On the Origin of the Standardization Sensitivity in RegEM Climate Field Reconstructions*

JASON E. SMERDON

Lamont-Doherty Earth Observatory, Columbia University, Palisades, and Barnard College, New York, New York

ALEXEY KAPLAN

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

DIANA CHANG

Barnard College, New York, New York

(Manuscript received 9 August 2007, in final form 30 April 2008)

ABSTRACT

The regularized expectation maximization (RegEM) method has been used in recent studies to derive climate field reconstructions of Northern Hemisphere temperatures during the last millennium. Original pseudoproxy experiments that tested RegEM [with ridge regression regularization (RegEM-Ridge)] standardized the input data in a way that improved the performance of the reconstruction method, but included data from the reconstruction interval for estimates of the mean and standard deviation of the climate field—information that is not available in real-world reconstruction problems. When standardizations are confined to the calibration interval only, pseudoproxy reconstructions performed with RegEM-Ridge suffer from warm biases and variance losses. Only cursory explanations of this so-called standardization sensitivity of RegEM-Ridge have been published, but they have suggested that the selection of the regularization (ridge) parameter by means of minimizing the generalized cross validation (GCV) function is the source of the effect. The origin of the standardization sensitivity is more thoroughly investigated herein and is shown not to be associated with the selection of the ridge parameter; sets of derived reconstructions reveal that GCV-selected ridge parameters are minimally different for reconstructions standardized either over both the reconstruction and calibration interval or over the calibration interval only. While GCV may select ridge parameters that are different from those that precisely minimize the error in pseudoproxy reconstructions, RegEM reconstructions performed with truly optimized ridge parameters are not significantly different from those that use GCV-selected ridge parameters. The true source of the standardization sensitivity is attributable to the inclusion or exclusion of additional information provided by the reconstruction interval, namely, the mean and standard deviation fields computed for the complete modeled dataset. These fields are significantly different from those for the calibration period alone because of the violation of a standard EM assumption that missing values are missing at random in typical paleoreconstruction problems; climate data are predominantly missing in the preinstrumental period when the mean climate was significantly colder than the mean of the instrumental period. The origin of the standardization sensitivity therefore is not associated specifically with RegEM-Ridge, and more recent attempts to regularize the EM algorithm using truncated total least squares could theoretically also be susceptible to the problems affecting RegEM-Ridge. Nevertheless, the principal failure of RegEM-Ridge arises because of a poor initial estimate of the mean field, and therefore leaves open the possibility that alternative methods may perform better.

* Lamont-Doherty Earth Observatory Contribution Number 7208.

Corresponding author address: Dr. Jason Smerdon, Lamont-Doherty Earth Observatory, P.O. Box 1000, Route 9W, Palisades, NY 10694.
E-mail: jsmerdon@ldeo.columbia.edu

DOI: 10.1175/2008JCLI2182.1

1. Introduction

Several recent attempts to perform climate field reconstructions (CFRs) of surface air temperature in the Northern Hemisphere (NH) have focused on a technique known as regularized expectation maximization (RegEM). This technique was developed by Schneider (2001) as a modification of a well-known family of expectation-maximization statistical techniques (e.g., Little and Rubin 2002) and was applied in the context of imputing missing values in a climate field. RegEM has subsequently been applied in various CFR applications (Mann and Rutherford 2002; Zhang et al. 2004; Rutherford et al. 2005, hereafter R05; Mann et al. 2005, hereafter M05, 2007a, hereafter M07a). Specifically within the context of NH CFRs of temperature, R05 used RegEM to derive a reconstruction of the NH temperature field back to A.D. 1400. This reconstruction was shown to compare well with the Mann et al. (1998) CFR and prompted R05 to argue that the comparison established a mutual validation of the two reconstructions.

In efforts to systematically evaluate the ability of the RegEM method to reconstruct past temperature fields, several tests have been performed using pseudoproxies (Mann and Rutherford 2002) derived from millennial integrations with general circulation models (GCMs; González-Rouco et al. 2003, 2006; von Storch et al. 2004; Ammann et al. 2007). M05 attempted to test the R05 RegEM method using pseudoproxies derived from the National Center for Atmospheric Research (NCAR) Climate System Model (CSM) 1.4 millennial integration. Subsequently, M07a have tested a different implementation of RegEM and shown it to perform favorably in pseudoproxy experiments. This latter study was performed in part because M05 did not actually test the R05 technique, which was later shown to fail appropriate pseudoproxy tests (Smerdon and Kaplan 2007). The basis of the criticism by Smerdon and Kaplan (2007) focused on a critical difference between the standardization procedures used in the M05 and R05 studies (here we define the standardization of a time series as both the subtraction of the mean and division by the standard deviation over a specific time interval). Their principal conclusions were as follows: 1) the standardization scheme in M05 used information during the reconstruction interval, a luxury that is only possible in the pseudoclimate of a numerical model simulation and not in actual reconstructions of the earth's climate; 2) when the appropriate pseudoproxy test of the R05 method was performed (i.e., the data matrix was standardized only during the calibration interval), the derived reconstructions suffered from warm biases and variance losses throughout the reconstruction interval; and 3) the similarity between the R05 and Mann et al. (1998) reconstructions, in light of the demonstrated problems with the R05 technique, suggests that both reconstructions may suffer from warm biases and variance losses.

The differences between the R05, M05, and M07a methods hinge on the details of the regularized regression that is used within the RegEM algorithm; regularization is typically necessary in CFR methods because the covariance matrix is rank deficient. Ridge regression was the regularization scheme used in the R05 and M05 RegEM approaches (RegEM-Ridge). Schneider (2001) also included truncated total least squares as an alternative regularization method within the RegEM algorithm (RegEM-TTLS), which is the basis of the new method applied by M07a. While comparisons between these various versions of RegEM should be an important subject of future work, we focus only on RegEM-Ridge within this study.

M07a attributed the problems with RegEM-Ridge to an imperfect selection of the regularization (ridge) parameter: ". . . we have found that the estimation of optimal ridge parameters is poorly constrained at decadal and longer timescales in our tests with pseudo-proxy data, and we have learned that earlier results using ridge regression (e.g. M05) are consequently sensitive to, e.g. the manner in which data are standardized over the calibration period." Mann et al. (2007b) further conclude that the selection of the ridge parameter using generalized cross validation (GCV), as performed in R05 and M05, is the source of the problem: "The problem lies in the use of a particular selection criterion (Generalized Cross Validation or 'GCV') to identify an optimal value of the 'ridge parameter', the parameter that controls the degree of smoothing of the covariance information in the data (and thus, the level of preserved variance in the estimated values, and consequently, the amplitude of the reconstruction)." The authors do not elaborate any further, however, making it unclear why such conclusions have been reached. Therefore, it is one of the principal goals of this manuscript to explore ridge parameter selection as the possible source of the differences in pseudoproxy tests of the R05 and M05 methods.

Within the context of NH CFRs, it is important to understand the source of the problems associated with the RegEM-Ridge method for the following two related reasons: 1) if RegEM-Ridge fails only because the method of regularization parameter selection is poor, a different selection procedure could be adopted to improve the application of the method in CFRs of the last

millennium, but 2) if the problem is not associated with RegEM-Ridge specifically and is instead a general feature of the RegEM method, similar pitfalls could be present in RegEM reconstructions that use alternative regularization schemes. In addition to the general motivation of the first reason, RegEM-Ridge was noted by Schneider (2001) for potential advantages over other regularization schemes, and thus may represent a specific approach worth better understanding and improving. The second reason above is also important because M07a used RegEM-TTLS to derive promising results. However, if, for instance, ridge regression was not the source of the problems in the R05 method, it would be an important and outstanding question as to why RegEM-TTLS succeeds while RegEM-Ridge does not. More generally, RegEM-Ridge has been applied to reconstruct other climate variables in different domains, for example, by Zhang et al. (2004) for North American drought reconstructions, making it important to understand unambiguously the strengths and weaknesses of the method in order to interpret reconstructions for which follow-up studies like M07a have not been performed. Ultimately, it is important to characterize the sources of problems in the performance of RegEM-Ridge within the context of paleoclimate CFRs to evaluate how alternative methods may or may not represent opportunities for improved approaches to the problem.

