Clim Dyn
DOI 10.1007/s00382-015-2684-0

Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era

Jason E. Smerdon[1] · **Sloan Coats**[1,2] · **Toby R. Ault**[3]

Received: 20 February 2015 / Accepted: 22 May 2015

Abstract The spatial skill of four climate field reconstruction (CFR) methods is investigated using pseudoproxy experiments (PPEs) based on five last millennium and historical simulations from the Coupled and Paleo Model Intercomparison Projects Phases 5 and 3 (CMIP5/PMIP3) data archives. These simulations are used for the first time in a PPE context, the frameworks of which are constructed to test a recently assembled multiproxy network and multiple CFR techniques. The experiments confirm earlier findings demonstrating consistent methodological performance across the employed methods and spatially dependent reconstruction errors in all of the derived CFRs. Spectral biases in the reconstructed fields demonstrate that CFR methods can alone alter the ratio of spectral power at all locations in the field, independent of whether there are any spectral biases inherent in the underlying pseudoproxy series. The patterns of spectral biases are model dependent and indicate the potential for regions in the derived CFRs to be biased by changes in either low or high-frequency spectral power. CFR methods are also shown to alter the pattern of mean differences in the tropical Pacific during the Medieval Climate Anomaly and the Little Ice Age, with some model experiments indicating that CFR methodologies enhance the statistical likelihood of achieving larger mean differences between independent 300-year periods in the region. All of the characteristics of CFR performance are model dependent, indicating that CFR methods must be evaluated across multiple models and that conclusions from PPEs should be carefully connected to the spatial statistics of real-world climatic fields.

Keywords Climate field reconstruction · Pseudoproxy · Last millennium · Climate model · PMIP3 · CMIP5

1 Introduction

The increasing availability of forced-transient simulations of the last millennium (e.g., Fernández-Donado et al. 2013; Masson-Delmotte et al. 2013; Otto-Bliesner et al. 2015) from fully-coupled General Circulation Models (GCMs) has greatly improved the potential for investigating multi-decadal-to-centennial-scale climate dynamics (e.g. Coats et al. 2013a, b, 2015a, b) and for comparing paleoclimate reconstructions with model output (e.g. Phipps et al. 2013; Schmidt et al. 2014; Hind et al. 2012, 2013; Coats et al. 2013a, b, 2015a, b; Anchukaitis et al. 2010; Seager et al. 2008; Fernández-Donado et al. 2013; Ault et al. 2013a, b; Lehner et al. 2012; Goosse et al. 2010, 2012; Sundberg et al. 2012; Berdahl and Robock 2013; Bothe et al. 2015). Among the currently available collection of simulations, the Coupled and Paleo Model Intercomparison Projects Phases 5 and 3 (CMIP5/PMIP3) have, for the first time, produced multiple last-millennium (LM), historical, and future simulations using the same model configurations and resolutions (Taylor et al. 2012). This development makes possible a wide range of model analyses and comparisons between paleoclimatic data and LM simulations, with direct quantitative applicability to historical simulations and future projections.

One important application of the LM simulations in the CMIP5/PMIP3 database is their use in pseudoproxy

✉ Jason E. Smerdon
jsmerdon@ldeo.columbia.edu

[1] Lamont-Doherty Earth Observatory of Columbia University, 61 Route 9 W, P.O. Box 1000, Palisades, NY 10964, USA

[2] NASA Goddard Institute for Space Studies, New York, NY, USA

[3] Cornell University, Ithaca, NY, USA

Published online: 06 June 2015

experiments (PPEs). Over the last decade, PPEs have emerged as an important tool for evaluating the performance of methods used to reconstruct the climate of the Common Era and the dependence of reconstruction fidelity on various characteristics of proxy networks, calibration data and proxy noise (for a review see Smerdon (2012)). Despite a considerable amount of research employing PPE frameworks, these efforts have overwhelmingly relied on only two fully-coupled forced transient simulations of the last millennium that have been available for a decade or more. The first simulation, called ERIK1, was completed with the GKSS ECHO-G model (González-Rouco et al. 2003) and initially used in PPEs by von Storch et al. (2004). Similarly, the NCAR CCSM1.4 model was used by Ammann et al. (2007) to perform a last-millennium simulation that was originally employed by Mann et al. (2005) in a PPE framework. Although some control simulations and shorter forced-transient simulations also have been used for PPEs, the ECHO-G and CCSM1.4 simulations, in addition to an updated ERIK2 LM simulation using the same ECHO-G model (González-Rouco et al. 2006), have formed the basis of almost all PPEs to date (Smerdon 2012).

