Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

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

By Alexey Kaplan†

Lamont-Doherty Earth Observatory of Columbia University

McShane and Wyner (2010; hereinafter MW2010) demonstrated that in many cases a comprehensive data set of p =1138 proxies (Mann et al., 2008) did not predict Northern Hemisphere (NH) mean temperatures significantly better than random numbers. This fact is not very surprising in itself: the unsupervised selection of good predictors from a set of $p \gg n$ proxies of varying sensitivities might be too challenging a task for any statistical method ($p/n_c \approx 10$; only n_c=119 out of total n=149 years were used for calibration in MW2010 cross-validated reconstructions). However, some types of noise[1] systematically outperformed the real proxies (see two bottom panels of MW2010 Figure 10). This finding begs further investigation: what do these random numbers have that real proxies do not?

To investigate this question, the present analysis uses ridge regression (RR, Hoerl and Kennard, 1970) instead of the Lasso.[2] The regression model used by MW2010 with Lasso and here with RR is

$$y = X\beta + \beta_0 \mathbb{1}_n + \varepsilon,$$

where y is a column vector of n observations (annual NH temperatures), ε is random error, X is a known $n \times p$ matrix of predictors (climate proxies). A vector of regression coefficients β and an intercept constant β_0 are to be determined. A column n-vector $\mathbb{1}_n$ has all components equal one. Proxy records

*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: Paleoclimate, Statistical Climate Reconstructions, Cross-Validation, Ridge Regression, Autoregressive Processes, Kriging

[1] Pseudoproxies used by MW2010 are called "noise" here; in climate research, pseudoproxies are synthetic combinations of a climate signal with some noise; without the former, it is a pure noise.

[2] The difference is in the penalty norm: Lasso uses L_1 while RR uses L_2. MW2010 have also argued that a rough performance similarity should exist between different methods for $p \gg n$ problems (p.25)

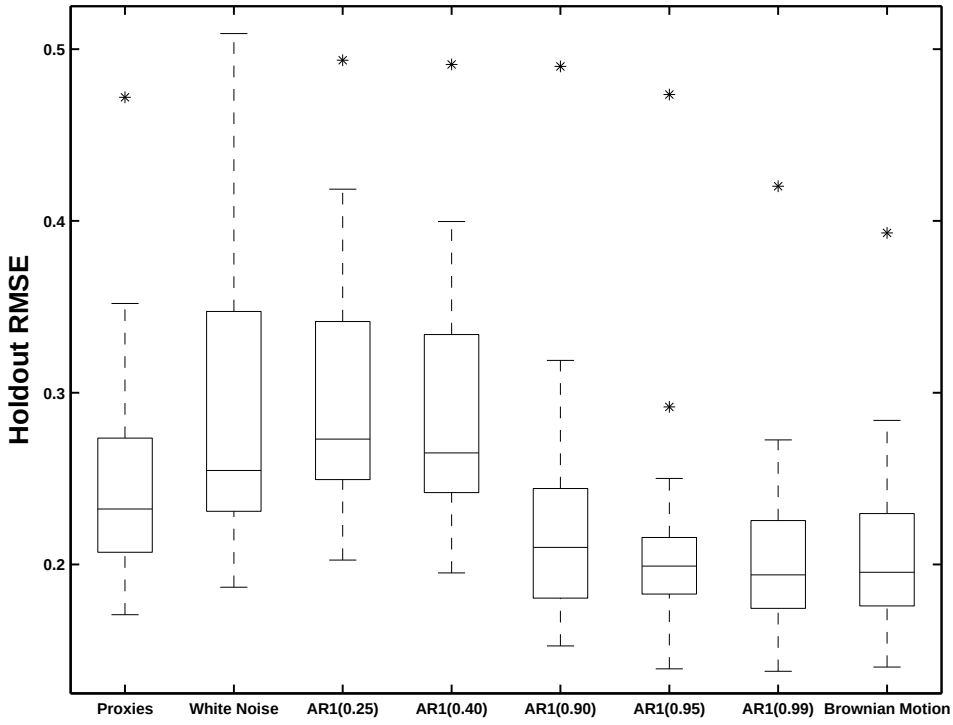

FIG 1. *Cross-validated RMSE on 120 30-year holdout blocks for the RR reconstructions from real climate proxies and from the random noise (one realization for each noise experiment); cf. MW2010, Figure 9.*

are standardized before use; in cross-validation experiments standardization is repeated for each calibration period.