With the above motivations in mind, we discuss several aspects of the RegEM formalism in section 2 of this manuscript and then explore ridge parameter selection in section 3 as the potential source of the differences in the R05 and M05 methods. We also use a direct optimization scheme as an alternative to GCV selection of the ridge parameter. Our results in section 3 demonstrate that ridge parameter selection is not the source of the standardization sensitivity in RegEM-Ridge. In section 4 we compare the two sets of reconstructions to show that the main differences between them arise because of the additional information introduced by the M05 standardization scheme (information that would not be available in real-world reconstructions). We summarize our results and provide conclusions in section 5.

2. RegEM formalism

At the heart of the RegEM method is a linear regression model that relates the missing variables to the available variables:

$$\mathbf{x}_m = \boldsymbol{\mu}_m + (\mathbf{x}_a - \boldsymbol{\mu}_a)\mathbf{B} + \mathbf{e}, \tag{1}$$

where $\mathbf{x}_m$ is the vector of missing variables, $\boldsymbol{\mu}_m$ is the vector of means of the missing variables, $\mathbf{x}_a$ is the vector of available variables, $\boldsymbol{\mu}_a$ is the vector of means of the available variables, $\mathbf{B}$ is a matrix of regression coefficients, and $\mathbf{e}$ is the assumed residual with unknown covariance matrix (Schneider 2001); vectors here are rows of length N_a and N_m for available and missing variables, respectively, and $\mathbf{B}$ has dimensions $N_a \times N_m$. Given the partitioned covariance matrix, the conditional maximum likelihood estimate of the regression coefficients can be written as

$$\hat{\mathbf{B}} = \hat{\boldsymbol{\Sigma}}_{aa}^{-1}\hat{\boldsymbol{\Sigma}}_{am}, \tag{2}$$

where $\hat{\boldsymbol{\Sigma}}_{aa}^{-1}$ is the estimated covariance matrix of the available variables and $\hat{\boldsymbol{\Sigma}}_{am}$ is the estimated cross covariance of the available and missing variables.

The regularization of the regression model in ridge regression is achieved by modifying the inverse matrix $\hat{\boldsymbol{\Sigma}}_{aa}^{-1}$ in Eq. (2) to be

$$(\hat{\boldsymbol{\Sigma}}_{aa} + h^2\hat{\mathbf{D}})^{-1}, \tag{3}$$

where $\hat{\mathbf{D}}$ in this case is chosen to be proportional to the covariance matrix of the observational error in $\mathbf{x}_a$ (following Golub et al. 1999) and h is a positive number called the ridge parameter [see Schneider (2001) for a detailed derivation and discussion of these equations]. When observational errors of different components of $\mathbf{x}_a$ are uncorrelated, $\hat{\mathbf{D}}$ is a diagonal matrix. Additionally, when error variances are proportional to the signal variances, as is the case for the pseudoproxies constructed by M05 and used in the present work, the diagonal of $\hat{\boldsymbol{\Sigma}}_{aa}$ can be used for $\hat{\mathbf{D}}$. Consistent with the M05 application of RegEM, the input data in this study are standardized, and therefore $\hat{\mathbf{D}} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The ridge term $h^2\,\hat{\mathbf{D}}$ in expression (3) inflates the diagonal of $\hat{\boldsymbol{\Sigma}}_{aa}$, consequently dampening the elements of matrix $\mathbf{B}$ in Eq. (2), and thus determines the degree of smoothing that is applied to the estimates of missing values. The selection of h is therefore crucial to the character of the derived reconstruction. As described by Schneider (2001) and applied by R05 and M05, this selection can be performed objectively by minimizing the GCV function.

In Fig. 1 we illustrate the structure of the M05 pseudoproxy data matrix in which there are a total of 1131 records (A.D. 850–1980) and 773 data points (669 instrumental grid cells and 104 pseudoproxies). Instrumental data points are considered available for the 1856–1980 "calibration period" and are reconstructed from A.D. 850 to 1855 (the "reconstruction interval"). The covariance matrix, with dimensions of 773×773, is

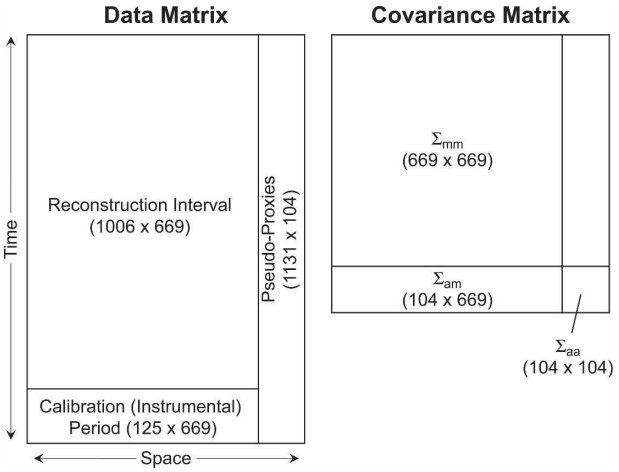

FIG. 1. Data and covariance matrices for the pseudoproxy experiments using CSM climate simulations. The reconstruction and calibration intervals span the time periods of A.D. 850–1855 and A.D. 1856–1980, respectively. There are a total of 104 pseudoproxy series and 669 grid cells in the Northern Hemisphere [based on selection criterion from Mann and Rutherford (2002)].

also schematically represented in Fig. 1. Each ridge regression performed in RegEM-Ridge uses three submatrices from this covariance matrix. These three submatrices—$\hat{\mathbf{\Sigma}}_{aa}$, $\hat{\mathbf{\Sigma}}_{mm}$, and $\hat{\mathbf{\Sigma}}_{am}$—are selected based on the available and missing observations in the data matrix (see Fig. 1) and correspond to the covariance matrices of the available observations, the missing observations, and the cross covariance of the available and missing observations, respectively. For instance, assume that a record of 10 observations is missing the observations 1, 4, and 9. In this case, the ridge regression will be performed using the following submatrices of the covariance matrix (a 10×10 matrix): $\hat{\mathbf{\Sigma}}_{aa}$ is a 7×7 matrix equal to $\|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i, j = 2, 3, 5, 6, 7, 8,$ and 10; $\hat{\mathbf{\Sigma}}_{mm}$ is a 3×3 matrix equal to $\|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i, j = 1, 4,$ and 9; and $\hat{\mathbf{\Sigma}}_{am}$ is a 7×3 matrix equal to $\|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i = 2, 3, 5, 6, 7, 8,$ and 10 and $j = 1, 4,$ and 9.