Herein, for the first time, five of the newly available CMIP5/PMIP3 LM simulations are used to derive a comprehensive set of PPEs, testing four widely applied climate field reconstruction (CFR) methods. Two categories of reconstruction techniques are often discussed in the literature, the first of which involves index methods that target indices such as Northern Hemisphere (NH; e.g. Moberg et al. 2005; Hegerl et al. 2007; Mann et al. 2008; Christiansen and Ljungqvist 2012) or global mean temperatures (Mann et al. 2008). CFR methods comprise the second category, which target spatial maps of reconstructed temperatures typically expressed on a regular latitude-longitude grid (e.g. Mann et al. 2009a; hereinafter M09). While both approaches have their merit, the promise of CFR methods is in their ability to estimate spatial patterns of temperature variability and change, thus providing more detailed dynamical insights. Our focus is therefore specifically on the field performance of derived CFRs, an emphasis that has only recently received attention in the pseudoproxy literature (Smerdon et al. 2008a, 2010a, 2011; Li and Smerdon 2012; Annan and Hargreaves 2012; Dannenberg and Wise 2013; Steiger et al. 2014; Wang et al. 2014; Evans et al. 2014). Note also that our discussion herein is specific to global and hemispheric temperature reconstructions and PPEs, in contrast to the many regional CFRs and PPEs that have targeted multiple climatic variables (e.g. Evans et al. 2002; Cook and Krusic 2004; Luterbacher et al. 2002, 2004; Pauling et al. 2003; Cook et al. 2010; Neukom et al. 2010; Riedwyl et al. 2009; Werner et al. 2013; PAGES 2k Consortium 2013; Anchukaitis et al. 2013; Tingley and Huybers 2010).

Given the importance of the spatial information estimated in CFRs, assessments of the field skill associated with applied CFR methods are critical for evaluating the robustness of the derived spatiotemporal information. There nevertheless are relatively few assessments that have used PPEs to test the spatial skill of CFR methods. While some studies have reported summaries of field statistics or provided spatial plots of limited assessment metrics (e.g. Mann et al. 2005, 2007; Rutherford et al. 2003), most evaluations of CFR methods using PPEs have focused primarily on their ability to derive skillful NH or global mean indices. Such evaluations are insufficient for assessing CFR spatial performance (Smerdon et al. 2011; Smerdon 2012). In the last several years, however, a growing number of studies have used PPEs to explicitly evaluate the spatial performance of hemispheric or global CFRs (Smerdon et al. 2008a, 2010a, 2011; Li and Smerdon 2012; Annan and Hargreaves 2012; Dannenberg and Wise 2013; Steiger et al. 2014; Wang et al. 2014; Evans et al. 2014; Guillot et al. 2015). Among the various conclusions of these studies, important spatial errors have been demonstrated in CFRs derived from a range of state-of-the-art methods and these errors are expressed relatively consistently across all techniques. The magnitude of error has been shown to be dependent on the character and level of noise in pseudoproxy networks, the pseudoproxy distributions and availability back in time, whether the pseudoproxies sample univariate or multivariate climate characteristics, and method-specific parameter choices such as the degree of regularization in multivariate regression formulations. This collection of studies therefore has pointed to the need to vet the spatial performance of currently applied CFR techniques using more realistic PPE designs, while more directly connecting outcomes of PPEs to interpretations of real-world reconstruction products.

Despite the growing body of work that focuses on the spatial performance of CFR methods, an underappreciated influence on these assessments is the spatiotemporal character of the modeled climate field that forms the basis of the PPE. An early-articulated (e.g. Mann et al. 2005) and fundamental assumption of PPE designs based on GCM simulations is that the models reasonably represent the spatiotemporal characteristics of the actual climate. With regard to the most common simulations used in PPE frameworks, namely forced-transient simulations that adopt plausible historical forcing scenarios, it is not necessary that the simulated trajectory or internal variability follow the exact historical trajectory of climate over the last millennium. Nevertheless, the relevance of PPE findings to interpretations of real-world reconstructions is dependent on the ability of models to reasonably capture the spatiotemporal characteristics of the targeted climate field. These characteristics include, *inter alia*, teleconnection patterns and

