# Climate Field Reconstruction under Stationary and Nonstationary Forcing

S. RUTHERFORD AND M. E. MANN

*Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia*

T. L. DELWORTH AND R. J. STOUFFER

*Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*

(Manuscript received 22 March 2001, in final form 9 August 2002)

ABSTRACT

The fidelity of climate reconstructions employing covariance-based calibration techniques is tested with varying levels of sparseness of available data during intervals of relatively constant (stationary) and increasing (nonstationary) forcing. These tests employ a regularized expectation-maximization algorithm using surface temperature data from both the instrumental record and coupled ocean–atmosphere model integrations. The results indicate that if radiative forcing is relatively constant over a data-rich calibration period and increases over a data-sparse reconstruction period, the imputed temperatures in the reconstruction period may be biased and may underestimate the true temperature trend. However, if radiative forcing is stationary over a data-sparse reconstruction period and increases over a data-rich calibration period, the imputed values in the reconstruction period are nearly unbiased. These results indicate that using the data-rich part of the twentieth-century instrumental record (which contains an increasing temperature trend plausibly associated with increasing radiative forcing) for calibration does not significantly bias reconstructions of prior climate.

## 1. Introduction

Numerous recent studies (e.g., Smith et al. 1996; Kaplan et al. 1997) have attempted reconstruction of past large-scale climate patterns [climate field reconstruction (CFR)] using covariance information from data-rich calibration or ''training'' intervals to fill in missing values in data-sparse ''reconstruction'' intervals. An extension of this approach was recently used to reconstruct climate patterns in past centuries from paleoclimate proxy data, using the twentieth century as a calibration interval in which the covariances between instrumental and proxy data are used to calibrate the proxy data (Mann et al. 1998, 1999).

One concern regarding the use of these methods is the possible introduction of a bias in the reconstructions if the basic boundary conditions of the climate system are changing (as might be expected in the response of the climate to anthropogenic increases in radiative forcing). In such cases the calibration interval may contain patterns of variability that did not exist during the reconstruction interval. Conversely, the reconstruction interval could potentially contain patterns that did not exist during the calibration interval. Covariance-based CFR methods assume that values are missing at random [i.e., that the probability that a value is missing is independent of the magnitude of the missing value (Little and Rubin 1987)]. Clearly this is not the case for the instrumental surface air temperature data because the data density increases as mean temperature and radiative forcing also increase (correlation coefficient $r = 0.65$ for correlation between cold-season mean of available data and number of missing values for each year). A similar concern applies to paleoclimate reconstructions based on proxy climate indicators wherein the presumably nonstationary twentieth century is used as a calibration period to reconstruct the climate in previous centuries. In both cases, there is a potential bias in employing a nonstationary overlap (training) interval to calibrate a sparse set of long instrumental or proxy climate series (predictors) against a more widespread but shorter-duration climate field in reconstructing the climate field (predictand) of interest. There are a number of issues worthy of investigation with regard to CFR, including sensitivity to the quality of instrumental and/or proxy data employed, the influences of specific spatial and seasonal sampling biases, and the relative performance of different CFR methods (e.g., Kaplan et al. 1997; Mann et al. 1998; Schneider 2001). Here, we specifically focus on the possible effects of nonstationarity on the performance of covariance-based methods of CFR. For this purpose, we only consider the accuracy of the reconstructed values.

*Corresponding author address:* Scott Rutherford, Dept. of Environmental Sciences, University of Virginia, Clark Hall, Charlottesville, VA 22903.
E-mail: srutherford@virginia.edu

We investigate this issue using the instrumental record and both control and forced integrations of the Geophysical Fluid Dynamics Laboratory (GFDL) R30 coupled ocean–atmosphere general circulation model (GCM; Delworth and Knutson 2000). We perform a set of experiments employing a recently proposed method of CFR (Schneider 2001) that is based on a regularized expectation-maximization (RegEM) algorithm and offers some theoretical advantages over previous methods of CFR (Smith et al. 1996; Kaplan et al. 1997; Mann et al. 1998). We note, however, that our results should broadly apply to other covariance-based methods because these other methods can be interpreted as approximations to the RegEM method (Schneider 2001). In this set of experiments, we focus on the surface temperature field but note the greater generality of the approach to large-scale sea level pressure and hydroclimatic fields provided that the underlying data series are, as is the surface temperature field, reasonably well described by a Gaussian distribution. We include the instrumental temperature field in our experiments to assess to what extent the results of experiments on the model output apply to actual climate data.

Our experimental design for both observational and model-generated data emulates the conventional situation encountered in CFR in that distinct data-rich and data-sparse intervals are created for the purposes of the experiments. The data-rich interval is used to establish covariance estimates between predictor and predictand data to be used for CFR (i.e., calibration), while the data-sparse interval is used to independently test the fidelity of the CFR during the reconstruction interval using additional instrumental data that were withheld from the calibration process (i.e., verification or cross-validation). Although the RegEM method uses all the available data to estimate the covariance matrix, in our experimental design (see section 4 and the appendix), only the data-rich calibration period can contribute to the covariance estimates between predictor and predictand because it is the only time interval in which the predictand data are available. Our experiments are set up so as to test the effects of nonstationarity during either, neither, or both calibration and reconstruction intervals. In section 2, we describe the instrumental and model data used in these experiments. In section 3, we describe the CFR method and diagnostics used for these experiments. In section 4, we describe the analytical approach used to test the effects of nonstationarity in CFR for both instrumental and model temperature fields, and we describe the results of these experiments in section 5. We discuss these results in section 6 and summarize our primary conclusions in section 7.

## 2. Data

### *a. Instrumental data*

The instrumental data used consist of monthly mean combined air temperature over land and sea surface temperature anomalies (relative to a 1961–90 reference period) from January 1856 through December 1998 on a 5° latitude by 5° longitude grid [see Jones et al. (1999) and references therein]. The data coverage is most complete in the twentieth century and decreases considerably in the nineteenth century. In addition, the data are sparse poleward of 70° latitude at all times. We averaged monthly data into annual calendar, boreal cold-season (October–March), and boreal warm-season (April–September) means if at least 50% of values needed to calculate the mean value for each year were available (e.g., a minimum of 6 months for the annual calendar mean case). If less than 50% of the values were available to calculate the average (annual, warm season, or cold season) for a specific year and grid point, the value for that year and grid point was classified as missing. After these averages were calculated we 1) completely removed all grid points poleward of 70° latitude (leaving 2016 grid points) and 2) culled all grid points that were less than 70% complete (see Fig. 1 for a flow chart of the process). Of the original 2016 grid points, the reduced dataset consisted of 1123 grid points in the cold season, 1118 grid points in the warm season, and 1312 for the annual calendar mean.

Computing statistics of reconstructive skill requires that we have a complete dataset for comparison with the dataset completed with the RegEM algorithm. For the GCM output this was not a problem because all grid points for all times are known. For the culled instrumental record, however, there are still missing data points. These were filled using the RegEM method to create a complete dataset of the reduced grid points for all times. Use of filling-in techniques can impose a bias in estimates of the underlying climate field; in the case of the RegEM method, imputed values will be biased toward the climatological mean to some degree (Schneider and Held 2001). However, the effects of these biases and those that may be introduced by culling of grid points (Fig. 1) are relatively unimportant for our purposes, which simply seek a self-consistency of the results of CFR resampling experiments with the original data rather than with the true (unknown, in the case of the observed) surface temperature field. Moreover, the potential impacts of such biases can be assessed directly in our analysis through parallel CFR experiments with model-generated climate field data. Note that we only use the original instrumental data in any statistical assessments of reconstructive skill.

### *b. GFDL coupled model*

We used air temperature from the lowest model level (25 m) in the GFDL coupled ocean–atmosphere model for this study. The model is global in domain and consists of general circulation models of the atmosphere (R30 resolution, corresponding to 3.75° longitude by 2.25° latitude, with 14 vertical levels) and ocean (1.875° longitude by 2.25° latitude, with 18 vertical levels). The

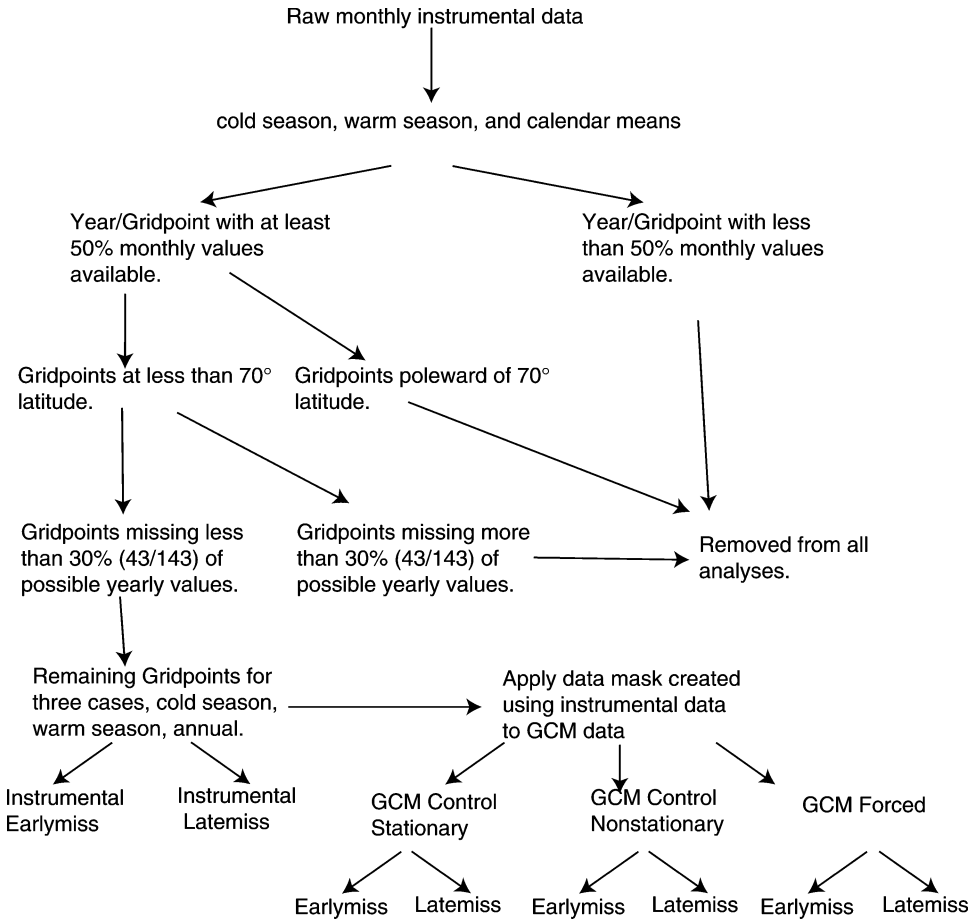


FIG. 1. Flowchart that shows the steps to remove grid points with sparse instrumental data and shows the test suites performed on the culled data.

model atmosphere and ocean communicate through fluxes of heat, water, and momentum at the air–sea interface. Flux adjustments are used to facilitate the simulation of a realistic mean state. A thermodynamic sea ice model is used over oceanic regions.