It is important to note that RegEM is an iterative algorithm and that the covariance matrix is changed upon each iteration. Additionally, the missing observations may be heterogeneously distributed within the dataset, making $\hat{\mathbf{\Sigma}}_{aa}$, $\hat{\mathbf{\Sigma}}_{mm}$, and $\hat{\mathbf{\Sigma}}_{am}$ different for each record (in the example above, the record is missing 1, 4, and 9, but other records could be missing any subset of 1–10). The most general application of RegEM-Ridge therefore requires an individual ridge regression per record and per iteration. This was the case discussed by Schneider (2001), in which RegEM-Ridge was used to impute missing instrumental data that were heterogeneously distributed within the data matrix. If the missing data occur in a regular block, as in the RegEM pseudoproxy reconstructions of M05, however, the submatrices of the covariance matrix are the same for each record with missing data. As indicated in Fig. 1, the missing observations for such records in the pseudoproxy data matrix extend from indices 1 to 669. Thus, the submatrices of the covariance matrix for all records requiring the reconstruction of missing data will be $\hat{\mathbf{\Sigma}}_{mm} = \|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i, j = 1$–669, $\hat{\mathbf{\Sigma}}_{aa} = \|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i, j = 670$–773, and $\hat{\mathbf{\Sigma}}_{am} = \|\hat{\mathbf{\Sigma}}_{ij}\|$, where $i = 670$–773 and $j = 1$–669. Consequently, the value of the GCV-optimized ridge parameter and the regression coefficients are the same for each record during a given iteration, allowing a significant computational savings compared to the general case (Little and Rubin 2002; Schneider 2001). Furthermore, because a single ridge parameter from the final iteration of RegEM-Ridge represents the optimized value for all of the records, comparisons between reconstructions and their final ridge parameters (h final) are straightforward: a single value of h final characterizes each reconstruction.

The final iteration of the RegEM algorithm selects the optimized ridge parameter and subsequently the final regression coefficient matrix. Assuming $\mathbf{P}$ to be a standardized $N_t \times N_p$ proxy matrix, where N_t is the number of time steps and N_p is the number of proxies, the imputed values within the RegEM algorithm can be written as a set of linear operators acting on the proxy data,

$$\mathbf{T}_{\text{recon}} = \mathbf{m} + \mathbf{PBS}, \tag{4}$$

where $\mathbf{m}$ is an $N_t \times N_s$ matrix of estimated field means for the complete data matrix in which each row is equal to one another (N_s is the number of grid locations), $\mathbf{B}$ is the $N_p \times N_s$ matrix of regression coefficients, and $\mathbf{S}$ is a diagonal matrix of the order of N_s containing the estimated standard deviations of the temperature field. During the first iteration of the RegEM algorithm, the $\mathbf{m}$ and $\mathbf{S}$ elements are obtained as the mean and standard deviation of the available values, respectively. In subsequent iterations, the values of $\mathbf{m}$ and $\mathbf{S}$ are reestimated from the results of the preceding iteration; $\mathbf{B}$ is determined from the ridge regression performed during each iteration. Equation (4) collects predictions of missing data from Eq. (1) (rows of the $N_t \times N_s$ matrix $\mathbf{T}_{\text{recon}}$ correspond to $\mathbf{x}_m$ vectors), with the added step of data standardization, because of which $\boldsymbol{\mu}_a = 0$ and the matrix $\mathbf{S}$ scales back to temperature; the rows of matrix $\mathbf{P}$ correspond to the standardized vector $\mathbf{x}_a$. The formulation of the RegEM solution in Eq. (4) has also been checked numerically and confirmed to exactly reproduce the imputed output of the RegEM code.

The utility of Eq. (4) lies in its representation of the

RegEM reconstruction in terms of parsed linear operators. This description will be used in section 4 to characterize differences between M05- and R05-derived reconstructions in terms of these individual operators. It also should be noted, however, that the representation of the RegEM-Ridge reconstruction given in Eq. (4) clearly allows a reconstruction to be performed during the calibration interval, and therefore the calculation of calibration interval statistics. These statistics have not been included for previous RegEM paleoclimate reconstructions (e.g. Zhang et al. 2004; R05; M05; Mann et al. 2007a), although they are common statistics used to evaluate the performance of traditional paleoreconstruction methods. Equation (4) makes it clear that a calibration interval reconstruction is straightforward to calculate using the final **m, S**, and **B** outputs of RegEM in this specific case of paleoreconstructions, and therefore calibration statistics should be reported for such results in the future.

3. Ridge parameter selection

In M05, the pseudoproxies and target field were standardized over the period of A.D. 850–1980, after which the target field was truncated prior to the calibration interval. This expanded standardization interval produced skillful pseudoproxy reconstructions and provided the basis of the conclusions in M05 that supported the use of RegEM-Ridge as a viable CFR method. In contrast, when the proxy and target field are standardized only during the calibration interval (A.D. 1856–1980), as in R05 and as required by any real-world CFR, pseudoproxy experiments with RegEM-Ridge yielded poor results (Smerdon and Kaplan 2007). The origin of this difference has been explained by M07a and Mann et al. (2007b) as resulting from problems in the selection of the ridge parameter using GCV. In this section we investigate the selection of the ridge parameter and its impacts on the derived reconstructions.

In all of our subsequent tests, the experimental design is very similar to that adopted by M05. We use the millennial integration from the NCAR CSM version 1.4 (CSM 1.4) GCM (Ammann et al. 2007). We employ the same 104 pseudoproxies used by M05, made publicly available at the M05 supplemental Web site (online at http://fox.rwu.edu/~rutherfo/supplements/Pseudoproxy05/). These pseudoproxies have been sampled from a $5° \times 5°$ grid interpolation of the originally resolved CSM field (Smerdon et al. 2008, manuscript submitted to *J. Geophys. Res.*) and reflect the real-world distribution of the Mann et al. (1998) proxy network. The pseudoproxies have all been degraded with white noise and set to have signal-to-noise ratios (SNRs), by standard deviation of infinity (no noise), 1.0, 0.5, and 0.25. Note that M05 presented so-called hybrid reconstructions, that is, the pseudoproxy and instrumental fields were split into high- and low-frequency domains divided at the 20-yr period; individual RegEM reconstructions were performed in each domain and the resulting split-domain reconstructions were recombined to derive the full-domain result. M05 note only small differences between the nonhybrid and hybrid reconstructions, a feature confirmed by our own experiments. To maintain simplicity of interpretation, we therefore have not used the hybrid implementation of RegEM and we perform nonhybrid RegEM reconstructions in all cases herein.

In Fig. 2 we plot reconstructions derived using the R05 and M05 standardization choices and the pseudoproxies from M05. Warm biases and variance losses clearly increase with noise level in these nonhybrid R05 reconstructions [similar to the hybrid versions shown in Smerdon and Kaplan (2007)], while the M05 reconstructions compare more closely with the mean NH time series of the model across all noise levels. For each of these reconstructions we plot the final GCV-selected ridge parameter in Fig. 3, a quick inspection of which confirms that the values of h final are almost identical for either standardization choice. These results therefore indicate that the differences in the selected ridge parameters are an unlikely source of the standardization sensitivity of RegEM-Ridge.

While there is little difference between h final in the two different standardization cases, it is possible that the GCV-optimized parameters are in fact far from the truly optimal value. To estimate the latter, we have performed RegEM-Ridge reconstructions for a range of fixed ridge parameters. For each ridge parameter value we have characterized the actual error of the reconstruction by calculating the RMS error and correlation coefficients between the mean NH time series of the reconstructed and the true model values. In Fig. 4, we plot these RMS errors and correlation coefficients as a function of the value of the ridge parameter for reconstructions with fixed h parameters (the value of h determined by GCV is also shown); these calculations were performed using the R05 standardization and pseudoproxies with SNR = 0.5. For both GCV-selected and fixed values of h, RegEM iterations stop when the relative difference in the Frobenius norm between imputed values of two sequential iterations drops below a set stagnation tolerance, set to 0.005 in this work following M05. (While this choice of stopping criterion allows consistent comparisons with the M05 results,