their temporal variability, the character of spectral power in the field and its spatial distribution, and the magnitude of simulated differences between the mean climate in the modern period (~1850-present), i.e. the typical calibration interval used in CFRs, and earlier periods during the last millennium. While it is impossible to perfectly characterize each of these climatic attributes—indeed a central pursuit of paleoclimatology is an attempt to provide estimates of these unknown quantities from proxy information—it is also true that state-of-the-art GCMs are variable in terms of how they simulate these various climate characteristics. For example, the spatial pattern and temporal stationarity of teleconnections between the tropical Pacific and regions of the global climate field can vary widely across the CMIP5/PMIP3 ensemble (Coats et al. 2013b, 2015a; Lewis and LeGrande 2015). In lieu of an ability to define with certainty the true spatiotemporal characteristics of the climate on multi-decadal-to-centennial timescales, the best option is therefore to test the dependence of PPE results over an ensemble of model simulations. Such a multi-model approach can highlight how the spatiotemporal characteristics of a given target field may influence the magnitude and spatial characteristics of the errors in CFRs as quantified by PPEs.

Multi-model approaches have not yet been widely applied in PPEs, particularly in attempts to assess the spatial performance of CFRs. Mann et al. (2007) presented PPE results using the ECHO-G ERIK1 and CCSM1.4 simulations and Christiansen et al. (2009) sampled a single simulation using a phase-randomizing procedure to create an ensemble of test climates characterized by the same spatiotemporal structures of the original simulation. Both of these studies noted the potential for PPE results to vary based on the simulation or ensemble member employed, but the studies focused primarily on the ability of CFRs to estimate robust hemispheric mean temperatures. Smerdon et al. (2011) first reported differences in the spatial performance of CFR methods that were dependent on the underlying model field using the ECHO-G ERIK2 and CCSM1.4 LM simulations. This preliminary result sets the stage for the current study in which an ensemble of CMIP5/PMIP3 simulations are used to construct a collection of PPEs in a homogenous framework to test four CFR methods. Similarities and differences in the performance of each CFR method are tracked across the model ensemble as a demonstration of how the underlying spatiotemporal characteristics of the target field may influence PPE assessments. We later focus specifically on the spectral fidelity of derived reconstructions and the model-dependent tendency for CFRs to enhance the mean difference between the Medieval Climate Anomaly (MCA) and the Little Ice Age (LIA). We conclude by providing recommendations on how to further test the influence of specific spatiotemporal characteristics in the target field on CFR performance and how our findings should be interpreted in terms of real-world reconstructions.

2 Data and methods

2.1 Model data and PPE design

The PPEs performed in this study use LM (note that we use LM as the acronym for the last millennium simulations used herein, but the CMIP5/PMIP3 designation for this particular modeling experiment is 'past1000') and historical simulations from five modeling centers as configured and implemented in CMIP5/PMIP3: the Beijing Climate Center CSM1.1 model (hereinafter BCC), the National Center for Atmospheric Research Community Climate System version 4 model (hereinafter CCSM), the Goddard Institute for Space Studies E2-R model (hereinafter GISS), the Institute Pierre-Simon Laplace CM5A-LR model (hereinafter IPSL) and the Max-Plank Institute ESM-LR model (hereinafter MPI). The LM simulations span the period 850–1850 C.E. and are forced with reconstructed time-varying exogenous forcings (Schmidt et al. 2011); the first ensemble member of the CMIP5 historical runs spanning the period 1850–2005 C.E. are appended to the LM simulations to produce model results from 850 to 2005 C.E. Although the appended simulations are not continuous, both the historical and LM simulations are generated using the same model configurations and resolutions. If the simulations have no drift, the discontinuity at 1850 C.E. should fall within the range of simulated climate variability, although all modes of variability will not be in phase across the discontinuity, particularly low-frequency modes such as the Atlantic Multidecadal Oscillation. A drift in the early centuries of the GISS LM simulation (Bothe et al. 2013) does not impact the discontinuity at 1850 C.E. and the uncorrected GISS LM simulation has been included in the collection of analyzed model output. The annual means of the modeled surface temperature fields are interpolated to even 5° latitude-longitude grids from which all samplings are performed (Smerdon et al. 2008b). For the reconstruction target field, each modeled temperature field is subsampled to approximate available instrumental temperature grids in the Brohan et al. (2006) surface temperature dataset resulting in a total of 1732 grid cells in the global field (Mann et al. 2008).