Output from two experiments is used in this study. The first experiment, referred to as control, was run for 900 yr with a constant atmospheric carbon dioxide ($CO_2$) concentration of 360 ppm. We used two 143-yr intervals from the control run to investigate the ability of the algorithm to reconstruct the temperature field when neither the reconstruction nor the calibration period contained nonstationary forcing, a test that cannot be conducted on the instrumental data. The first 143-yr period is a relatively trend-free interval of natural variability, whereas the second period exhibits a natural secular cooling trend of approximately 0.3°C (Fig. 2a) over model years 75–120. Thus, we were able to test the performance of the CFR technique in the presence of a trend that is internal in origin (and is thus associated simply with limited temporal sampling of an underlying stationary system), rather than a trend that is externally forced (and thus associated with fundamentally nonstationary behavior).

The second experiment, referred to as the forced experiment, starts from an arbitrary initial condition in the control integration. For the period 1866–1990 the model incorporates estimates of the observed time-varying greenhouse gas concentration and sulfate aerosols. From 1990 onward atmospheric greenhouse gases increase at 1% per year. Delworth and Knutson (2000, and references therein) provide further details on the model and experiments. We selected the 143-yr interval corresponding to model years 1948–2090 because this interval presents the RegEM method with a worst-case scenario of highly nonstationary mean response to external forcing (Figs. 2b,c). For both GCM cases, we regridded the data to a 5° by 5° resolution corresponding to the instrumental data.

## 3. Method

Here we present only a brief overview of the method used for CFR and the statistical diagnostics used to as-

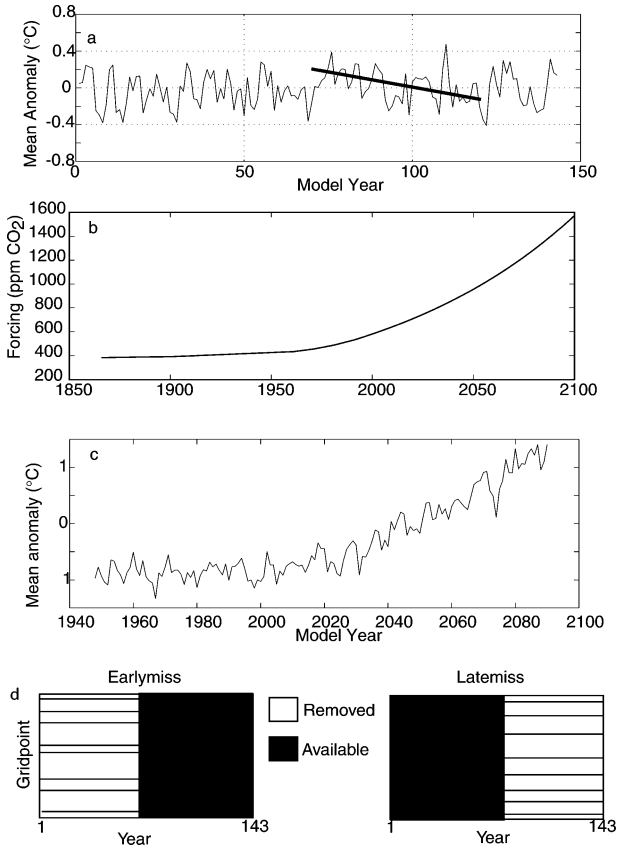


FIG. 2. (a) Test interval of the 900-yr GCM control run (no sulfate aerosol or $CO_2$ forcing) showing the multidecadal cooling trend that occurs in the absence of nonstationary forcing. Model year is arbitrary. (b) Forcing function for the forced GCM experiment in terms of ppm $CO_2$, and (c) response of the GCM (Northern Hemisphere mean) to the forcing function in (b). The forcing includes both sulfate aerosols and atmospheric $CO_2$. Tests are concentrated on modeled years 1948–2090. (d) Earlymiss and Latemiss tests showing the data available during the calibration period (solid) and the verification period (striped).

sess the skill of the reconstruction. Details of the CFR method and the statistical diagnostics are in the appendix. The expectation-maximization (EM) method is a standard method for filling in missing values in a dataset (Little and Rubin 1987). Like most statistical methods, EM only works if the problem is well posed, meaning that the data of interest have fewer variables than records. Typically, CFR problems are ill posed, meaning that the data of interest have more variables than records. In our case, for example, there are over 1000 variables (grid points) and only 143 records (time steps). Thus, some method of regularization must be used to solve the problem. The RegEM algorithm of Schneider (2001) uses ridge regression to regularize the problem and applies an iterative solution method. The method begins by filling in the missing values with a first guess. Typically, the first guess is based, in some way, on the mean of the available data (see the appendix for a discussion of possible choices). Next, the mean and covariance matrices are estimated from the completed dataset. Using the estimated mean and covariance matrices, the filled-in values are adjusted and a revised data matrix is created. The revised data matrix is then used in the next iteration to adjust the estimates of the mean and covariance matrices. The iterations continue until the solution ceases to change appreciably (see the appendix for a discussion of stopping criteria).

We evaluate the skill of the reconstruction using the reduction-of-error statistic (RE), or beta ($\beta_v$), the coefficient of efficiency (CE), and a relative root-mean-square error (Rrmse; see the appendix for details). Each reconstruction can be thought of as containing three components: 1) the mean of the reconstruction period relative to the calibration period, 2) the trend of the reconstruction period, and 3) the interannual/decadal variability. In the case of $\beta_v$, each component contributes to the score. In the case of CE, however, only components 2 and 3 contribute to the score. When the means of the calibration and reconstruction periods are similar, the $\beta_v$ and CE scores are similar (for example in the control GCM case). If the means of the calibration and reconstruction periods differ and the reconstruction correctly estimates the mean of the reconstruction period, $\beta_v$ will be larger than CE. For a perfect reconstruction, both $\beta_v$ and CE will be equal to 1. A random reconstruction with neither trend nor differing means between the calibration and reconstruction period will have $\beta$ and CE equal to $-1$. We calculated $\beta_v$ and CE for each experiment for both the multivariate case (i.e., the statistics for the single time series constructed from the mean of the reconstructed grid points). Since spatial averaging sharply reduces the degrees of freedom, spatial means of the reconstruction typically resolve a significantly larger fraction of data variance than do individual grid points. Thus, one may expect $\beta_v$ and CE for the mean time series to be considerably larger than for the multivariate case. Rrmse is fundamentally different in that it normalizes by the variance over the entire time period and does not differentiate between calibration and verification. Whereas $\beta$ and CE will be influenced by the differences in mean or trend between verification periods, Rrmse will not. See the discussion in the appendix for details.

Although the verification $\beta_v$, CE, and Rrmse statistics provide a measure of skill in the CFR experiments, additional diagnostics of the CFR verification period must be used to examine whether there is any evidence for systematic bias in the CFR results. These diagnostics involve examining the verification-period residuals (the time series of differences between imputed and actual values during the verification period). To conclude that there is no evidence of bias, these residuals should 1) be normally distributed, 2) have zero mean, 3) exhibit no statistically significant trend, and 4) exhibit a white noise spectrum. We thus evaluated the residuals via $\chi$-squared normality tests (at the $\alpha = 0.05$ significance level), $t$ tests for zero mean ($\alpha = 0.05$) and trend ($\alpha =$

TABLE 1. Values of $\beta_v$, CE, and Rrmse resolved variance statistics and the results of tests on the residuals of the mean time series. Note that large $\beta_v$ and CE indicate a better reconstruction, whereas a smaller Rrmse implies a better reconstruction.

| Percent missing grid points | Multivariate | | | Mean | | Residuals | | |
|---|---|---|---|---|---|---|---|---|
| | $\beta_v$ | CE | Rrmse | $\beta_v$ | CE | Normal (pass/fail) | Mean = 0 (pass/fail) | Trend (yes/no) |
| | | | | GCM control, no trend | | | | |
| Early 70% | 0.58 | 0.57 | 0.69 | 0.93 | 0.93 | P | P | N |
| 85% | 0.50 | 0.50 | 0.72 | 0.86 | 0.86 | P | P | N |
| 90% | 0.49 | 0.48 | 0.75 | 0.86 | 0.86 | P | P | N |
| 95% | 0.37 | 0.37 | 0.79 | 0.79 | 0.79 | P | P | N |
| 98% | 0.22 | 0.20 | 0.87 | 0.62 | 0.62 | P | P | N |
| | | | | GCM control, trend | | | | |
| Early 70% | 0.60 | 0.60 | 0.69 | 0.92 | 0.92 | P | P | N |
| 85% | 0.55 | 0.54 | 0.73 | 0.89 | 0.88 | P | P | N |
| 90% | 0.55 | 0.55 | 0.74 | 0.89 | 0.88 | P | P | N |
| 95% | 0.48 | 0.48 | 0.78 | 0.84 | 0.84 | P | P | **Y** |
| 98% | 0.23 | 0.23 | 0.89 | 0.51 | 0.50 | P | **F** | N |
| Late 70% | 0.56 | 0.55 | 0.69 | 0.85 | 0.85 | P | **F** | N |
| 85% | 0.56 | 0.56 | 0.71 | 0.85 | 0.84 | P | **F** | N |
| 95% | 0.47 | 0.47 | 0.77 | 0.81 | 0.81 | **F** | **F** | **Y** |
| | | | | GCM forced | | | | |
| Early 70% | 0.79 | 0.61 | 0.53 | 0.99 | 0.95 | P | **F** | N |
| 85% | 0.77 | 0.56 | 0.56 | 0.99 | 0.90 | P | P | **Y** |
| 90% | 0.77 | 0.56 | 0.57 | 0.99 | 0.93 | P | P | N |
| 95% | 0.74 | 0.51 | 0.60 | 0.99 | 0.79 | P | P | N |
| 98% | 0.60 | 0.23 | 0.70 | 0.97 | 0.35 | P | P | N |
| Late 70% | 0.78 | 0.68 | 0.60 | 0.95 | 0.94 | P | **F** | **Y** |
| 85% | 0.78 | 0.68 | 0.59 | 0.95 | 0.94 | P | P | **Y** |
| 90% | 0.80 | 0.65 | 0.57 | 0.99 | 0.99 | P | **F** | **Y** |
| 95% | 0.77 | 0.60 | 0.60 | 0.99 | 0.98 | P | P | **Y** |
| 98% | 0.69 | 0.46 | 0.69 | 0.96 | 0.96 | P | P | **Y** |
| | | | | Instrumental | | | | |
| Early 70% | 0.44 | 0.25 | 0.91 | 0.98 | 0.91 | P | P | N |
| 85% | 0.44 | 0.13 | 0.94 | 0.95 | 0.72 | P | P | **Y** |
| 90% | 0.45 | 0.17 | 0.94 | 0.97 | 0.79 | P | **F** | N |
| 95% | 0.38 | 0.09 | 0.96 | 0.95 | 0.69 | **F** | P | N |
| Late 70% | 0.56 | 0.38 | 0.77 | 0.96 | 0.87 | P | **F** | **Y** |
| 85% | 0.55 | 0.38 | 0.81 | 0.97 | 0.87 | P | **F** | **Y** |
| 90% | 0.50 | 0.27 | 0.81 | 0.84 | 0.86 | P | **F** | N |
| 95% | 0.45 | 0.22 | 0.84 | 0.93 | 0.72 | **F** | **F** | **Y** |
| Warm 85% | 0.41 | 0.20 | 0.83 | 0.71 | 0.70 | P | **F** | **Y** |