Let w be a column n_c-vector such that $w^T \mathbb{1}_{n_c} = 1$. Define matrix-valued functions $\mathcal{W}[w] = I - \mathbb{1}_{n_c} w^T$ and $\mathcal{R}[S, \lambda, w] = S_{vc}(S_{cc} + \lambda I)^{-1} \mathcal{W}[w] + \mathbb{1}_{n_v} w^T$, where S is a positive semidefinite $n \times n$ matrix, $\lambda > 0$ is the ridge parameter found as a minimizer of the generalized cross-validation function (GCV, Golub et al. 1979), matrix (or vector) subscripts c or v hereinafter indicate submatrices corresponding to the calibration or validation periods, respectively. The RR reconstruction $\hat{y}_v$ of temperatures in the validation period (a "holdout block" of n_v=30 consecutive years) is a linear transformation: $\hat{y}_v = \mathcal{R}[S_p, \lambda, e] y_c$, where $S_p = \widetilde{X} \widetilde{X}^T / p$, $\widetilde{X}$ is the standardized version of X, and $e = n_c^{-1} \mathbb{1}_{n_c}$.

Using these formulas, the RR version of the MW2010 cross-validation tests were performed for real proxies and for some noise types. Results are shown in Figure 1. The cross-validated RMSE of the RR reconstructions are smaller