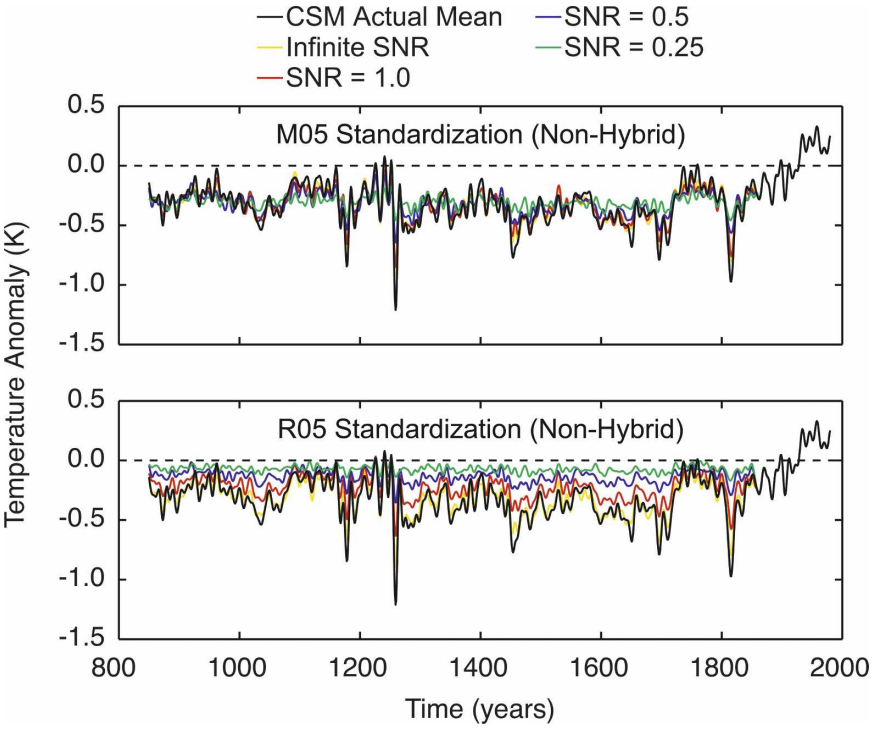

FIG. 2. Nonhybrid RegEM reconstructions of the CSM mean NH climate using (top) the M05 convention, which standardized the instrumental and proxy data over the entire simulation interval (A.D. 850–1980), and (bottom) the R05 standardization convention, which standardized the instrumental and proxy data over the A.D. 1856–1980 calibration interval.

larger values are likely permissible, causing minimal impact on the derived results and improved convergence times.) As demonstrated in Fig. 4, the GCV-selected value of h is within the range of optimal ridge parameters for the correlation coefficient statistic: an optimal value lies between 0.7 and 0.8 and the GCV-selected value was 0.72. The RMS error optimization, however, yields a smaller estimate of $h = 0.31$, less than half the GCV-selected value of 0.72. Nevertheless, this different value of h does not have a large impact on the derived reconstruction; in Fig. 5 we plot the reconstructed NH means derived using the fixed ridge parameter of 0.31 and the GCV-selected value of 0.72 and note only small differences between the two. These results indicate that in addition to the fact that the ridge parameter is not the source of the standardization sensitivity, GCV in particular selects h parameters that yield reconstructions that are quite similar to those determined with directly optimized values of h. These findings therefore give no support to the idea that ridge parameter selection in general, or GCV-selected parameters specifically, are the sources of the standardization sensitivity observed in the performance of RegEM-Ridge.

4. Characterizing the differences between R05 and M05 RegEM-derived reconstructions

Given the results of the preceding section, the source of the standardization sensitivity in RegEM-Ridge has yet to be elucidated. The purpose of this section is to investigate the differences between RegEM-Ridge reconstructions derived with the R05 and M05 standardizations in order to deduce the source of the observed standardization effect. We use the parsed description of the RegEM reconstruction given by Eq. (4) to analyze the role of the individual operators and derive insights into the sources of the differences in the RegEM-Ridge reconstructions.

a. Reconstructed mean and standard deviation fields

In Fig. 6 we plot the difference between the temporal means of the actual model field and RegEM-Ridge reconstructions using R05 and M05 standardizations for pseudoproxies with SNRs of 1.0 and 0.5 (all comparisons are for the reconstruction interval from A.D. 850 to 1855). Table 1 presents the range of differences in the

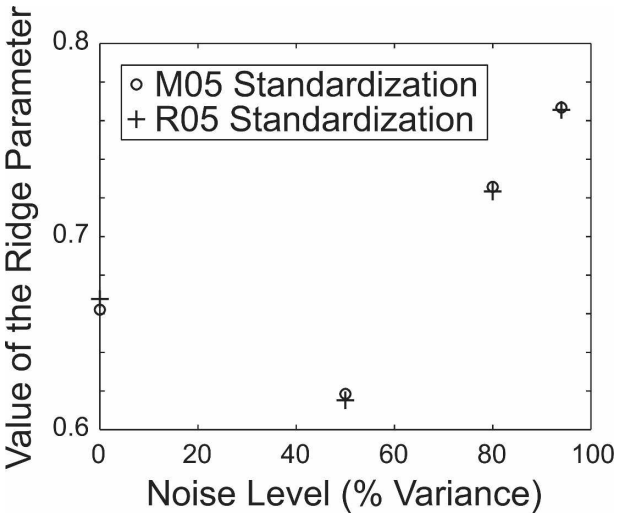

FIG. 3. The value of the ridge parameter determined for the final iteration of the reconstructions shown in Fig. 2. Noise levels are plotted as percent variance in which SNR values of infinity, 1.0, 0.5, and 0.25 correspond to percent variance values of 0%, 50%, 80%, and 94% noise by variance, respectively.

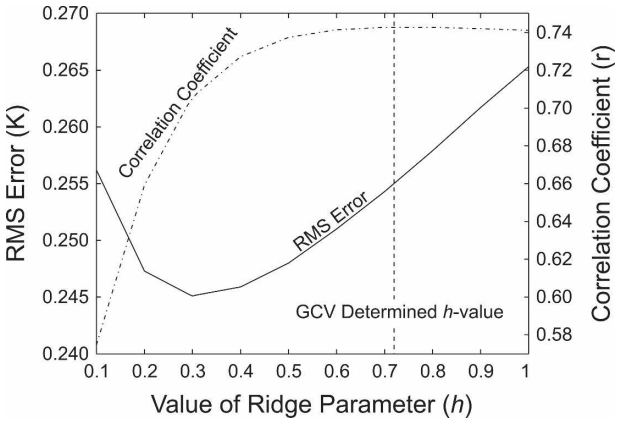

FIG. 4. The RMS error and correlation coefficients associated with fixed ridge parameter RegEM reconstructions. Both the RMS error and correlation coefficients are determined from the mean NH time series of the reconstructions and the known model series during the reconstruction interval (A.D. 850–1855). The optimal values of h were 0.31 and between 0.7 and 0.8 for the respective RMS and correlation optimization schemes, as compared to the value of 0.72 selected by GCV in the final iteration of RegEM. Results were derived for the SNR = 0.5 pseudoproxies.

temporal means and the spatially averaged mean difference in the field for each SNR level and each standardization choice. Table 2 presents the means of the NH time series for the known model field and the set of R05 and M05 reconstructions shown in Fig. 2. All time series have been referenced to the calibration interval mean, and thus the reconstruction interval means are negative. Similar to the comparison of the mean NH time series shown in Fig. 2, the M05 standardization choice yields a mean field that is very similar to the known mean field of the model (Fig. 6, top panel). By contrast, the R05 standardization choice yields progressively larger differences in the mean field for all reconstructions, reaching maximum differences for a SNR of 0.25 that range from −1.22 to 0.21 K, with a mean difference of −0.28 K.

In the no-noise case, RegEM-Ridge reconstructs well the actual model mean using either the M05 or the R05 standardization. For progressively higher noise levels, however, the R05 reconstructions tend away from the actual model mean and toward the mean of the calibration interval, while the M05 reconstructions stay very close to the actual model mean during the reconstruction interval. The explanation of this effect is rooted in the fact that for high noise levels the reconstructions tend toward the mean used in the initial standardization. With each iteration RegEM performs a ridge regression that, at high noise levels, yields reconstructions that are not significantly different from the sample mean of the imputations from the prior iteration. In such high noise cases, the mean corresponding to the initial standardization therefore is carried through iterations without much change to the final RegEM solution.