0.05), and the general consistency of the power spectrum with a white-noise null hypothesis. Note that these tests are not all independent of each other. For example, residuals with a significant trend might be expected to exceed the white-noise significance levels at lower frequencies. In addition, if the residuals fail the normality test, one of the assumptions of the *t* test and trend test is violated. Although *t* tests are relatively robust to modest violations of the normality assumption, these tests should be interpreted with some caution if the residuals fail the normality test or if they do not behave as white noise.

## 4. Analytical approach

### *a. Instrumental data*

We make use of a filled-in (spatially and temporally complete) version of the instrumental record, which is then resampled to provide distinct data-sparse and data-rich subsets during either the earlier (relatively stationary) or latter (relatively nonstationary) halves of the full data interval. For simplicity, the data series are resampled in such a way that they provide a spatially sparse but temporally complete data field during the data-sparse subinterval. Two test suites were subsequently performed on each dataset (Figs. 1 and 2d, and Table 1). In the first suite, we removed 100% of the values from the earliest 72 yr (e.g., 1856–1927) at randomly selected grid points, with the latest 71 yr (e.g., 1928–98) being complete. The individual members of the suite had varying percentages of missing data. This suite is referred to as Earlymiss (Fig. 2d). In the second suite, referred to as Latemiss, we removed all values in time from the latter half (e.g., 1928–98) of randomly selected grid points and left data for 1856–1927 complete. The instrumental Earlymiss suite is applicable to filling in

of the sparse early instrumental data. In contrast, the Latemiss suite has no direct real-world application because the instrumental data are relatively complete over 1928–98. The purpose of the Latemiss suite is to assess potential reconstruction biases that may be attributed to calibrating over a period during which the mean appears to be relatively stationary while attempting to reconstruct a period during which the mean appears to be nonstationary.

In our experimental design, the only information available for estimating covariance between predictors and predictands is during the data-rich calibration period. Thus, although RegEM uses all the available data, there is no useful covariance information in the data-sparse verification period to estimate covariances between predictors and predictands.

### b. GFDL coupled model

With the model surface temperature data, unlike the instrumental data, we are free to test multiple possible scenarios, including cases in which nonstationarity is present during neither or either of the calibration and verification subintervals (see Fig. 1 and Table 1 for a breakdown of the tests). As with the instrumental data, both Earlymiss and Latemiss suites were performed on the control and forced GCM experiments (though the distinctions between the Earlymiss and Latemiss experiments obviously become less meaningful in the case of the control simulation experiments). Furthermore, the Latemiss/Earlymiss experiments were performed using two distinct 143-yr intervals of the control simulation, one in which the mean is relatively constant, and the other in which there is a multidecadal trend present in one-half of the full interval. These latter additional experiments allow us to test the fidelity of the CFR when secular trends associated with patterns of (stationary) low-frequency internal climate variability, rather than (nonstationary) secular forcing changes, are present and impart a mean trend during either the calibration or reconstruction interval.

To make the instrumental and GCM tests comparable, we first removed from the GCM domain all grid points poleward of 70° latitude, and those grid points that were less than 70% complete were removed from both the instrumental and GCM datasets and were not considered in any of the tests. With the reduced ''complete'' dataset determined in this manner, we then performed CFR experiments on incomplete versions of the dataset through the entire removal of randomly selected grid points from either the early or late subintervals, employing varying percentages of missing spatial data. In the Earlymiss suite, we removed all values in time from the earliest 72 yr of randomly selected grid points while the latest 71 yr were complete. In the Latemiss suite, we again removed all values in time from the later half of randomly selected grid points and left the earlier half complete.

## 5. Results

We present the results separately for the two GCM runs (control and forced) and the instrumental record. We focus on the cold-season results, but note that warm-season and annual mean results are similar unless otherwise stated. Verification statistics and tests of residuals are summarized in Table 1.

### a. GCM control experiments

The control experiment wherein no natural trend is present during either half of the full (143 yr) interval can serve as a baseline for comparison with all of the other cases. The results of this experiment are summarized in Table 1 and Fig. 3. In this case, we only show results for the Earlymiss test, but, as is expected from symmetry considerations, similar results (and similar $\beta_v$, CE, and Rrmse statistics) are achieved for Latemiss tests. A slight deterioration of the mean time series reconstructions occurs as the number of missing grid points increases from 70% to 90% and the multivariate verification $\beta_v$ drops from 0.58 to 0.49. At 98% missing grid points, the multivariate $\beta_v$ has dropped to 0.22 and the mean series $\beta_v$ is 0.62. Both CE and Rrmse show a similar pattern. Note that a smaller Rrmse and a larger CE and $\beta_v$ indicate a better reconstruction.

The second time interval analyzed from the GCM control experiment contains a multidecadal (roughly 50-yr) cooling trend of approximately 0.3°C, which occurs without external forcing of the model (Fig. 2a). The trend is best defined as occurring between model years 75 and 120 of the interval studied and is preceded by a period of relatively trend-free variability. Because our breakpoint for the Earlymiss/Latemiss tests is at year 72, the trend is completely contained within either the calibration or reconstruction/verification subperiod.

Examples of mean reconstructed time series for the Earlymiss and Latemiss tests are shown in Fig. 4. In the Earlymiss tests, the trend does not appear to affect the fidelity of the reconstructions, because the diagnostics are as good as or better than those for the control experiment without the trend (Table 1). The residuals for the mean time series are consistent with white noise and generally pass the normality, zero-mean, and trend tests (although with 95% and greater missing grid points, failures occur).

The residuals for the mean time series show evidence for a nonzero mean in the Latemiss experiment, and the residuals are inconsistent with white noise at the lowest frequencies (Fig. 4d). On closer inspection, one observes that the reconstruction underestimates the magnitude of several warm peaks in the early part of the reconstruction interval. Thus the reconstruction is unable to reconstruct fully the magnitude of the abrupt shift that occurs between model year 70 and 75 (see Fig. 2a). To verify the impact of the warm peaks between model years 72 and 85 on the evaluation of residuals,

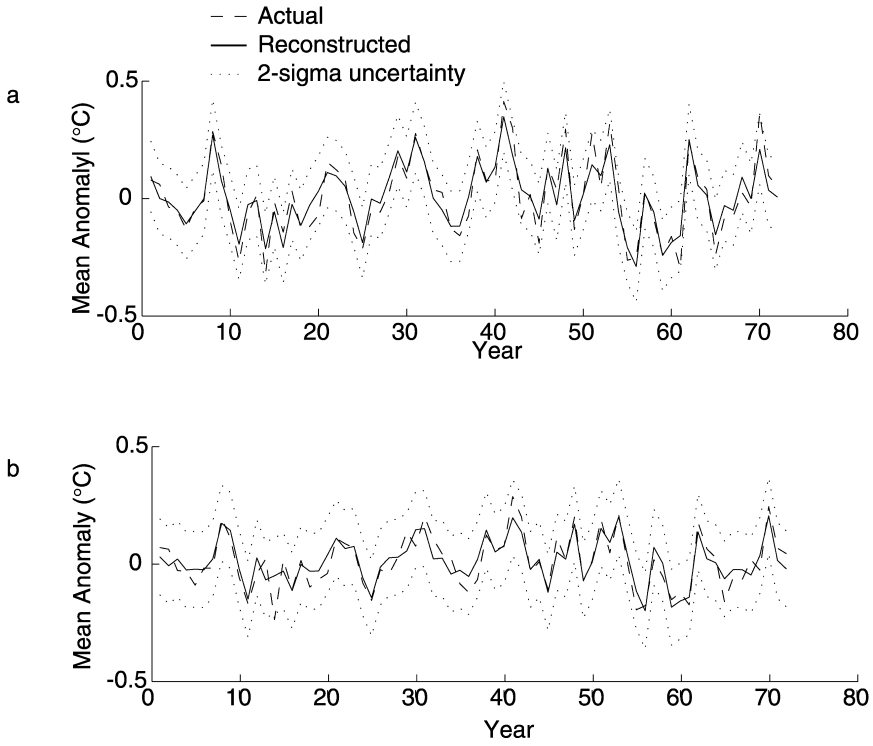


FIG. 3. Earlymiss mean time series reconstruction for the control GCM (cold season) experiment with no multidecadal trend present: mean of filled-in values for (a) 85% and (b) 95% missing grid points. In both cases the resulting reconstruction is nearly unbiased. In this and all subsequent figures showing reconstructed time series, the solid line is the reconstruction, the dashed line is the actual mean time series, and the dotted lines are the $2\sigma$ uncertainties on the reconstruction.

we conducted tests of the residuals from model year 85 through 143. In this case the residuals consistently passed all tests.

Figure 5 shows the gridpoint CE and root-mean-square error (rmse) with 95% missing grid points for the Earlymiss case. Areas with relatively high CE are often associated with high densities of available grid points. This is not always true, however. In many cases, such as Europe (both east and west) and the Indian Ocean, a few grid points are associated with relatively high CE over a large area. In contrast, available grid points in the Southern Ocean and southwest Atlantic tend to produce very localized regions of relatively high CE.