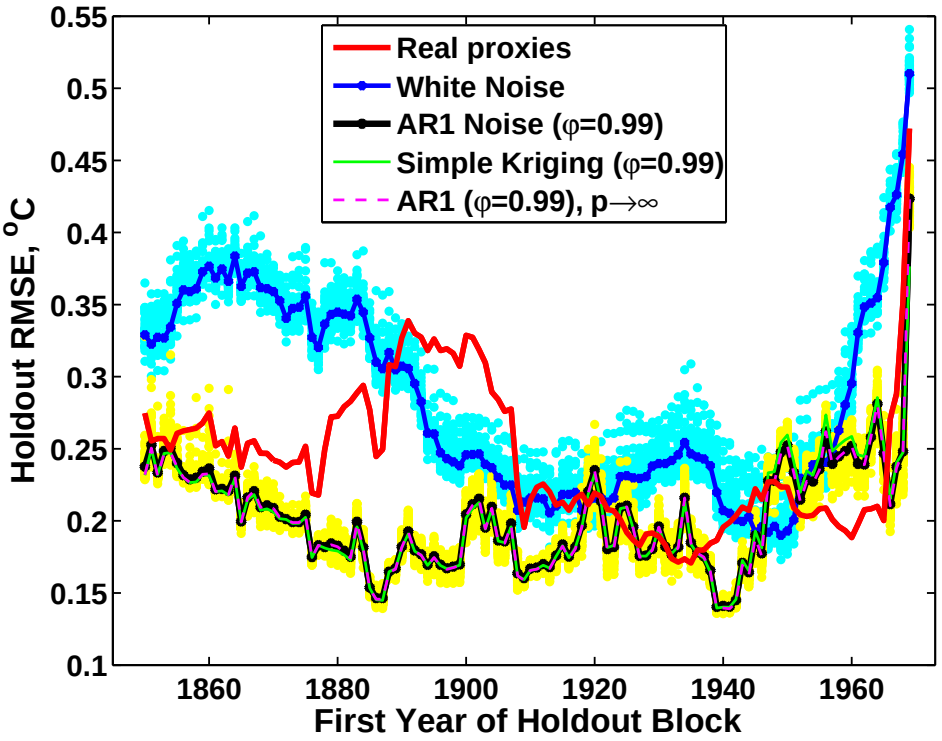

FIG 2. *Holdout RMSE for RR reconstructions as a function of time for real proxies (red) and two 100-member ensemble means: white noise (blue) and AR(1) noise with* φ *=0.99 (black). Individual ensemble members are shown by magenta and yellow dots, respectively. The probability limit* ($p \to \infty$) *for the latter is shown by magenta dashes. Holdout RMSE for simple kriging of the NH mean temperature index using an exponential semivariogram (Le and Zidek, 2006)* $\gamma(\tau) = \lambda_{min} + 1 - \exp[\tau \ln \varphi]$ *with the GCV-selected nugget* $\lambda_{min} = \ell(\Phi, 0)$ *and long decorrelation scale* $-1/\ln(\varphi) = 99.5$ *years (*τ *is time in years) is shown by green line.*

than Lasso values (cf MW2010, Figure 9), but the relative performance in different experiments appears consistent between RR and Lasso. As in the Lasso case, noise with high temporal persistence ($\varphi \geq 0.9$ and Brownian motion) outperformed the proxies. Figure 2 illustrates the time dependence of the holdout error for the real-proxy, white-noise, and $\varphi = 0.99$ AR(1) cases. There is a general similarity between these and the corresponding curves in Figure 10 by MW2010.

Note that a traditional approach to hypothesis testing would evaluate an RMSE corresponding to a regression of temperature data (y) on real proxies (X) in the context of the RMSE probability distribution induced by the assumed distribution of y under the hypothesized condition (e.g., $\beta = 0$). However, MW2010 evaluate the RMSE of real proxies in the context of the RMSE distribution induced by random values in X, not y. Such an approach to testing a null hypothesis would be appropriate for an inverse relationship, that is $X = y\beta^T + \mathbb{1}_n\beta_0^T + \varepsilon$. When used with a direct regression model here, however, it results in the RMSE distribution with a surprising feature: when $p \to \infty$, RMSE values for individual realizations of the noise matrix X converge in probability to a constant.

This convergence occurs because the columns x of X in the noise experiments are i.i.d. from the noise distribution; AR(1) with $\varphi = 0.99$ is considered here: $x \sim \mathcal{N}(0, \Phi)$, $\Phi = (\varphi^{|i-j|})$. The columns of $\widetilde{X}$ are i.i.d. too, hence the random matrix $S_p = \widetilde{X}\widetilde{X}^T/p$ is an average of p i.i.d. variates $\tilde{x}\tilde{x}^T$. Expectation $\Psi = \mathbf{E}\tilde{x}\tilde{x}^T$ exists; its elements are computed as expectations of ratios and first inverse moments of quadratic forms in normal variables (Jones, 1986, 1987). The weak law of large numbers applies, so $S_p \xrightarrow{P} \Psi$. Since the GCV function depends on S and w as well as on λ, its minimizing λ will depend on these parameters too: $\lambda_{\min} = \ell[S, w]$. Here GCV is assumed well-behaved, so that ℓ is a single-valued function, continuous at (Ψ, e). From the definition of $\mathcal{R}$, $\mathcal{B}[S, e] \equiv \mathcal{R}[S, \ell[S, e], e]$ will also be continuous at $S = \Psi$, thus $S_p \xrightarrow{P} \Psi$ implies $\hat{y}_v = \mathcal{B}[S_p, e]y_c \xrightarrow{P} \mathcal{B}[\Psi, e]y_c$.