In the case of the R05 standardization, the initial mean field is determined for the calibration interval and the reconstructions therefore tend toward the relatively warm mean of that period; for high noise levels this tendency amounts to a systematic warm bias of the reconstructed field. For the M05 standardization, however, the reconstructions are initialized with the mean of the full model period, which is not significantly different from the mean of the reconstruction interval (the mean of the area-weighted NH time series referenced to the calibration interval is −0.30 and −0.33 K for the complete dataset and reconstruction-only periods, respectively). Tendency toward this nearly correct mean field results in the artificially unbiased character of the M05 reconstruction. Note that the mean values in Table 2 for the M05 reconstructions essentially converge to the A.D. 850–1980 period mean as noise levels increase. As such, the success of the M05 method, specifically with regard to the lack of warm bias, is entirely due to the additional information (i.e., nearly correct mean) included in the extended standardization choice.

Reconstructed fields (rows of $\mathbf{T}_{\text{recon}}$) discussed in this work represent the estimated means of ensembles of possible states distributed around these means with a reconstruction error covariance $\mathbf{C}$ (Schneider 2006).

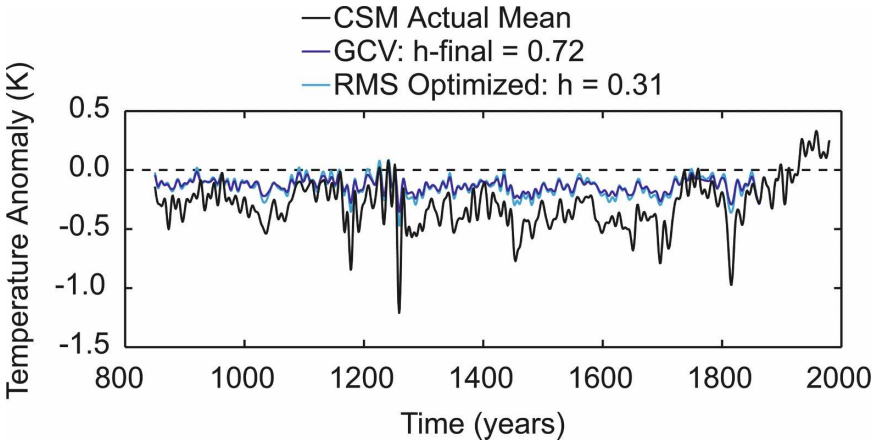

FIG. 5. Comparison of the mean NH time series reconstructed using a GCV-selected and a fixed ridge parameter of 0.31 determined from the RMS optimization. Each reconstruction was performed using the SNR = 0.5 pseudoproxies.

The full covariance of the true state therefore can be estimated as

$$\mathbf{P} = \frac{1}{N_t - 1}\mathbf{SB}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{PBS} + \mathbf{C}, \tag{5}$$

where the first term on the right-hand side is a sample covariance of the reconstructed states, while the residual covariance **C** is due to the deviation of the true states from their reconstructed versions. The covariance matrix is produced by the RegEM algorithm (Schneider 2001) and is the same for all records with missing data in the reconstruction problem considered herein – the larger the reconstructed error variance in the diagonal of matrix **C**, the smaller the sample variance of the reconstructed means. Although the latter cannot serve as a reasonable estimate of the variance of the system, it can provide a convenient measure of the skill of the reconstruction because reduced sample variances, relative to the full system variance [the diagonal of the **P** matrix in Eq. (5)], correspond to increased error variances.

In Fig. 7 we plot ratios between the sample standard deviation fields (computed as the square roots of the sample variance estimates discussed above) of four reconstructions (the same as in Fig. 6) and those of the known model field; summaries of the standard deviations for all reconstructed fields are given in Table 1. Both the R05 and M05 standardizations yield reconstructions with reduced variance in the field. For an SNR of 1.0 the M05- and R05-standardized reconstructions yield ranges of standard deviation ratios from 0.24 to 0.67 and from 0.22 to 0.62, with mean ratios of 0.44 and 0.38, respectively. These variance losses increase with noise; for an SNR of 0.5 the respective ranges of standard deviation ratios for the two reconstructions are 0.18–0.48 and 0.15–0.40, with mean ratios of 0.31

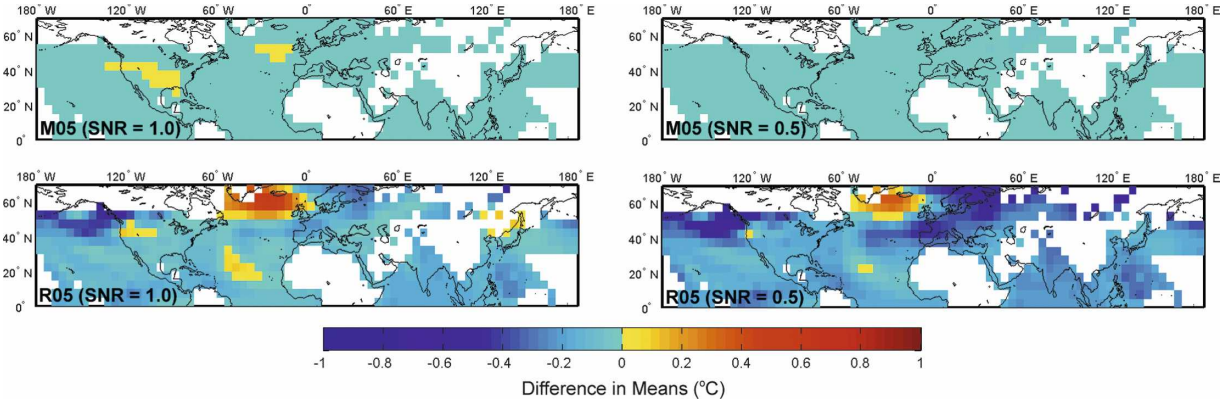

FIG. 6. Mean differences between the reconstructed and known model fields for R05 and M05 standardizations and pseudoproxies with SNRs of 1.0 and 0.5.

TABLE 1. Summary of the differences between the known model field and the R05- and M05-derived reconstructions. All quantities are determined for the reconstruction interval from A.D. 850 to 1855.

Reconstruction	Range of differences in the temporal mean (K)	Mean difference in the temporal means (K)	Range of temporal std dev ratios	Average std dev ratio
M05 (SNR = infinity)	−0.026 to 0.005	−0.006	0.265–0.773	0.511
M05 (SNR = 1.0)	−0.031 to 0.005	−0.008	0.241–0.673	0.443
M05 (SNR = 0.5)	−0.056 to −0.001	−0.015	0.176–0.480	0.313
M05 (SNR = 0.25)	−0.096 to −0.007	−0.028	0.131–0.300	0.204
R05 (SNR = infinity)	−0.353 to 0.840	0.006	0.259–0.760	0.480
R05 (SNR = 1.0)	−0.657 to 0.582	−0.103	0.215–0.623	0.382
R05 (SNR = 0.5)	−1.010 to 0.375	−0.207	0.148–0.398	0.235
R05 (SNR = 0.25)	−1.2161 to 0.2051	−0.275	0.114–0.224	0.157

and 0.24. In general, the patterns of the variance underestimation are similar in the two sets of reconstructions, although the reconstructions derived from the R05 method lose more variance over certain regions, with the oceans being the clearest example. These losses of variance are also seen in the NH mean time series shown in Fig. 2 (see Table 2 for summary statistics) and those discussed by Smerdon and Kaplan (2007) for the hybrid versions of the RegEM-Ridge reconstructions. The expression of these losses is different in the area-weighted NH mean time series and in the field ratios reported above because of the scaled weighting associated with the NH time series. Nevertheless, the NH mean time series displays a similar progressive variance loss with increased noise level: the standard deviation ratios for the M05 and R05 versions of the reconstructions with an SNR of 1.0 are 0.76 and 0.61, respectively, and for an SNR of 0.5 are 0.61 and 0.38, respectively. Generally, the smaller standard deviations of the R05 reconstructions, relative to those of the M05 versions, are expected because of larger variability of the full model dataset (used in M05 standardizations) compared to that of the calibration-only period (used by the R05 standardization).