Pseudoproxies are sampled once from the 283 grid points that contain at least one proxy in the most populated nest of the M09 multiproxy network; all pseudoproxies are taken as available for the entire reconstruction interval and all are constructed from the mean annual surface temperature field of each simulation. This framework updates

Table 1 Ridge parameters selected using minimization of the GCV function for each of the ridge regression CFRs at all noise levels for each model PPE

SNR	BCC	CCSM	GISS	IPSL	MPI
Inf.	0.82	0.61	0.78	0.59	0.79
1.0	1.38	1.22	1.41	1.23	1.42
0.5	1.58	1.37	1.58	1.41	1.58
0.25	1.80	1.53	1.96	1.56	1.70

previous PPEs that have tested CFR performance using the Mann et al. (1998) multi-proxy network as the basis for pseudoproxy sampling (e.g. von Storch et al. 2004, 2006; Mann et al. 2005, 2007; Smerdon and Kaplan 2007; Smerdon et al. 2008a, 2010a, b, 2011; Christiansen et al. 2009), while complementing more recent work that either included preliminary results using pseudoproxy sampling approximating the M09 network (Smerdon et al. 2011) or used more comprehensive and realistic PPE designs that emulate the M09 network within the older CCSM1.4 (Wang et al. 2014) and ECHO-G ERIK2 (Evans et al. 2014) LM simulations.

The sampled pseudoproxies are perturbed at four white-noise levels to construct the pseudoproxy networks: signal-to-noise ratios (SNRs) of infinity (no noise), 1.0, 0.5 and 0.25, by standard deviation. Typical proxy records are estimated to have SNRs in the range of 0.5–0.25 (e.g. Mann et al. 2007; Wang et al. 2014). In all model cases, the same realization of 283 Gaussian white-noise series are used to perturb the pseudoproxy network and each noise level is achieved by rescaling the variance of the noise matrix to produce the desired SNR (Smerdon 2012). All tested methods are calibrated from 1850 to 1995 C.E., in keeping with the calibration interval employed by M09. Note that this again is an updated convention relative to many previous PPE frameworks that used a shorter calibration interval from 1856 to 1980 C.E.; all validation statistics are calculated during the reconstruction interval from 850 to 1849 C.E.

The above conventions are simplifications of real-world conditions. The noise in real proxies is typically multivariate (i.e. sensitive to climate variables in addition to temperature), non-stationary, and autocorrelated (e.g., Jacoby and D'Arrigo 1995; Briffa et al. 1998; Esper et al. 2005; Evans et al. 2002; Anchukaitis et al. 2006; Franke et al. 2013), while proxy sensitivity is typically seasonally dependent (e.g. Pauling et al. 2003; St. George et al. 2010). The modeled climates are considered to reasonably mimic real-world field statistics, but important features such as the strength and character of teleconnections vary across simulations and can be different from observations (e.g. Coats et al. 2013b). The adopted experimental setup therefore can be considered a best-case scenario for real-world conditions, whereas additional modifications to the PPE framework to more fully mimic real-world proxies will only degrade the CFR skill (e.g. von Storch et al. 2004, 2006; Mann et al. 2007; Wang et al. 2014; Evans et al. 2014).

2.2 CFR methods

Multivariate linear regression is the underlying formalism of most CFR methods used to date, although the last several years have seen important applications of newly emerging techniques (e.g. Tingley and Huybers 2010; Tingley et al. 2012; Steiger et al. 2014; Guillot et al. 2015). The basic approach of these linear regression methods relates a matrix of climate proxies to a matrix of climate data during a common time interval (generally termed the calibration interval) using a linear model. While this formalism is straightforward and well documented, it works best when the problem is overdetermined; that is, the time dimension is much larger than the spatial dimension, which allows the covariances to be more reliably estimated. This is typically not the case for most CFR scenarios, however, in which the number of target variables usually exceeds the time dimension, yielding a rank-deficient problem. For most global or NH CFRs, the number of grid cells in the climate field is typically on the order of many hundreds or a few thousands, while the observational record usually contains 150 annual fields or less. These conditions thus warrant some form of regularization. Published linear regression methods for large-scale temperature CFRs vary primarily in the form of this regularization and the manner in which the amount of regularization is chosen. We employ herein several of

Table 2 CCA dimensional reductions (d_{cca}, d_p, d_t) for the absolute minimum cross-validation errors and for the preferred solutions (in parentheses) based on the first local minimum in the cross-validation errors