### *b. GCM-forced experiments*

The model experiments with changing radiative forcing allow us to test the algorithm with varying degrees of externally forced trend. We present here the results of a worst-case scenario in which the anthropogenic (combined greenhouse–surface aerosol)-forced climate response is larger than that inferred (e.g., Crowley 2000) for the instrumental record. This time period represents modeled years 1948 to 2090 (Figs. 2b,c).

#### 1) EARLYMISS

In the Earlymiss experiments, the global/hemisphere mean time series show no evidence of bias regardless of the number of missing grid points (Fig. 6). Residuals typically pass tests of normality, zero-mean, trend, and white noise (Table 1 and Figs. 6c,d). Furthermore, the interannual variability of the global, Northern Hemisphere, and Southern Hemisphere mean time series is captured with a high degree of skill. Unlike the control experiments, CE statistics tend to be lower than the $\beta_v$ statistics, because $\beta_v$ is ''rewarded'' for identifying the significant change in mean that occurs between the two subperiods in the case of the forced model experiments. The value of $\beta_v$ in the forced Earlymiss is larger than that for the control Earlymiss for the same reason. Comparing the multivariate CE and Rrmse between forced and control model runs for the Earlymiss tests reveals that, even with a large trend present in the calibration period, the reconstruction skill is similar to or better than that achieved with the control no-trend tests. However, once the percentage of missing grid points becomes extremely high (98% or so) the CE for the mean time series drops precipitously to 0.23 and the Rrmse jumps to 0.89.

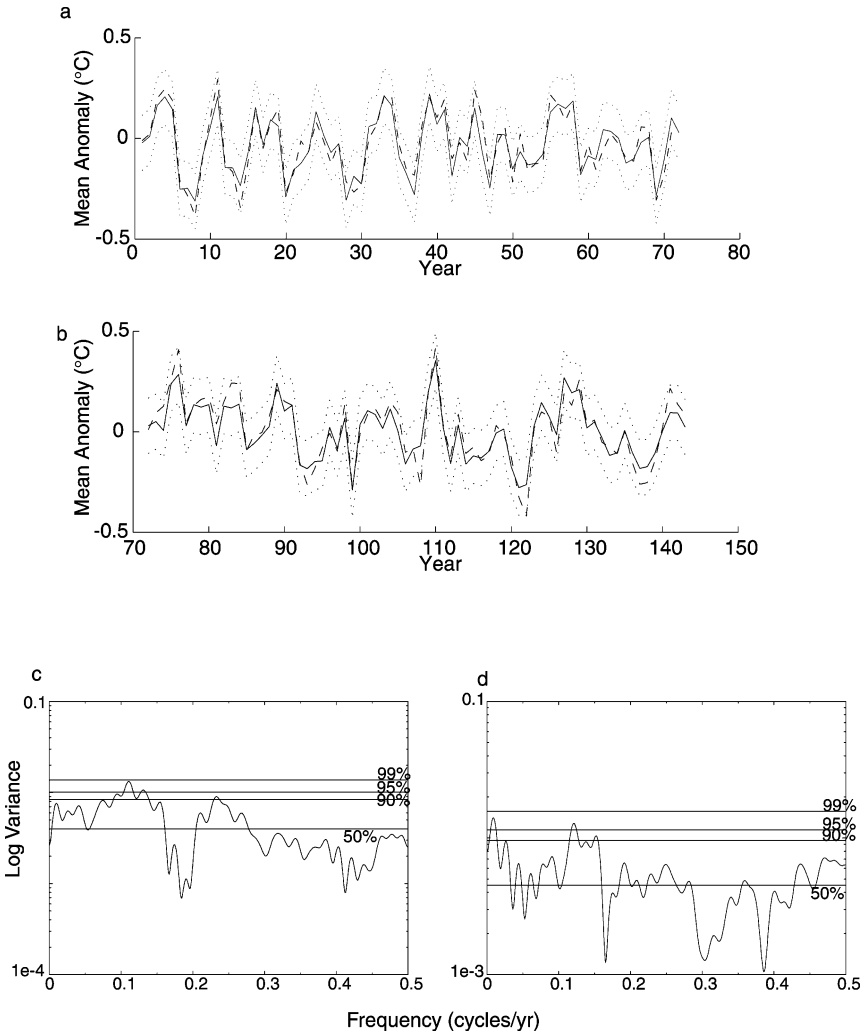


FIG. 4. Earlymiss and Latemiss (cold season) mean time series reconstruction for the GCM control experiment containing a multidecadal secular trend. The trend extends from model year 75 to 120. Filled-in mean time series for 85% missing grid points in (a) the Earlymiss test and (b) the Latemiss test. In the Latemiss test the magnitude of the first three warm peaks is underestimated. Power spectrum of the mean time series residuals for the (c) Earlymiss and (d) Latemiss cases, respectively. In both cases the residuals are consistent with white noise.

### 2) LATEMISS

In contrast to the Earlymiss experiments, the Latemiss tests show evidence of systematic bias in the reconstructed mean time series (Figs. 7a,b). The residuals are inconsistent with white noise (Figs. 7c,d) and exhibit a statistically significant trend (Table 1). Generally, the multivariate CE for the Latemiss tests are greater than those for the Earlymiss tests (Table 1). This result is not unexpected, because CE is normalized by the variance of the reconstruction period. If one verification period contains a trend and the other does not, then the verification period with the trend will tend to have a larger CE, all else being equal. Both $\beta_v$ and Rrmse indicate that the Latemiss reconstructions are no better than (90%, 95%, and 98% missing grid points) or worse than (70% and 85% missing grid points) the Earlymiss reconstructions, with the exception of $\beta_v$ for the case of 98% missing grid points,

The grid point CE and rmse for 95% missing grid points are shown in Fig. 8. These maps illustrate the general importance of gridpoint location. Note the relatively low CE and rmse statistics over parts of North America and the western North Atlantic, regions that had no available grid points in the reconstruction period. In contrast, large areas of Europe and the Indian Ocean exhibit relatively high verification scores even though there are few available grid points in these regions. The maps do not illustrate the bias that exists in the mean time series for the Latemiss case (Table 1, Fig. 7).

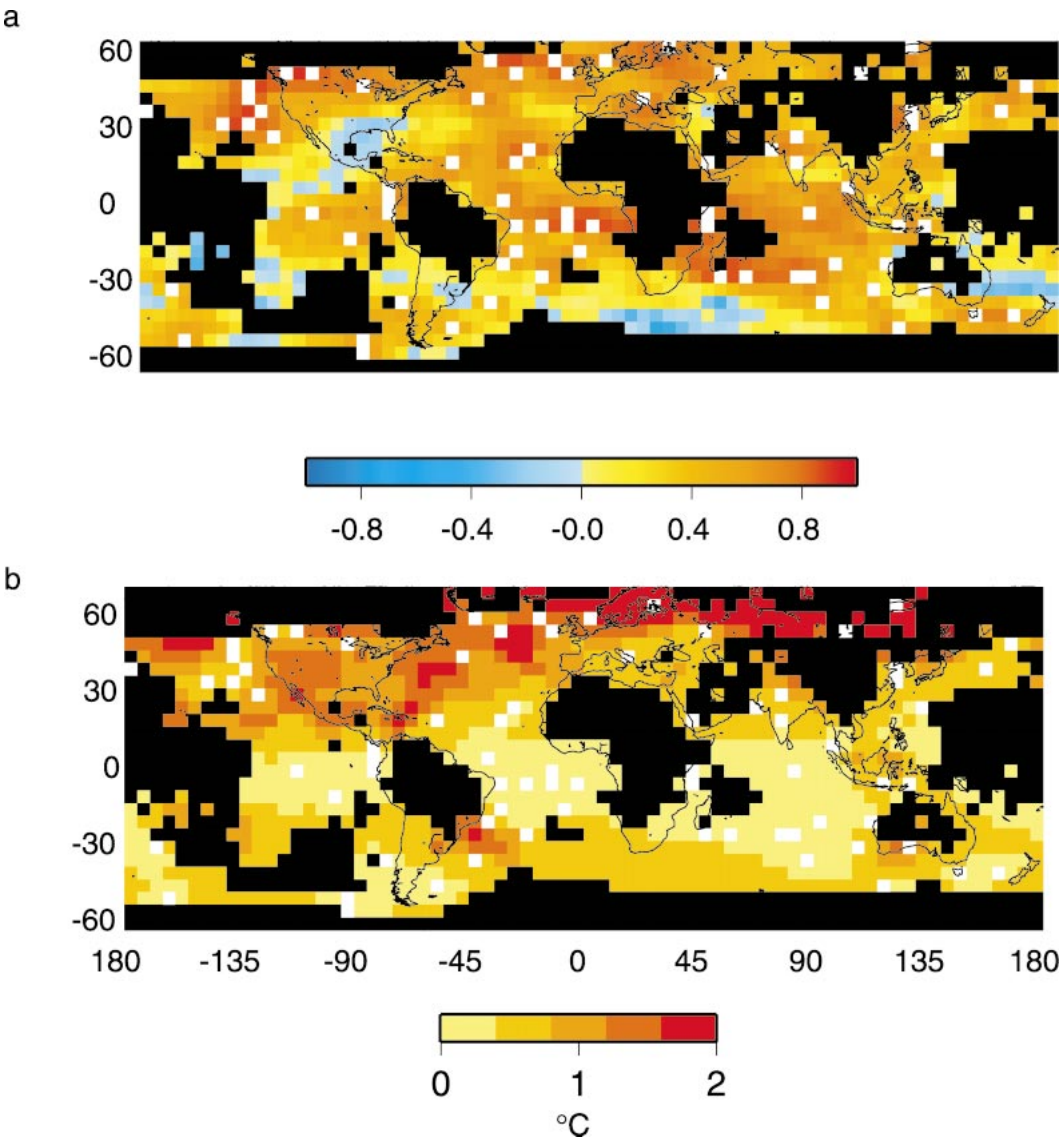


FIG. 5. Maps of (a) verification CE, a measure of relative error, and (b) rmse (°C), a measure of absolute error, for the Earlymiss (cold season) control GCM reconstruction with multidecadal trend and 95% missing grid points. Black boxes indicate grid points that were removed because they are less than 70% complete in the instrumental record. White boxes indicate grid points that were available during the data-sparse reconstruction period. As expected for the case in which the calibration and verification periods have nearly the same mean, $\beta$ (not shown) and CE are similar.