b. Regression coefficients

The crucial component of the reconstruction in Eq. (4) is the regression coefficient matrix **B**. This matrix is determined from ridge regression in the final iteration. As we demonstrated in section 3, the final values of h in each pair of reconstructions are very similar, and we therefore expect the **B** matrices for both the M05 and R05 standardization choices to closely compare. In Fig. 8 we plot comparisons between all 69 576 elements of the **B** matrices for the R05 and M05 reconstructions shown in Fig. 2. Within each scatterplot we note the correlation coefficient between the two sets of matrix elements as well as the line representing a one-to-one correspondence. The correlations are all highly significant, ranging from 0.89 to 0.98, but differences do exist in the two sets of matrix elements that increase with noise levels.

We have directly tested the impact of the differences in the **B** matrices by computing the correlation coeffi-

TABLE 2. Summary of the comparisons between the known NH mean time series and those of the various reconstruction schemes. All statistics are for individual time series (not for their differences) and are computed for the reconstruction interval (A.D. 850–1855). Means are computed relative to the actual NH mean of the model in the calibration interval (A.D. 1856–1980).

Reconstruction scheme	Correlation	Mean (K)	Std dev (K)
CSM actual	1.0	−0.333	0.208
M05 (SNR = infinity)	0.913	−0.327	0.169
M05 (SNR = 1.0)	0.861	−0.325	0.159
M05 (SNR = 0.5)	0.737	−0.319	0.126
M05 (SNR = 0.25)	0.508	−0.307	0.078
R05 (SNR = infinity)	0.907	−0.323	0.168
R05 (SNR = 1.0)	0.861	−0.225	0.127
R05 (SNR = 0.5)	0.743	−0.133	0.080
R05 (SNR = 0.25)	0.493	−0.072	0.047
M05 w/ R05 coefficient (SNR = infinity)	0.908	−0.328	0.172
M05 w/ R05 coefficient (SNR = 1.0)	0.862	−0.320	0.135
M05 w/ R05 coefficient (SNR = 0.5)	0.744	−0.309	0.087
M05 w/ R05 coefficient (SNR = 0.25)	0.494	−0.301	0.054

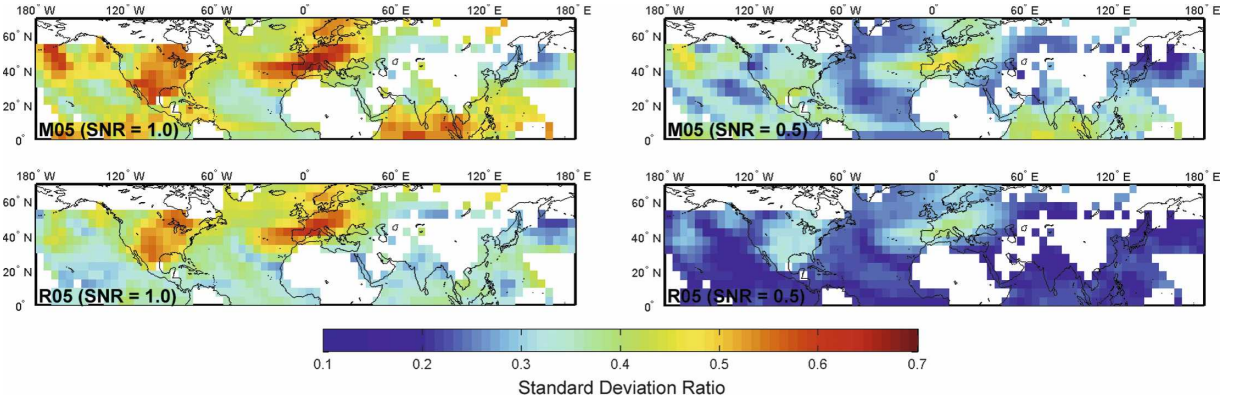

FIG. 7. Same as in Fig. 6, but for standard deviation ratios.

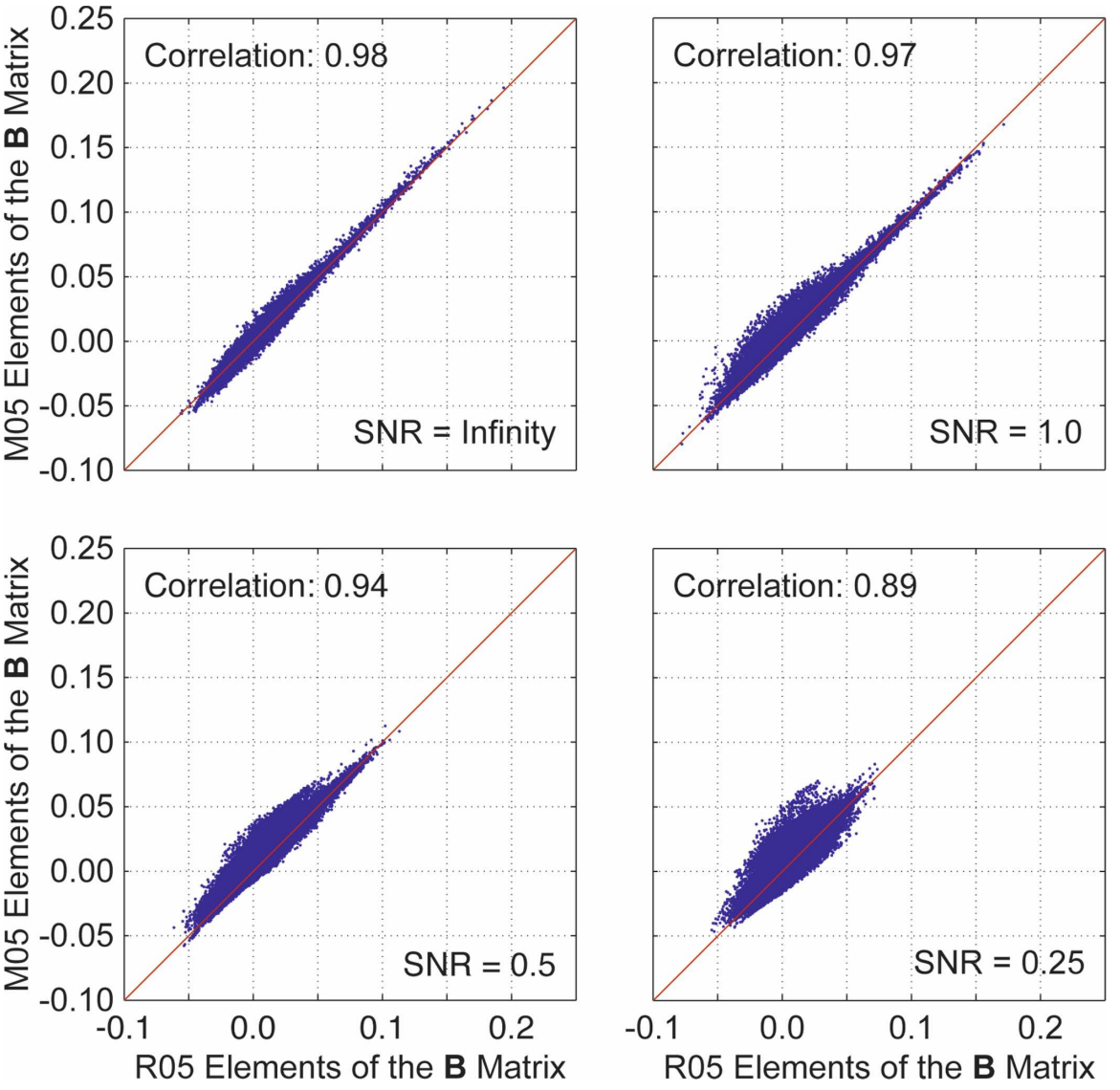

FIG. 8. Comparison of the elements in the **B** matrices for the reconstructions with R05 and M05 standardizations. The total number of elements contained in each **B** matrix is 69 576. Red lines shown in the plots are the one-to-one lines. The correlations between the two sets of **B** elements are shown in the upper-left-hand corner of each plot.