SNR	BCC	CCSM	GISS	IPSL	MPI
Inf.	35, 35, 50 (21, 33, 33)	36, 45, 50 (21, 27, 38)	29, 35, 49 (29, 35, 49)	34, 37, 50 (34, 37, 50)	37, 42, 50 (21, 22, 49)
1.0	37, 38, 50 (29, 38, 50)	41, 48, 50 (26, 48, 34)	31, 31, 50 (26, 26, 34)	40, 50, 50 (30, 50, 33)	27, 27, 49 (25, 27, 49)
0.5	32, 40, 33 (24, 42, 29)	34, 48, 34 (10, 48, 11)	29, 50, 35 (11, 50, 14)	25, 39, 38 (18, 49, 20)	29, 41, 32 (10, 50, 11)
0.25	10, 45, 10 (7, 45, 7)	11, 40, 12 (5, 44, 5)	14, 48, 14 (3, 48, 3)	12, 44, 16 (2, 47, 2)	10, 41, 11 (7, 41, 9)

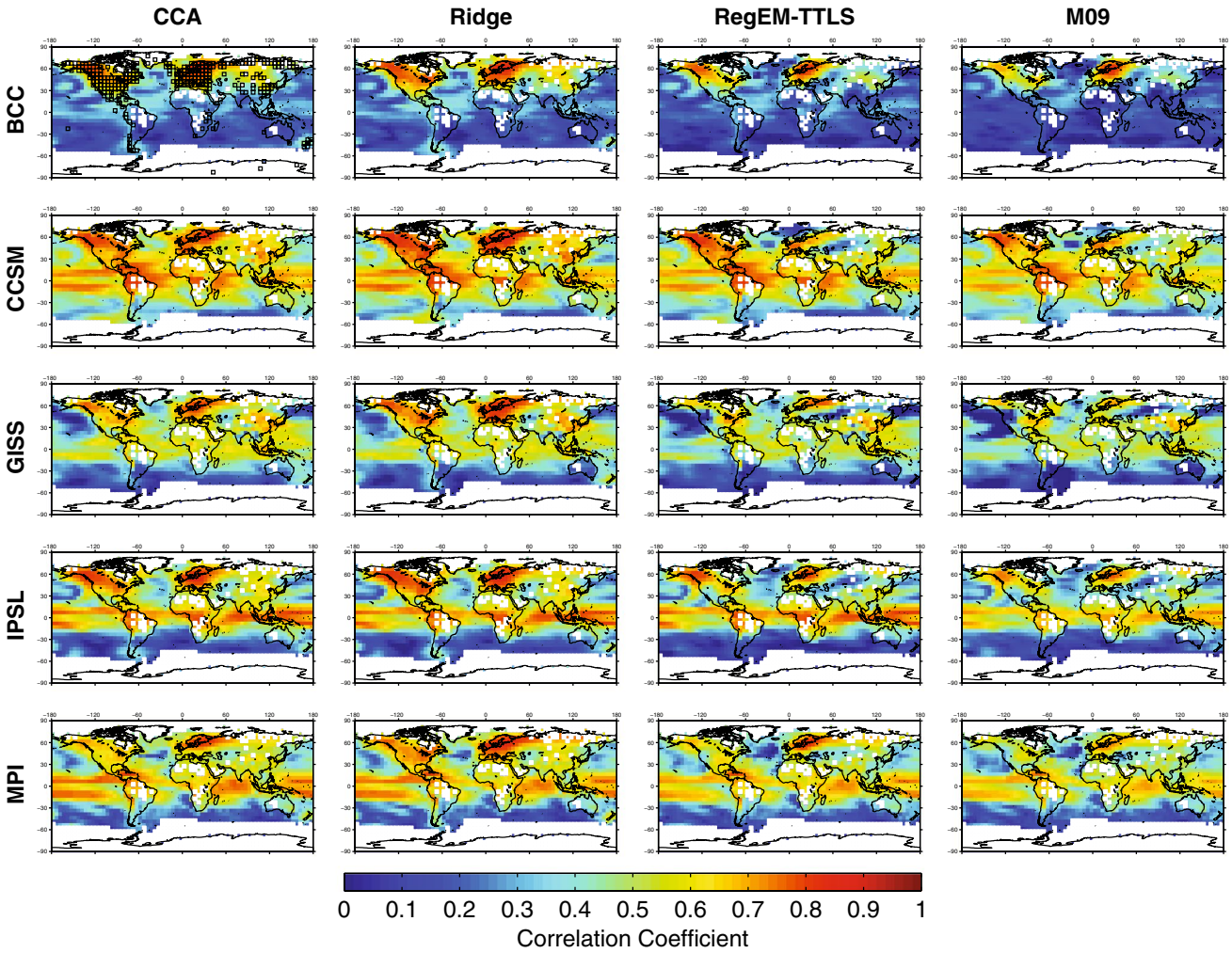

Fig. 1 Local correlation coefficients (Pearson's r) for the four CFR methods using pseudoproxies with SNRs of 0.5 and the five CMIP5/PMIP3 model fields. All methods use the same pseudoproxies, target field, and intervals for calibration (1850–1995 CE). Grid-point locations of the pseudoproxies used in all the PPEs are shown in the *upper left panel* as *open grid squares*, which approximate the distribution of the M09 network. Correlation coefficients are 95 % significant (two-sided p value < 0.05) in PPEs for values of about 0.06 or higher

the most commonly applied regularized linear regression techniques and describe their basic approaches in the following subsections.

2.2.1 *RegEM-TTLS*

We perform CFRs using two versions of regularized expectation maximization (RegEM) that both employ truncated total least squares (RegEM-TTLS) for regularization (Schneider 2001; Mann et al. 2007). The first is a standard version of RegEM-TTLS as originally described by Schneider (2001) and the second is the hybrid version applied by M09. The hybrid convention calibrates the multiproxy network on the target temperature field in split spectral domains by first separating the target field and the multiproxy (or pseudoproxy) network into high and low-frequency components. We follow the M09 convention by splitting these two domains at the 20-year period using a ten-point butterworth filter. The hybrid reconstruction is then derived by calibrating the pseudoproxy network in the two frequency domains using the RegEM-TTLS algorithm and subsequently combining the reconstructions from each domain to derive a complete field [see Mann et al. (2005, 2007) for further description of the hybrid method]. Throughout the remainder of this study, the standard RegEM-TTLS approach will be referred to as the RegEM-TTLS method and the hybrid approach as the M09 method. Note also that differences between reconstructions derived from the hybrid and standard versions of the RegEM method have been reported to be minimal (Rutherford et al. 2005; Mann et al. 2005, 2007; Smerdon et al. 2011), although the importance of hybrid calibrations