### *c. Instrumental data*

Diagnostics for the instrumental data were only calculated over all randomly selected missing grid points that were initially 95% complete. However, a residual bias may still exist in the earliest (i.e., nineteenth century) instrumental data (e.g., Folland and Parker 1995). We thus used a more conservative abbreviated verification period of 1900–28 for the Earlymiss tests to evaluate the residuals for normality, zero-mean, and trend but nonetheless used the full 1856–1928 verification period for the white-noise tests to provide adequate estimates of the spectra of the residuals. We also note the likelihood of data issues in the nineteenth century that might influence our results as discussed below.

#### 1) EARLYMISS

The Earlymiss experiments using the instrumental data show limited or no evidence of bias (depending on both the number and specific distribution of available grid points). With 85% (Fig. 9a), 90% (not shown), and even 98% (not shown) missing grid points, there is no evidence of a warm or cold bias in the mean reconstruction. However, with the particular sampling of grid

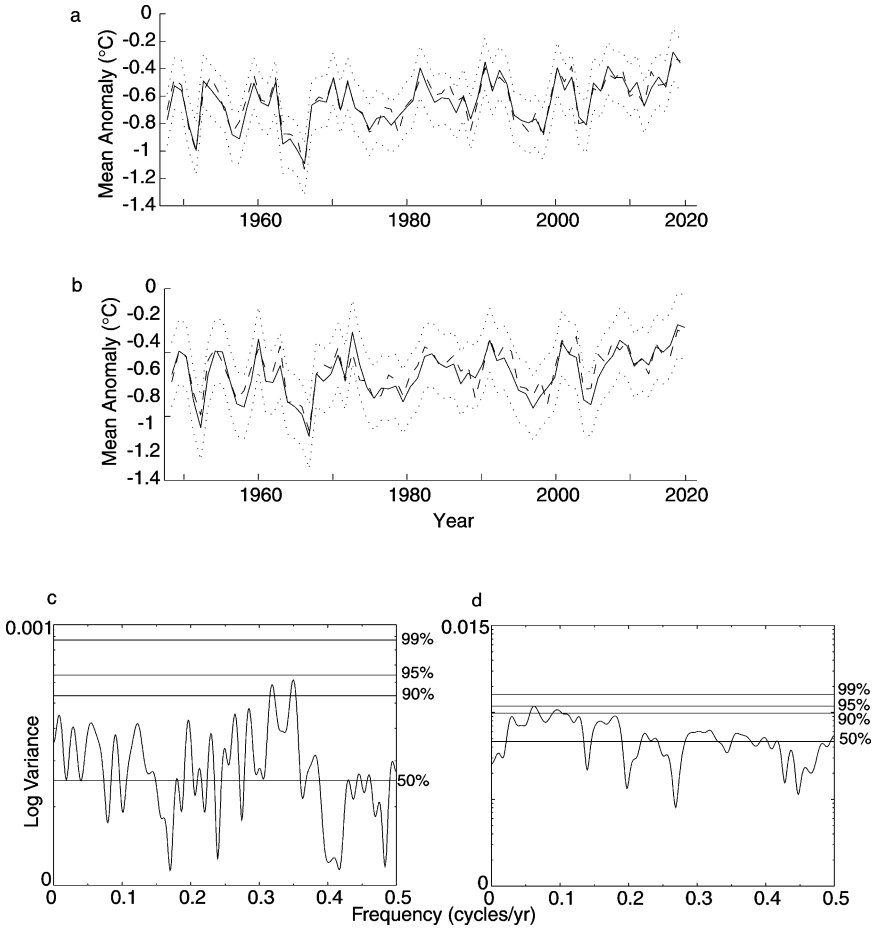


FIG. 6. Earlymiss (cold season) (a), (b) mean time series reconstructions and (c), (d) white-noise residual tests for the forced GCM with (a), (c) 85% and (b), (d) 95% missing grid points. The 50%, 90%, 95%, and 99% significance levels are indicated on the power spectra. The results are nearly ubiased, and the residuals are consistent with white noise.

points available in the 95% missing gridpoint experiment, the mean reconstruction was found to be too warm (primarily between 1865 and 1895) and the spectrum of residuals is inconsistent with white noise at the lowest frequencies at the 0.01 significance level (Figs. 9c,d). The normality, zero-mean, and trend tests on the residuals (1900–28) are mixed but are generally consistent with the forced GCM Earlymiss results, suggestive of little or no bias in the verification residuals.

#### 2) LATEMISS

In the instrumental Latemiss reconstructions we encountered clear evidence of bias in the reconstructions (Figs. 10a,b), consistent again with the forced model (Latemiss) simulation results. The cold bias in the reconstruction (which is evident in the 85% missing data experiment and increases with the number of missing grid points) is clearly visible in the last 20 yr of the reconstruction interval. The residuals in this case typically fail the zero-mean and trend tests (Table 1), and the spectrum is typically inconsistent with the white-noise null hypothesis (Figs. 10c,d). The warm-season reconstruction exhibits a bias that is similar to that of the cold season, though a lesser level of skill is evident (Table 1). Though the residual tests indicate a bias in the Latemiss reconstructions, the multivariate verification statistics indicate that the Latemiss reconstructions are better than the Earlymiss. This contrasts with the forced GCM case, suggesting that nineteenth-century data quality and availability are a factor in our experiments.

## 6. Discussion

### *a. Reconstruction bias*

Significant differences are found among the different experiments in the diagnostics used to check for bias in the verification residuals that require further discussion. In nearly all cases, the verification residuals pass the normal distribution test, but this is not necessarily true

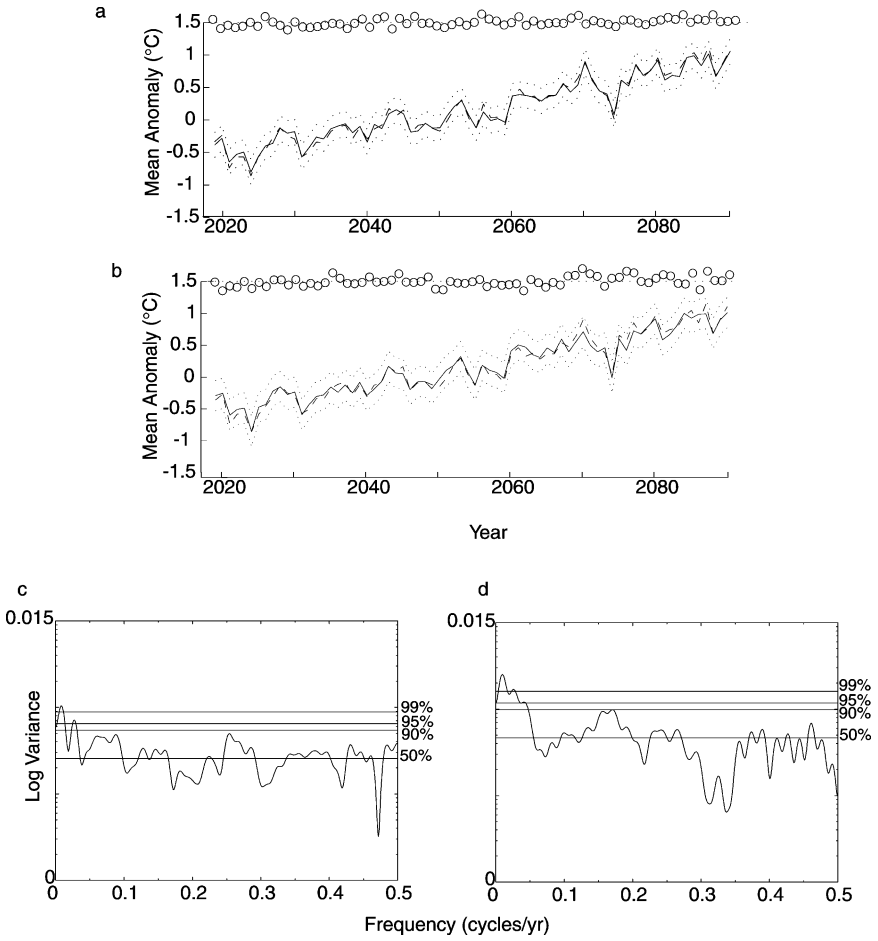


FIG. 7. Latemiss (cold season) (a), (b) mean time series reconstructions and (c), (d) white-noise residual tests for the forced GCM with (a), (c) 85% and (b), (d) 95% missing grid points. The open circles are the residuals, which have been arbitrarily offset for clarity. The zero line for the residuals is shown. Note that the reconstructed mean time series tend to overestimate the true mean temperature in the early part of the reconstruction and to underestimate it in the latter part of the reconstruction, particularly with 95% missing grid points. In both cases, the residuals are inconsistent with white noise.

of the other diagnostics. The control GCM, forced GCM, and instrumental Earlymiss residuals all pass, in general, the battery of tests (zero-mean, trend, and white noise) for unbiased CFR. The Latemiss control GCM experiment shows some moderate evidence of bias with respect to the zero-mean test (which may be associated with an underestimate in the amplitude of a few individual temperature peaks), but the white noise and trend test do not show any evidence of any systematic bias in the estimate of low-frequency variability. In contrast, residuals from the Latemiss instrumental and forced GCM typically fail either two or three of the tests. Our results thus suggest that employing a calibration period that contains a forced (i.e., nonstationary) trend is likely to produce a nearly unbiased CFR, whereas using a stationary calibration period to reconstruct the climate field over a period containing a forced trend may produce a systematic underestimate of the trend. This conclusion is broadly consistent with that of Schneider and Held (2001), who used a GCM simulation with twentieth-century forcing and the evolving data mask of the actual instrumental record to examine the potential bias that might be expected from filling in the instrumental data. They found that a 25-yr forced warming trend in the simulated twentieth-century monthly means was underestimated by 10%–18% by the imputed values, which is similar to the results for our Latemiss instrumental and forced GCM experiments (Figs. 7 and 10). Our results also suggest that reconstructing a natural trend of a magnitude and duration that may be expected to occur based on the control GCM is less problematic than reconstructing a forced trend. Experiments using output from a long GCM control run (e.g., 900+ yr) that exhibit long-term drift are under way to assess the robustness of this result to trend duration.