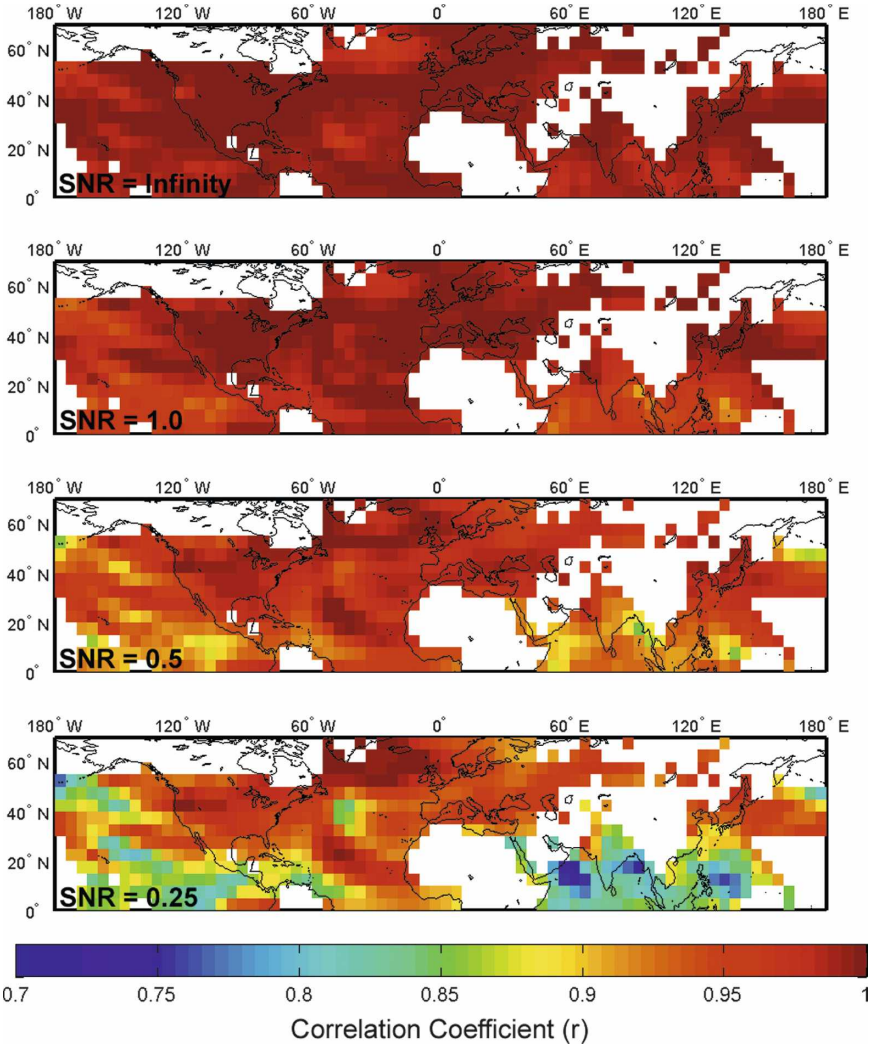

FIG. 9. Correlation fields between the R05- and M05-standardized reconstructions. (from top to bottom) Reconstructions with pseudoproxy SNRs of infinity, 1.0, 0.5, and 0.25, respectively. All correlations are highly significant (note that the scale spans only the range of correlations observed), although they decrease with increasing noise level.

cient fields between the M05 and R05 reconstructions. These fields are shown in Fig. 9 and indicate that the spatial and temporal variability in the two sets of reconstructions are very similar, despite any differences between the regression coefficients. Correlation coefficients never fall below 0.96, 0.91, 0.85, and 0.71 for the SNR reconstructions of infinity, 1.0, 0.5, and 0.25, respectively (note that the scale in Fig. 9 only shows the range of correlation values from 0.7 to 1). As one would expect, the weakest correlations (although still highly significant) occur for the reconstructions that used pseudoproxies with an SNR of 0.25, in which case the elements of **B** were most different.

The overwhelming conclusion from the comparisons in Fig. 9, in conjunction with the previous comparisons of the reconstructed means and standard deviations, is that the two sets of reconstructions are virtually identical, except for the differences in their means and to a lesser extent their variability. As a final exercise, however, we cast the comparison between the two sets of R05 and M05 reconstructions in a different light by computing reconstructions with Eq. (4) in which we use **m** and **S** from the M05 reconstructions, but the **B** matrices from the R05 reconstructions. If the different features in the two sets of reconstructions were a result of differences in the **B** matrices, we would expect these features to remain in the reconstructions derived from this mixed set of operators. In Fig. 10 we plot the NH area-weighted time series for the reconstructions with M05 and R05 standardizations from Fig. 2 and the new

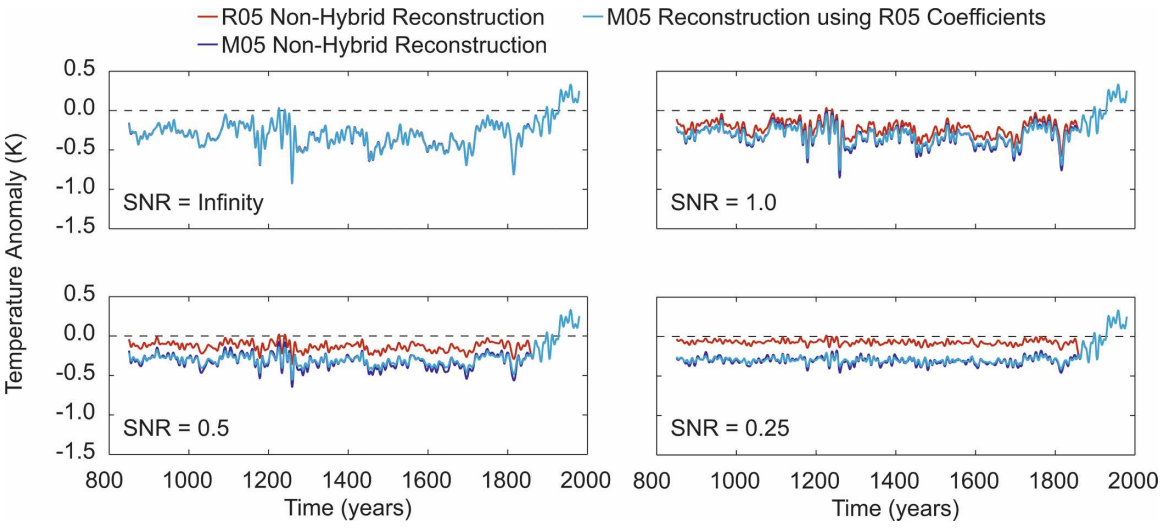

FIG. 10. NH mean time series for reconstructions with R05 and M05 standardizations (repeated from Fig. 2) and the results of the mixed-operator reconstructions in which **m** and **S** from the M05-standardized reconstructions were combined with **B** from the R05-standardized reconstruction [see Eq. (4)].

results from the mixed-operator reconstruction. Table 2 also includes the means and standard deviations of the area-weighted mean time series in the NH during the reconstruction interval, as well as the correlation coefficients between the actual CSM NH mean time series and each of the reconstructed mean time series. In all of the cases these results demonstrate that the M05-standardized reconstructions and the mixed-operator reconstructions are virtually indistinguishable, although the standard deviations of the latter are somewhat reduced toward the R05-standardized reconstructions. These results are a final indication that the **B** matrices, the most nontrivial elements of the reconstructions, are identical for all practical purposes because they produce essentially the same reconstructions. The differences between the two sets of reconstructions are only maintained if the means and standard deviations from the M05 reconstruction are used to provide the additional information from the reconstruction period.