on the skill of the derived reconstructions has been debated (Rutherford et al. 2010; Christiansen et al. 2010).

A linear fit to the log-eigenvalue spectrum is used to determine the truncation parameter for the RegEM-TTLS CFRs in the same manner that was advocated by Mann et al. (2007) for the high-frequency component of their derived hybrid reconstructions. For the M09 CFRs, a linear fit to the log-eigenvalue spectrum was again used to determine the truncation parameter for the high-frequency component of the reconstructions, while the low-frequency truncation was determined by selecting the eigenvalue rank yielding 33 % of the cumulative variance in the low-frequency field. This percentage of retained cumulative variance is reduced from 50 %, as originally adopted by Mann et al. (2007); the value of 33 % has since been advanced by Rutherford et al. (2010) and M09 as more appropriate. A value of 10^{-4} was used for the stagnation tolerance and the inflation parameter was set to one for all versions of the RegEM CFRs.

2.2.2 *Ridge regression*

We apply standard ridge regressions (Hoerl and Kennard 1970) for the ridge regression CFRs in this study. The application of a single ridge regression was used by Smerdon et al. (2011), but is otherwise a break from earlier studies that have used ridge regression as the form of regularization in the iterative RegEM algorithm (RegEM-Ridge). The application of RegEM-Ridge for the purpose of CFRs for the Common Era has been discussed in detail in various publications (Schneider 2001; Mann et al. 2005; Smerdon and Kaplan 2007; Lee et al. 2007; Smerdon et al. 2008a, 2010a; Christiansen et al. 2009). We use standard ridge regression instead of RegEM-Ridge herein, because the iterative RegEM result converges to the single ridge regression result in the special case of our PPE design, namely when missing values comprise a single and regular block in the data matrix. Again in keeping with Smerdon et al. (2011), we determine the value of the ridge parameter for the single ridge regressions in the same manner applied by Schneider (2001) in RegEM-Ridge, that is, by minimization of the generalized cross validation (GCV) function (Golub et al. 1979). GCV selections of the ridge parameter in the ridge regression CFRs at all SNRs are provided in Table 1.