With regard to the case of using a stationary interval

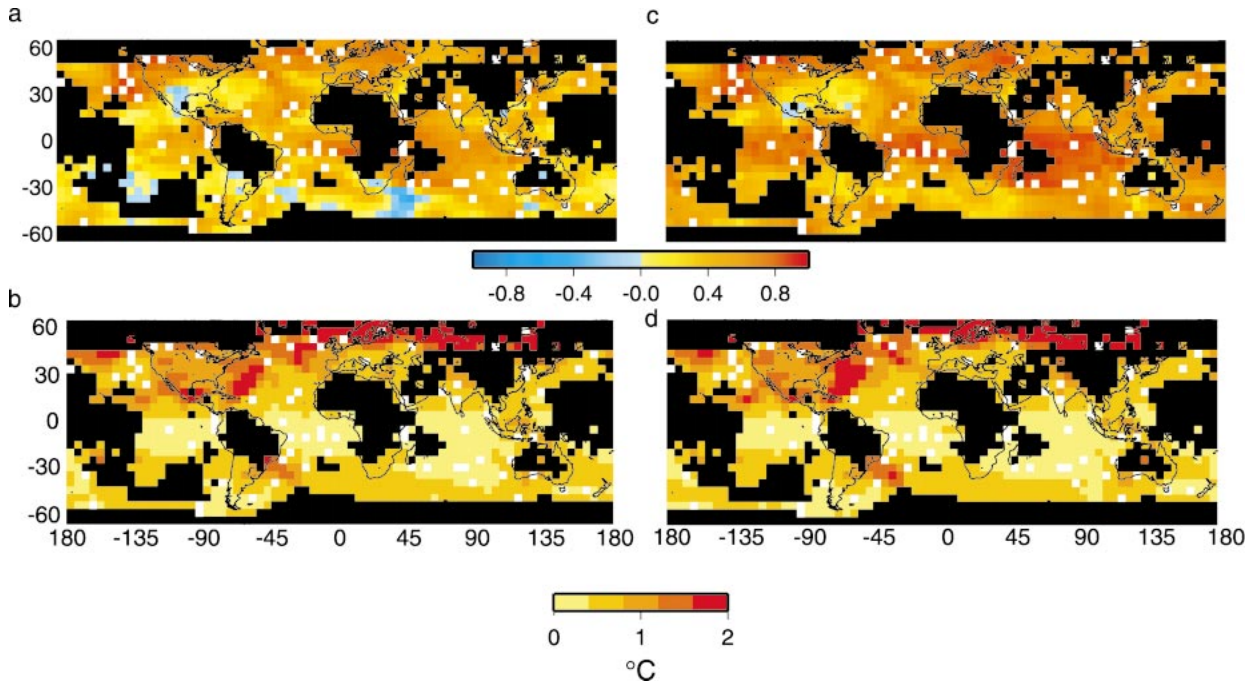


FIG. 8. Gridpoint (a), (c) CE and (b), (d) rmse (°C) for the forced-GCM (a), (b) Earlymiss and (c), (d) Latemiss reconstructions with 95% missing grid points. Here, CE is larger in the Latemiss case because it is normalized by the verification period variance, which is greater in the Latemiss case. These results also display the importance of available gridpoint location. The lowest CE statistics occur where there are large regions without available grid points. Although CE indicates a low relative error at high northern latitudes, the rmse is large.

to reconstruct the behavior over a nonstationary interval, we interpret the bias as resulting from the inability of the covariance-based algorithm to appropriately represent the pattern associated with anthropogenic forcing (i.e., a large-scale pattern of warming, which dominates the late twentieth century) in terms of a linear superposition of the more spatially heterogeneous patterns associated with natural variability. In the Latemiss experiments, the reconstruction attempts to approximate a sparse set of grid points that sample the large-scale warming with the spatially heterogeneous patterns that dominate the earlier calibration period. In so doing, the intrinsic scale of the warming pattern is underestimated, and the large-scale warming trend is underestimated. A longer calibration period that includes part of the warming pattern might alleviate this problem, but the utility of such an approach will depend upon the relative importance of the pattern as compared with the natural patterns as the calibration interval is extended. Thus the length of the calibration period necessary will likely be application specific. On the other hand, the Earlymiss tests show that if the calibration period contains the forced (anthropogenic) pattern, it is not used during the reconstruction period if there is no evidence it is important then. In essence, the amplitude of the forced pattern can readily be damped during the reconstruction interval if this is what the available data merit, but it is far more difficult for a CFR algorithm to identify and use in later reconstructions a forced pattern that is weak or essentially absent during the calibration interval.

### *b. Reconstruction skill*

The reconstruction skill, as expected, decreases systematically as the number of grid points available for CFR decreases. Our results indicate that a significant threshold is reached from roughly 95% to 98% missing grid points in both the instrumental and GCM tests. Somewhere within this range, the diagnosed skill decreases dramatically. This phenomenon can readily be understood in terms of a sampling saturation effect wherein the number of available grid points is potentially considerably larger than the effective spatial degrees of freedom as defined by Schneider [2001, his Eq. (21)]. At 95% missing data in the Earlymiss forced-GCM experiment for example, the RegEM algorithm estimates 23 effective spatial degrees of freedom in the field, with 173 grid points available. In contrast, for 98% missing data, the number of available grid points has decreased by nearly a factor of 6 (to 30), but the effective spatial degrees of freedom are cut in half (12). Thus, each spatial degree of freedom is now only sampled about 2 times on the average, imposing a high probability that a given spatial degree of freedom is left unsampled.

For related reasons, a decrease in the number of avail-

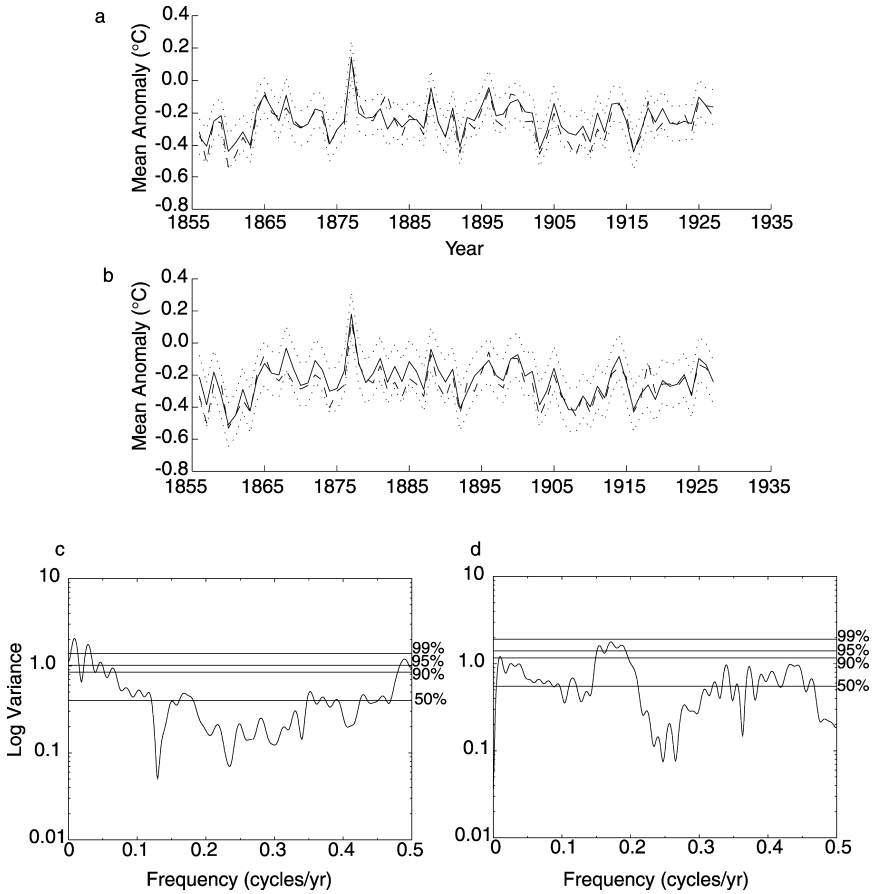


FIG. 9. Earlymiss instrumental mean time series reconstructions with (a), (c) 85% and (b), (d) 95% missing grid points. Note that the residuals fail the white-noise test at low frequencies. The power spectra were calculated over 1856–1928. In contrast, residual tests that are limited to a 1900–28 verification period pass zero-mean and trend tests (Table 1). The discrepancy may be due to variable instrumental data quality in the nineteenth century.

able grid points does not necessarily lead to a decrease in the skill of the reconstruction for any given realization (Table 1). The precise distribution of available grid points can significantly influence the reconstructive skill in CFR, particularly at larger percentages of missing data for which the probability that a given (potentially important) climatic degree of freedom remains unsampled in any given sampling realization. Multiple realizations were performed for several of the experiments, using independent sets of available grid points in the reconstruction interval. These experiments typically yield a range of values of the multivariate $\beta_v$ with a standard deviation of about 0.04. A few well-placed grid points capture a particular spatial pattern more completely than many poorly placed grid points. Certain surface temperature grid points (e.g., those in the eastern tropical Pacific) are located, for example, in particularly important locations for resolving ENSO. In contrast, grid points with large seasonal variance but low signal-to-noise ratios with regard to interannual variability provide relatively little information to constrain the behavior of patterns important for interannual-and-longer-timescale variability. Zwiers and Shen (1997) demonstrated the ability of a small data network to capture large-scale patterns in a GCM and showed furthermore that both the size of the network and the location of the available grid points play a role. Taking this notion one step further, it may be possible to use the instrumental record to identify regions that are particularly important for CFR and to target these regions for the development of proxy climate networks.