5. Discussion and conclusions

This study has produced a convenient description of the RegEM method for a typical problem of CFR from paleoproxies. We have taken advantage of the fact that only a single regression per iteration is necessary for datasets where the missing data precisely fall within a rectangular block of the data matrix (Schneider 2001). This feature allows for a straightforward comparison between ridge parameters in different RegEM-Ridge reconstructions, and helps to speed up the algorithm considerably for such a dataset. We have also provided a formulation for the final RegEM output in terms of several linear operators acting upon the proxy matrix. This formulation allows a transparent interpretation of the source of the skill in RegEM CFRs, as well as provides a means of performing a reconstruction during the calibration interval. This later advancement allows RegEM reconstructions to be evaluated in terms of traditional in-sample regression skill diagnostics.

By taking advantage of the above-mentioned simplifications, we have shown that ridge parameter selection is not the source of the differences between the R05 and M05 CFR results. No notable differences exist in the selected ridge parameters when the two different methods are employed. We have also shown that GCV specifically is an unlikely source of the problem. While ridge parameter values selected using the GCV procedure might be different from the truly optimal values, the reconstructions derived using the differently selected values are almost identical. These results lend no credence to the statement of either M07a or Mann et al. (2007b) that ridge parameter selection in general, or GCV selection of the ridge parameter specifically, is the source of the standardization sensitivity observed in RegEM-Ridge.

We also have performed experiments that assessed the ultimate source of the differences between R05- and M05-derived reconstructions. These experiments show that the principal source of the standardization sensitivity discussed by Smerdon and Kaplan (2007) is the inclusion of additional information in the M05 standardization choice. We have proven this by demonstrating that the R05 and M05 reconstructions are vir-

tually identical except for their means, and to a lesser extent their standard deviations. Similarly, one can achieve reconstructions that are very similar to the M05 reconstructions using the **B** matrices from the R05 reconstructions. These results collectively rule out explanations of the standardization sensitivity in RegEM-Ridge that hinge on the selection of the regularization parameter, and point directly to the additional information (i.e., the mean and standard deviation fields of the full model period) included in the M05 standardization as the source of the differences between M05- and R05-derived reconstructions. It should be noted further that this information, especially in terms of the mean, happens to be "additional" only because of a special property of the dataset to which RegEM is applied herein: missing climate data occur during a period with an average temperature that is significantly colder than the calibration period. This property clearly violates an assumption that missing values are missing at random, which is a standard assumption of EM (Schneider 2006). If the missing data within the climate field were truly missing at random, there presumably would not be a significant systematic difference between the M05 and R05 standardizations, and hence corresponding reconstructions. The violation of the randomness assumption, however, is currently unavoidable for all practical problems of CFRs during the past millennium and thus its role needs to be evaluated for available reconstruction techniques.

Finally, when the application of RegEM-Ridge is appropriately confined to the calibration interval the method is particularly sensitive to high noise levels in the pseudoproxy data. This sensitivity causes low correlation skill of the reconstruction and thus a strong "tendency toward the mean" of the regression results. It therefore will likely pose some challenges to any regularization scheme applied to this dataset when the SNR in the proxies is high. We thus expect RegEM-TTLS, which according to M07a does not show standardization sensitivity, to have significantly higher noise tolerance and skill than RegEM-Ridge. The precise reasons and details of this skill increase is a matter for future research. It remains a puzzling question, however, as to why the R05 historical reconstruction that was derived using RegEM-Ridge and the calibration-interval standardization (thus expected to be biased warm with dampened variability) and the M07a historical reconstruction that used RegEM-TTLS (thus expected not to suffer significantly from biases) are not notably different. The absence of a demonstrated explanation for the difference between the performance of RegEM-Ridge and RegEM-TTLS, in light of the new results presented herein, therefore places a burden of proof on the reconstruction community to fully resolve the origin of these differences and explain the present contradiction between pseudoproxy tests of RegEM and RegEM-derived historical reconstructions that show little sensitivity to the method of regularization used. Such efforts should be given high priority, and further tests of the RegEM algorithm are highly warranted before great confidence can be placed in RegEM-derived CFRs.

Acknowledgments. We gratefully acknowledge Mike Evans for his helpful comments on this manuscript, as well as Tapio Schneider and one anonymous reviewer for their constructive criticism and useful suggestions. This research was supported in part by the National Science Foundation by Grant ATM04-07909 to AK, by the National Oceanic and Atmospheric Administration by Grant NA07OAR4310060 to JES, and by Grant NA03OAR4320179 (CICAR). DC was additionally supported by a research internship from the Hughes Science Pipeline Project and Department of Environmental Science, both at Barnard College.

REFERENCES

Ammann, C. M., 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, doi:10.1073/pnas.0605064103.

Golub, G. H., P. C. Hansen, and D. P. O'Leary, 1999: Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.,* **21** (1), 185–194.

González-Rouco, F., H. von Storch, and E. Zorita, 2003: Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. *Geophys. Res. Lett.,* **30,** 2116, doi:10.1029/2003GL018264.

——, H. Beltrami, E. Zorita, and H. von Storch, 2006: Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling. *Geophys. Res. Lett.,* **33,** L01703, doi:10.1029/2005GL024693.

Little, R. J. A., and D. B. Rubin, 2002: *Statistical Analysis with Missing Data*. 2nd ed. John Wiley and Sons, 318 pp.

Mann, M. E., and S. Rutherford, 2002: Climate reconstruction using 'Pseudoproxies'. *Geophys. Res. Lett.,* **29,** 1501, doi:10.1029/2001GL014554.

——, 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. Rutherford, E. Wahl, and C. Ammann, 2005: Testing the fidelity of methods used in proxy-based reconstructions of past climate. *J. Climate,* **18,** 4097–4107.

——, ——, ——, and ——, 2007a: Robustness of proxy-based climate field reconstruction methods. *J. Geophys. Res.,* **112,** D12109, doi:10.1029/2006JD008272.

——, ——, ——, and ——, 2007b: Reply. *J. Climate,* **20,** 5671–5674.

Rutherford, S., M. E. Mann, 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 methodology, predictor network, target season, and target domain. *J. Climate,* **18,** 2308–2329.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate,* **14,** 853–887.

——, 2006: Analysis of incomplete data: Readings from statistical literature. *Bull. Amer. Meteor. Soc.,* **87,** 1410–1411.

Smerdon, J. E., and A. Kaplan, 2007: Comments on "Testing the fidelity of methods used in proxy-based reconstructions of past climate": The role of the standardization interval. *J. Climate,* **20,** 5666–5670.

——, J. F. González-Rouco, and E. Zorita, 2008: Comment on "Robustness of proxy-based climate field reconstruction methods," by Mann et al. *J. Geophys. Res.,* **113,** D18106, doi:10.1029/2007JD009542.

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

Zhang, Z., M. E. Mann, and E. R. Cook, 2004: Alternative methods of proxy-based climate field reconstruction: Application to the reconstruction of summer drought over the conterminous United States back to 1700 from drought-sensitive tree ring data. *Holocene,* **14,** 502–516.