2.2.3 *Canonical correlation analysis*

Canonical correlation analysis (CCA) was applied as described in Smerdon et al. (2010a). Dimensions of the proxy and instrumental fields were both reduced by eigenvalue truncation, as were the number of retained canonical coefficients. These dimensional reductions were selected based on 'leave-half-out' cross-validation statistics, as described by Smerdon et al. (2010a). The CCA dimensional selections are given in Table 2 for the minimum cross-validation root mean squared error (RMSE), as well as those achieved for preferred dimensions taken as the first local minimum of the cross-validation RMSE to guard against artificial skill (Smerdon et al. 2010a). The preferred dimensions are used in all of the CCA CFRs in this study. For those cases in which the preferred dimensional selections were different from those of the absolute minimum RMSE, the mean RMSE is increased by only several thousandths of a Kelvin degree.

3 Evaluations of reconstruction performance

3.1 Correlation coefficients

Maps of the correlation coefficients (Pearson's r) calculated during the reconstruction interval (850–1849 C.E.) between the derived CFRs and the true model targets are plotted in

Table 3 Mean local correlation coefficient during the verification interval for all reconstructions and models in this study

SNR	BCC	CCSM	GISS	IPSL	MPI
CCA					
Inf	0.472	0.725 (0.642)	0.618	0.669	0.621
1.0	0.412	**0.666** (0.566)	0.542	0.587	0.588
0.5	**0.302**	**0.569** (0.463)	0.437	**0.478**	**0.495**
0.25	0.152	0.395 (0.281)	0.279	0.289	0.354
Ridge regression					
Inf.	**0.551**	**0.771**	**0.668**	**0.714**	**0.690**
1.0	**0.418**	**0.666**	**0.552**	**0.592**	**0.592**
0.5	0.295	0.560	**0.439**	0.475	0.491
0.25	0.154	0.382	0.269	0.295	0.336
RegEM-TTLS					
Inf.	0.283	0.614	0.459	0.499	0.521
1.0	0.264	0.567	0.436	0.469	0.483
0.5	0.231	0.532	0.377	0.432	0.461
0.25	**0.159**	0.398	0.320	**0.329**	**0.378**
M09					
Inf.	0.228	0.591	0.436	0.486	0.507
1.0	0.208	0.549	0.405	0.443	0.479
0.5	0.194	0.525	0.357	0.413	0.440
0.25	0.151	**0.430**	**0.324**	0.325	0.368

Numbers given in bold correspond to the best performance within each model experiment for each noise level

Numbers shown in parenthesis are from Smerdon et al. (2011) for a CCA PPE that used the older CCSM1.4 LM simulation, the same spatial sampling convention used herein for the pseudoproxy network and target field, and a shorter calibration interval (1856–1980 C.E.)