### *c. Seasonal and instrumental–GCM comparisons*

Some significant differences exist between the results of the warm-season and cold-season Latemiss instrumental experiments (Table 1). These differences likely arise from the greater level of spatial organization associated with cold-season interannual variability (e.g., El Niño teleconnections and the North Atlantic Oscillation), which may provide more skillful reconstructions of cold-season spatial patterns in the instrumental data

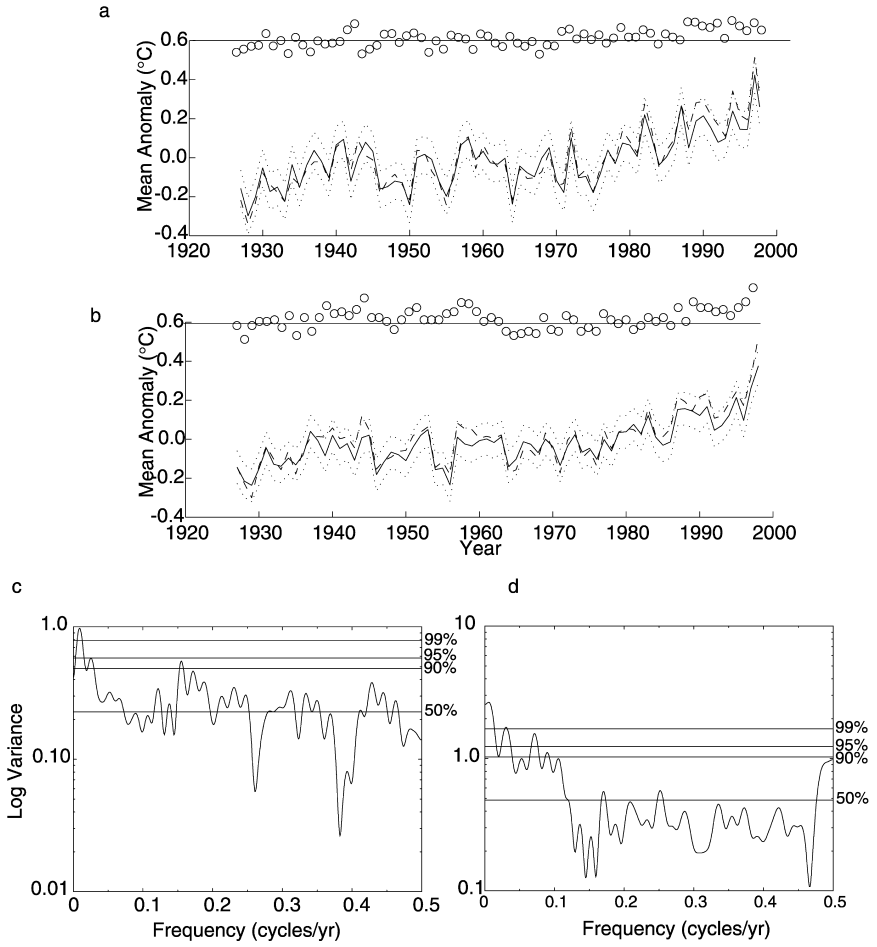


FIG. 10. Reconstructed mean time series and spectra of residuals for the Latemiss tests using instrumental data with (a), (c) 85% and (b), (d) 95% missing grid points. Note the underestimate of the warming trend in the latter part of the reconstruction.

given a particular sparse available sampling of the spatial field. The relative importance of the anthropogenic pattern in both warm and cold seasons may also be an important factor. The greater amplitude of the anthropogenic pattern during the cold season (Schneider and Held 2001) may increase its detectability during the (late nineteenth/early twentieth century) calibration interval of the Latemiss experiment, when this signal, presumably, is beginning to emerge from the background of natural variability.

The control GCM experiment, for which the calibration period contains a natural multidecadal trend, warrants some additional comment. The Latemiss reconstruction in this case underestimates the magnitude of several large peaks in the early part of the reconstruction interval, leading to a very modest apparent bias in the verification residuals. In contrast to the forced Latemiss experiments, for which we attribute the more substantial observed bias to a fundamental weakness of the use of covariance information from a stationary field to reconstruct the behavior of a nonstationary field, this more modest apparent bias is consistent with the known tendency of the RegEM algorithm to impute values that are slightly biased toward the mean.

It is noteworthy that the GCM experiments yield better verification scores than do the instrumental results for the same experiment (e.g., Earlymiss, 85% missing). We attribute this in large part to the existence of nonclimatic sources of small-scale variance in the instrumental record [i.e., measurement error, systematic measurement bias, and sampling error (grid boxes contain varying numbers of observing stations)] that are intrinsically impossible to describe in terms of large-scale patterns, particularly earlier in the record, for which the differences in these scores are greatest. Furthermore, it is likely that poor data quality and sparse coverage of the nineteenth century contribute to the differences observed between the instrumental Earlymiss and Latemiss tests relative to the same tests for the forced GCM. If we had 143 years of actual, complete, high-quality instrumental data, the GCM and instrumental verification scores may have been closer.

## 7. Conclusions

Our climate field reconstruction experiments using the RegEM method lead us to the following conclusions:

1) Under stationary boundary conditions (as deduced from experiments with a control GCM simulation with no changes in radiative forcing) relatively unbiased CFR results are achieved even with extremely low density of available data. This remains true if the calibration interval contains a secular trend that is entirely natural in origin (arising from the finite sampling of low-frequency natural variability). If instead the verification interval contains such a trend, there is some tendency to underestimate the variance of individual fluctuations but there is no evidence of systematic bias in the low-frequency variability.
2) When changing forcing leads to nonstationary behavior during the calibration interval (as deduced from the forced Earlymiss GCM experiments), CFR during a prior interval is nearly unbiased. Similar, though not quite as definitive, conclusions are reached for the actual temperature record, based on Earlymiss experiments using the filled-in instrumental surface temperature data.
3) When changing forcing leads to nonstationary behavior during the reconstruction, rather than calibration interval (as deduced from the forced Latemiss GCM experiments), CFR shows evidence of systematic bias. Similar conclusions are reached for Latemiss experiments using the filled-in instrumental surface temperature data.

These results thus indicate that seasonal and annual reconstructions of sparse early (early or pre-twentieth century) climate fields from relatively complete recent (late twentieth century) instrumental data should produce nearly unbiased reconstructions. Furthermore, it is evident from our experiments that a relatively few well-chosen predictors (e.g., our 95% missing temperature gridpoints experiment, which used 73 grid points to reconstruct a surface temperature field of 1123 grid points) can produce a skillful CFR, yielding, in particular, relatively reliable estimates of global and hemispheric means. Work is in progress to extend these experiments to the estimate of past surface temperature fields with predictors containing significant levels of observational error and/or bias, which will have more direct relevance to the problem of proxy-based climate field reconstruction.

*Acknowledgments.* We thank Tapio Schneider for his help in many aspects of this project, including adapting the RegEM algorithm to our needs and providing helpful comments on the manuscript. We also received helpful comments from three anonymous reviewers that greatly improved the manuscript. Authors SR and MEM acknowledge support from NOAA and NSF through the Earth Systems History Program.

## APPENDIX

### Regularized Expectation Algorithm

The RegEM algorithm is an iterative method for the estimation of mean values and covariance matrices from incomplete data and for using the resulting covariance estimates to fill in missing data with imputed values (Schneider 2001). The RegEM algorithm is based on iterated regression analyses of the variables (e.g., grid points or proxies) with missing values at each record (i.e., time step) on the variables with available values. In each iteration, the regression of variables with missing values on variables with available values is computed for each record from estimates of the mean and of the covariance matrix. Missing values are then filled in with imputed values predicted from the regression model. Once all missing values in the data have been computed, new estimates of the mean and of the covariance matrix are determined and are used in the next iteration to adjust the imputed values. The iterations continue until the estimates of the missing values, mean, and covariance matrix have reached an already specified level of convergence. The regression coefficients in each iteration are estimated by ridge regression, a regularized regression method in which a continuous regularization parameter controls the filtering of the noise in the data (Hansen 1997; Hoerl and Kennard 1970a,b; Schneider 2001, and references therein). In each iteration and for each record, the regularization parameter is chosen adaptively by generalized cross validation (Golub et al. 1979; Hansen 1997). Thus the regularization can adapt to the density and pattern of available values in the data. [See Schneider (2001) for details on the RegEM algorithm, for a juxtaposition with other methods, and for a comparison between ridge regression and principal component regression.]

We used the RegEM algorithm with multiple ridge regressions, in which a single regularization parameter is estimated for each record with missing values. It is also possible to estimate an individual regularization parameter for each missing value, but this approach is computationally more expensive. The choice of initial estimates for the mean and covariance matrix will be discussed below. As stopping criterion for the iterations, we used

$$\left[\sum_{1}^{N}\sum_{1}^{T}(x_{n,t}^{k}-x_{n,t}^{k-1})^2\Big/\sum_{1}^{N}\sum_{1}^{T}(x_{n,t}^{k-1})^2\right]^{1/2}<5\times10^{-3},$$

where $x$ is the data ($T$ records by $N$ variables) with the imputed values of the $k$th iteration filled in for missing values, and with the sums extending over all missing values.

#### a. Initial values

The first iteration begins by assigning initial values for missing data. Typically, each missing data point is

initially assigned the time-mean value of all the available data for that grid point. We used a slightly different approach by using the mean of all available data (over space and time) from the reconstruction period, after removing the climatological mean, as the initial value for every missing value. In tests in which the reconstruction period exhibited a trend (e.g., in the instrumental record when the 1856–1928 data were complete and we were imputing missing grid points between 1929 and 1998) this resulted in a much faster convergence of the algorithm. Figure A1a shows intermediate results over 60 iterations when the mean of all available data for grid point $g$ is given as the initial guess for every missing value at grid point $g$. The slope of the time series is changing slowly as the number of iterations increases. The stopping criterion has not been reached after 60 iterations, and, based on the results at 20, 40, and 60 iterations, several hundred iterations would be necessary (computational constraints prohibit running these cases for several hundred iterations). By using the mean of all available values (over space and time) in the reconstruction period as the initial guess, the stopping criterion is reached more quickly (Fig. A1b). For the kind of situation we are testing (with distinct calibration and reconstruction intervals), the mean of all the available data over the reconstruction period is a better choice for a first guess at the missing values because it uses the available information from the time period of interest (a reasonable alternative approach might use the spatial mean at time $t$ as the initial guess for all missing values at time $t$). When extended to include climate proxy data, one could simply use the mean of the standardized proxy series as the initial guess for the missing values.

### b. Regularization parameter

In CFR methods that use truncated principal components to regularize the ill-posed regression problem, the decision of where to truncate is important. Typically these methods use some rule, such as Preisendorfer's ''rule N'' (Preisendorfer 1988), to choose the number of principal components to retain, and higher-order principal components are discarded. In the RegEM algorithm, higher-order principal components are smoothly filtered out with a continuous regularization parameter controlling the strength of the filtering instead of being abruptly truncated. The regularization parameter is chosen by generalized cross validation (GCV). Frequently, we found that with greater than 85% missing grid points (and rarely with less than 85% missing grid points) the GCV selection of the regularization parameter is too large because of a shallow or nonexistent minimum in the GCV function. This resulted in the need for an upper bound on the regularization parameter. We determined an upper bound on the regularization parameter by specifying the minimum fraction of total variation in the standardized available data that must be retained in the regularization. We tried a number of test runs using the forced GCM output and found that the best results are achieved when the minimum fraction of total variation to be retained is 95%. In many cases, this upper bound is never reached by the GCV selection; thus, the results are relatively insensitive to the value chosen. Even when the bound is reached, whether the upper bound is based on 95% or 90% variance to be retained makes little difference in the solution.