When p is finite but large, like $p = 1138$, reconstructions based on individual realizations of a noise matrix X are dominated by their constant components, especially when $\varphi \approx 1$: note the small scatter of RMSE values in the ensemble of AR(1) with $\varphi = 0.99$ (yellow dots in Figure 2). The probability limit $\hat{y}_v = \mathcal{B}[\Psi, e]y_c$ yields RMSE values (magenta dash in Figure 2) that are very close ($1.3{\cdot}10^{-3}$ °C RMS difference) to the ensemble mean RMSE (black curve in Figure 2). To interpret this non-random reconstruction, consider its simpler analogue, using neither proxy standardization nor a regression intercept (β_0). Then, if the assumptions on the GCV function change accordingly, $\hat{y}_v \xrightarrow{P} \mathcal{B}[\Phi, 0]y_c = \Phi_{vc}[\Phi_{cc} + \ell(\Phi, 0)I]^{-1}y_c$, i.e., a predic-

tion of y_v from y_c by "simple kriging" (Stein 1999, p.8), which in atmospheric sciences is called objective analysis or optimal interpolation (Gandin, 1965). The RMSE corresponding to this solution is shown in Figure 2: it is also close to the ensemble mean RMSE for AR(1) noise with $\varphi = 0.99$ (RMS difference is $5.4{\cdot}10^{-3}\,^{\circ}$C). The solution $\mathcal{B}[\Psi, e]y_c$, to which the noise reconstructions without simplifications converge as $p \to \infty$, is more difficult to interpret. Still, it has a structure of an objective analysis solution and its results are quite close to simple kriging: the RMS difference between the two reconstructions over all holdout blocks is $7.7{\cdot}10^{-3}\,^{\circ}$C.

Due to the large value of p in the MW2010 experiments, their tests with the noise in place of proxies essentially reconstruct holdout temperatures by a kriging-like procedure in the temporal dimension. The covariance for this reconstruction procedure is set by the temporal autocovariance of the noise. Long decorrelation scales ($\varphi \geq 0.95$) gave very good results, implying that long-range correlation structures carry useful information about predictand time series that is not supplied by proxies. By using such a noise for their null hypothesis, MW2010 make one skillful model (multivariate linear regression on proxies) compete against another (statistical interpolation in time) and conclude that a loser is useless. Such an inference does not seem justified.

Modern analysis systems do not throw away observations simply because they are less skillful than other information sources: instead, they combine information. MW2010 experiments have shown that their multivariate regressions on the proxy data would benefit from additional constraints on the temporal variability of the target time series, e.g., with an AR model. After proxies are combined with such a model, a test for a significance of their contributions to the common product could be performed.

Acknowledgements. Generous technical help and many useful comments from Jason Smerdon and very helpful presentation style guidance from Editor Michael Stein are gratefully acknowledged.

SUPPLEMENTARY MATERIAL

Supplement A: Codes, data, and detailed derivations (http://www.imstat.org/aoas/supplements/default.htm). Temporarily it is stored at `http://rainbow.ldeo.columbia.edu/~alexeyk/MW2010discAOAS/kaplan_discMW2010_code_final_1Nov2010.tar.gz`

References.

[1] Gandin, L.S. (1963). *Objective Analysis of Meteorological Fields.* Gidrometeorologicheskoye Izdatel'stvo, Leningrad. Translated from Russian, Israeli Program for Scientific Translations. Jerusalem, 1965.

[2] Golub, G., M. Heath, and G.Wahba (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-224.
[3] Hoerl, A.E., and R.W.Kennard (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55-67.
[4] Jones, M.C., (1986). Expressions for inverse moments of positive quadratic forms in normal variables. *Austral. J. Statist.* **28**, 242-250.
[5] Jones, M.C., (1987). On moments of ratios of quadratic forms in normal variables. *Statistics & Probability Letters* **6**, 129-136.
[6] Le, N.D. and J.V.Zidek (2006). *Statistical Analysis of Environmental Space-Time Processes* Springer, New York.
[7] 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.
[8] 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.
[9] Stein M.L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging* Springer, New York.
[10] Tibshirani, R. (1996). Regression Shrinkage and Selection via Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Lamont-Doherty Earth Observatory
61 Route 9W
P.O. Box 1000
Palisades, NY 10964
E-mail: alexeyk@ldeo.columbia.edu