turn perform worse than the AR1(Emp) and Brownian Motion null models tested by MW10. Additionally, simple area-weighted composite-plus-scale (CPS) ‘reconstructions’ from the Northern Hemisphere (NH) subset of the 86% red-noise and 94% white-noise predictor networks yield NH mean indices that

compare well with the target (Figure 1b; respective correlations between the target and CPS reconstructions are 0.73 and 0.65), indicating that skillful reconstructions are possible from networks with such noise levels. These findings are fundamental to the MW10 argument that the proxies are poor temperature recorders because they do not perform better than some noise models. To the contrary, the results that I present demonstrate that predictor networks explicitly containing temperature signals – perturbed with approximate proxy noise levels – also do not beat the AR1(Emp) and Brownian Motion noise models in cross-validation experiments and that skillful CPS reconstructions can be derived from such predictors. The appropriate conclusion is therefore not that the proxies are limited in their ability to predict NH temperatures, but that the test performed by MW10 is subject to Type II errors and is unsuitable for measuring the degree to which the proxies sample temperature. Note that this conclusion, although challenged by MW10, also supports arguments by Ammann and Wahl (2007) about the dangers of Type II errors in this paleoclimatic context.

It is worth considering the likely reasons why the AR1(Emp) and Brownian Motion noise models perform better than predictors explicitly containing temperature signals. As discussed by MW10, the large amounts of persistence in these models are good approximations of a principal characteristic of the target time series, namely its temporal autocorrelation. This fact, combined with the short cross-validation period, allows highly persistent time series to test well. But this success is likely also dependent on selections from many noise draws. MW10 have focused on the Mann et al. (2008) study that includes 1,209 total proxies (1,138 if the Lutannt series are excluded) and thus a large number of possible predictors in the MW10 noise experiments. In contrast, other NH temperature reconstructions have been successfully cross validated using only a few tens of proxies (e.g. Esper et al. 2002; Moberg et al. 2005; Hegerl et al. 2007). It therefore is yet unclear how the MW10 cross-validation tests would compare in scenarios using far fewer predictors.

MW10 present a number of experiments that deserve further testing and analyses. The issues that I raise similarly come with their own set of caveats that cannot be explored in a short discussion paper. For example, a field sampling reflecting the true Mann et al. (2008) proxy locations with reduced ocean sampling and regional clustering might worsen the cross-validation skill of the instrumental predictors compared to the random sampling used in my experiments. Conversely, the true proxy distribution is more concentrated in the NH, which may improve prediction of the NH mean index. I have also sampled each grid cell once, as opposed to multiple sampling

reflecting the occurrence of several proxies in a single grid cell. This latter sampling will in effect reduce the noise in the relevant cells, thus making the noise dependence of cross-validation skill less straightforward to interpret. These dependencies should be tested in future work. Nevertheless, the preliminary results that I have outlined suggest that the MW10 hypothesis test is subject to Type II errors and thus is not suitable for evaluating the reliability of proxy archives as temperature predictors.

Acknowledgements. I thank Alexey Kaplan for providing many insightful discussions on the subject of this manuscript and Editor Michael Stein for inviting my contribution. Code and data files for my experiments are posted online at www.ldeo.columbia.edu/~jsmerdon/2010_aoas_supplement.html

References.

- [1] AMMANN, C.M. AND E.R. 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.
- [2] ESPER, J., E.R. COOK AND F.H. SCHWEINGRUBER (2002). Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295** 2250–2253.
- [3] HEGERL, G.C., T. CROWLEY, M. ALLEN, W.T. HYDE, H.N. POLLACK, J. SMERDON, AND E. ZORITA (2007). Detection of human influence on a new, validated 1500-year temperature reconstruction. *Journal of Climate* **20** 650–666.
- [4] MANN, M. E., Z. ZHANG, M. K. HUGHES, R. S. BRADLEY, S. K. MILLER, S. RUTHERFORD, AND F. NI (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences USA* **105**, 36, 13252–13257.
- [5] MANN, M. E., S. RUTHERFORD, E. WAHL, AND C. AMMANN (2007). Robustness of proxy-based climate field reconstruction methods. *Journal of Geophysical Research* **112**, D12109, doi:10.1029/2006JD008272.
- [6] MCSHANE, B.B. AND A.J. WYNER (2010). A Statistical Analysis of Multiple Temperature Proxies: Are Reconstructions of Surface Temperatures Over the Last 1000 Years Reliable? *Annals of Applied Statistics* To appear.
- [7] MOBERG, A, D.M. SONECHKIN, K. HOLMGREN, N.M DATSENKO, W. KARLEN (2005). Highly variable Northern Hemisphere temperature reconstructed from low- and high-resolution proxy data. *Nature* **433**, 7026, 613–617.

LAMONT-DOHERTY EARTH OBSERVATORY
 61 ROUTE 9W
 P.O. Box 1000
 PALISADES, NY 10964
 E-MAIL: jsmerdon@ldeo.columbia.edu