### c. Inflation factor

Another parameter that can be adjusted in the regularized EM algorithm is an inflation factor. The inflation factor adjusts the residual covariance matrix (Schneider 2001, his Eq. 3) for the underestimation of the imputation error due to the regularization (Schneider 2001). The inflation factor is important for error estimation [see the Schneider (2001) appendix] because it accounts for error in the ridge regression coefficients that is then propagated to the errors in the imputed values [see the Schneider (2001) appendix for a discussion of the estimated error in the imputed values]. Relative to the effect on the estimated error, small changes in the inflation factor have a minor effect on the imputed values. Using the GCM control simulation, we conducted multiple tests to determine an inflation factor for different percentages of missing grid points [as suggested by Schneider (2001)]. The inflation factor was adjusted until the estimated (calibration) error and observed (verification) error were nearly the same (precise equating of values would require a large number of trials with slightly different inflation factors, which was not considered to be a worthwhile expenditure of computational resources). These inflation factors were then applied to the forced GCM and instrumental analyses.

### d. Statistical diagnostics

Strictly speaking, the regularized EM algorithm does not contain a calibration period or verification period in the sense defined for standard multivariate regression approaches. However, a calibration period can be defined in the context of our experiments in terms of the distinct data-rich period that is used to estimate the covariance between predictor and predictand that is ultimately used to impute missing values, and a verification period can be defined in terms of the distinct data-sparse period during which imputed values are compared with withheld data to assess the CFR skill. As a measure of reconstructive skill, we evaluate three distinct measures of resolved variance [see Cook et al. (1994) for a review of reduction of error and coefficient of efficiency] during the verification period.

These include the conventional reduction-of-error statistic, defined here by the symbol $\beta_v$ [as in Mann et al. (1998)]:

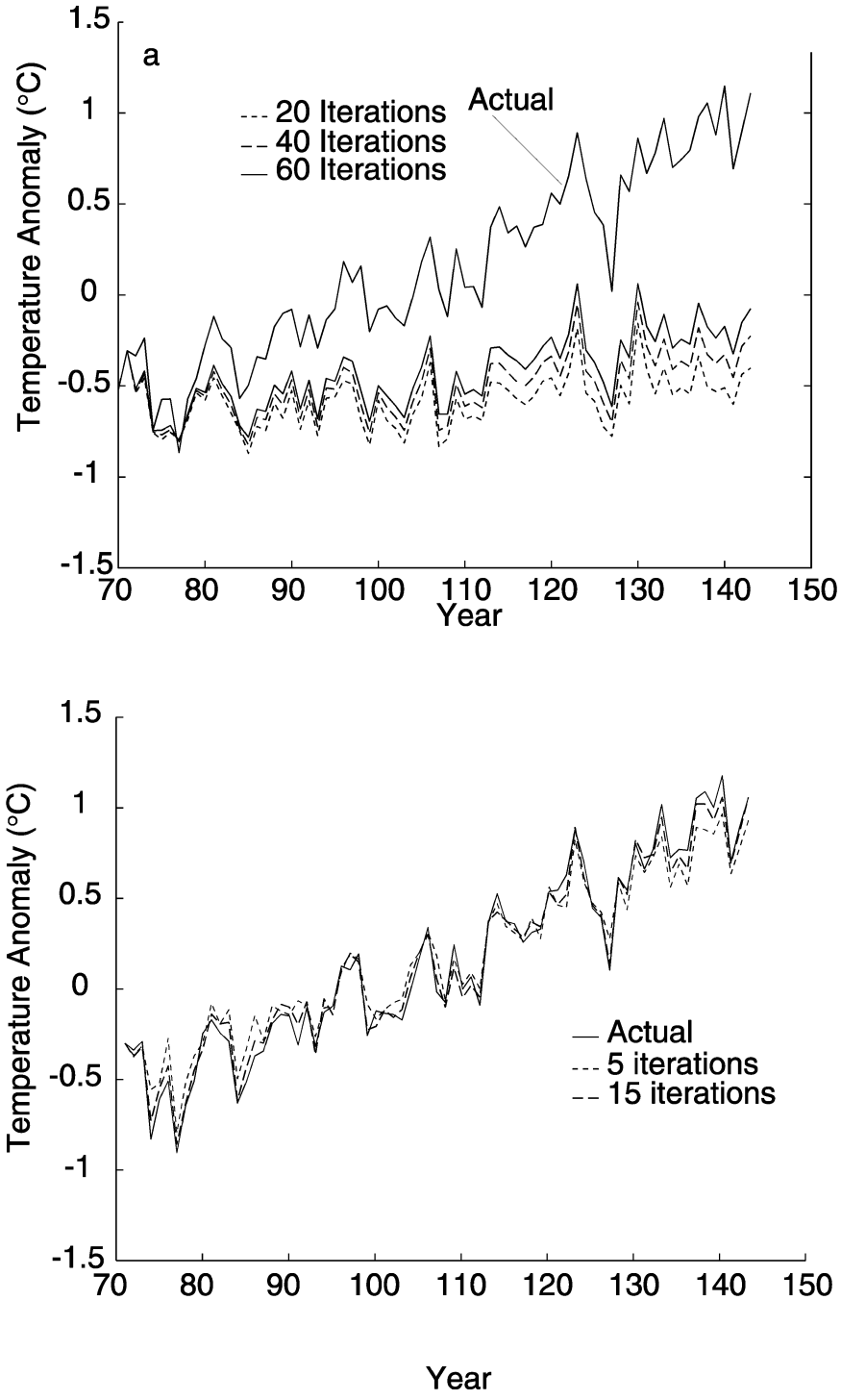


FIG. A1. Mean time series from convergence tests using the forced GCM simulation: (a) the climatological mean for each grid point $g$ was used as the initial value for all missing values at that grid point; (b) the mean of all available values in the reconstruction period was used as the starting point. In (a), the solution is still changing after 60 iterations and the stopping criterion has not been reached. In (b), the stopping criterion has been reached after only 15 iterations.

$$\beta_v = 1 - \left[\sum_1^N \sum_1^T (x_{n,t} - \hat{x}_{n,t})^2\right] \Big/ \left[\sum_1^N \sum_1^T (x_{n,t} - \overline{x}_c)^2\right]$$

and the coefficient of efficiency:

$$\mathrm{CE} = 1 - \left[\sum_1^N \sum_1^T (x_{n,t} - \hat{x}_{n,t})^2\right] \Big/ \left[\sum_1^N \sum_1^T (x_{n,t} - \overline{x}_v)^2\right],$$

where the sums are over the imputed grid points $N$ and time $T$, $\overline{x}_c$ is the mean of the calibration period, $\overline{x}_v$ is the mean of the verification period, $\hat{x}$ is the imputed value, and $x$ is the actual value. Whereas $\beta_v$ assigns skill for estimating differences in mean between verification and calibration period, CE does not.

In addition, we also calculate a relative root-mean-square error for the multivariate case:

$$\mathrm{Rrmse} = \left[\frac{1}{M}\sum_1^N \sum_1^T (x_{n,t} - \hat{x}_{n,t})^2/(\sigma_n)^2\right]^{1/2},$$

where $M$ is the total number of filled-in values and $\sigma_n$ is the standard deviation of the $n$th grid point over the entire test period, not just the calibration or verification period as with $\beta_v$ and CE. Also shown in Figs. 5 and 8 is the standard root-mean-square error (rmse) as a measure of absolute error (CE, $\beta$, and Rrmse are all measures of relative error).

For each CFR experiment, we calculated all diagnostics for the full multivariate field and calculated $\beta_v$ and CE for the single time series (in which case the expressions for $\beta_v$ and CE above contain only a temporal sum) of the global mean (areally weighted mean of all *filled-in* grid points). For the GCM experiments, we calculated verification $\beta_v$, CE, and Rrmse over all filled-in values. For the instrumental data, recall that we initially culled all instrumental grid points that were less than 70% complete and filled the remaining missing values using the RegEM algorithm. This leads to the EM algorithm sometimes filling in randomly selected grid points that were previously partially filled using the same method. Therefore, we calculated statistics only using available values from grid points that were initially 95% complete during the period of 1856–1998 (less than 7 missing out of 143 annual/seasonal values), thereby only comparing imputed values with values that were initially present in the raw temperature data. We furthermore restricted the mean calculation in the instrumental data to include only those grid points that were initially 95% complete, and we used only values that were initially available in the instrumental data to calculate the mean for each year.

The RE, or $\beta$, statistic in traditional multivariate regression can be directly related to the estimates of error in the regression model (e.g., Mann et al. 1998). We could choose to define, in terms of the estimated (calibration) errors in the values imputed by the RegEM approach, an analogous measure of calibration-resolved variance ($\tilde{\beta}_c$):

$$\tilde{\beta}_c = 1 - \left[\sum_1^N \sum_1^T (\tilde{x}_{n,t})^2\right] \Big/ \left[\sum_1^N \sum_1^T (x - \overline{x}_c)^2\right],$$

where $\tilde{x}$ is the estimated error of the imputed value. To avoid confusion with the more conventional measure of calibration-resolved variance, we focus, however, only on the *verification* statistics.

## REFERENCES


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

Crowley, T. J., 2000: Causes of climate change over the past 1000 years. *Science,* **289,** 270–277.

Delworth, T. L., and T. R. Knutson, 2000: Simulation of early 20th century global warming. *Science,* **287,** 2246–2250.

Folland, C. K., and D. E. Parker, 1995: Correction of instrumental biases in historical sea surface temperature data. *Quart. J. Roy. Meteor. Soc.,* **121,** 319–367.

Golub, G. H., M. T. Heath, and G. Wahba, 1979: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21,** 215–223.

Hansen, P. C., 1997: *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM Monographs on Mathematical Modeling and Computation,* No. 4, Society for Industrial and Applied Mathematics, 247 pp.

Hoerl, A. E., and R. W. Kennard, 1970a: Ridge regression: Applications to non-orthogonal problems. *Technometrics,* **12,** 69–82.

——, and ——, 1970b: Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics,* **12,** 55–67.

Jones, P. D., M. New, D. E. Parker, S. Martin, and J. G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Rev. Geophys.,* **37,** 173–199.

Kaplan, A., Y. Kushni, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J. Geophys. Res.,* **102** (C13), 27 835–27 860.

Little, R. J. A., and D. B. Rubin, 1987: *Statistical Analysis with Missing Data.* Series in Probability and Mathematical Statistics, John Wiley and Sons, 278 pp.

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.

——, ——, and ——, 1999: Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.,* **26,** 759–762.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography.* Elsevier Science, 425 pp.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate,* **14,** 853–871.

——, and I. M. Held, 2001: Discriminants of twentieth-century changes in earth surface temperatures. *J. Climate,* **14,** 249–254.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate,* **9,** 1403–1420.

Zwiers, F. W., and S. S. Shen, 1997: Errors in estimating spherical harmonic coefficients from partially sampled GCM output. *Climate Dyn.,* **13,** 703–716.