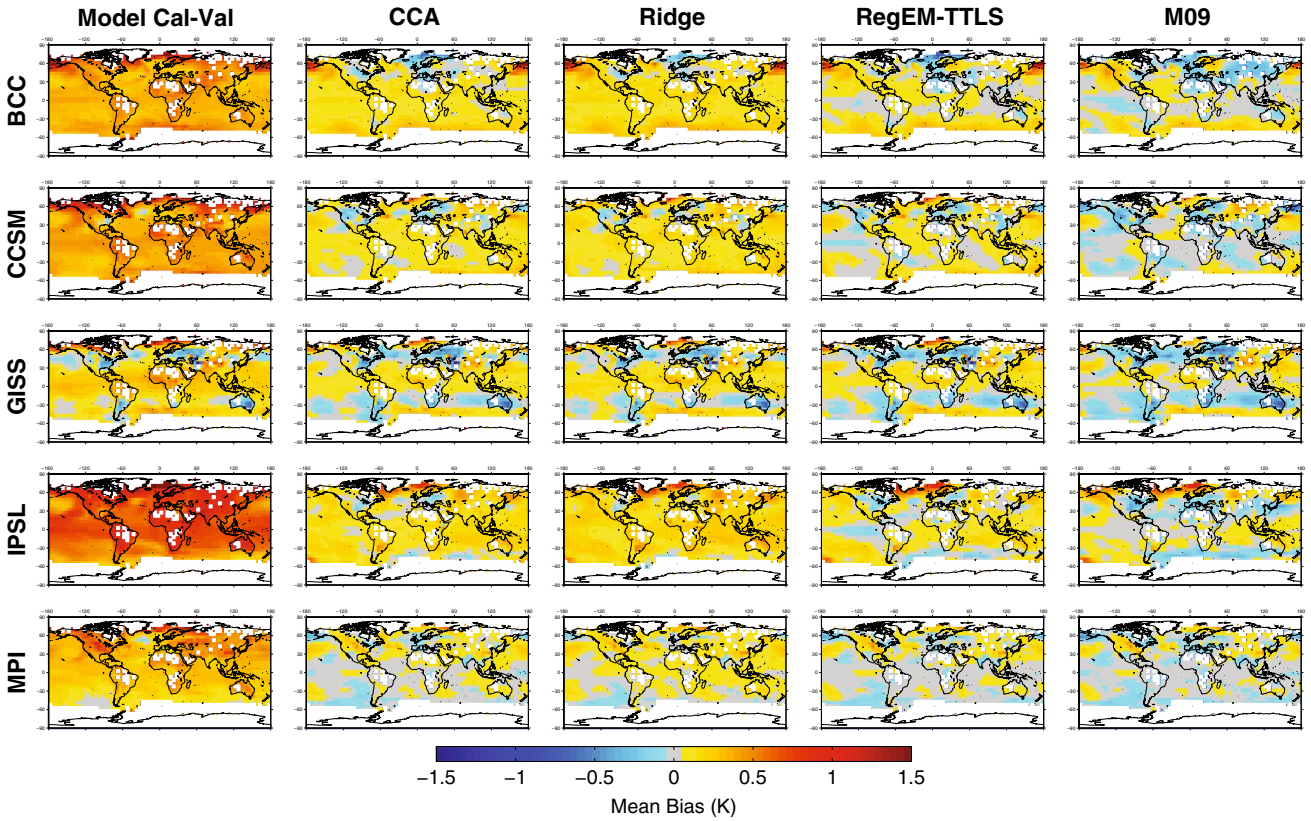

Fig. 2 Same as in Fig. 1, but for biases, that is, the mean differences between the CFRs and the true model fields during the reconstruction interval. The additional left column is the true mean difference between the calibration and validation intervals in each model

Fig. 1 using pseudoproxies with a SNR of 0.5; the spatially averaged mean results are given in Table 3 for all noise levels. The spatial patterns in the derived correlation fields are relatively consistent across all four methods within each model experiment and share many similarities with the patterns presented by Smerdon et al. (2011) using earlier LM simulations, a comparable collection of methods, and a sparser pseudoproxy network approximating the Mann et al. (1998) multiproxy distribution. One exception is that in all but the BCC model experiment, the correlation patterns reveal significant skill in tropical regions where limited pseudoproxy sampling exists. This skill throughout the tropics is in contrast to the more limited reconstruction skill in the same regions presented by Smerdon et al. (2011), indicating either improved sampling of the tropical teleconnections in the newer M09 network or stronger teleconnections between the tropics and pseudoproxy sampling locations in the new generation of models. The BCC experiments, and to some extent those from the GISS model, nevertheless indicate that the enhanced tropical skill is model dependent and therefore not solely a product of the spatial sampling of the pseudoproxy network or CFR method.

Additional features of the correlation maps in Fig. 1 include the now well-noted tendency for all methods to yield the largest correlation coefficients in regions of dense pseudoproxy sampling (Smerdon et al. 2011; Annan and Hargreaves 2012; Dannenberg and Wise 2013; Steiger et al. 2014; Wang et al. 2014; Evans et al. 2014). This robust performance feature is not only consistent across all methods, it is observed for all model experiments in this study. The regions containing the lowest correlation coefficient values are again in keeping with the earlier Smerdon et al. (2011) results and occur over the sparsely sampled Southern Hemisphere, while some experiments yield low values in the extratropical oceans as well. These reconstruction deficiencies occur despite a broader spatial sampling in the emulated M09 pseudoproxy network and highlight the continued need for improved proxy sampling of the Southern Hemisphere and oceans. Overall methodological performance, as represented by the mean of the global correlation coefficient fields (Table 3), is consistent across all of the model experiments. CCA and ridge regression modestly outperform the RegEM methods, but the latter methods collectively return the largest mean correlations across all models for the highest noise level. Interestingly, Smerdon et al. (2011) performed one CCA PPE with exactly the same pseudoproxy sampling used herein, but employed a shorter calibration interval (1856–1990 CE) and the older CCSM1.4 simulation. The mean results for this experiment are reported in Table 3 and indicate consistently less skill than

derived in the newer CCSM PPE, providing another example of the differences that arise based on the underlying spatiotemporal characteristics of the model simulation (although some reductions are likely due to a calibration interval that was shorter by 11 years).

3.2 Mean biases

Mean biases during the reconstruction interval are present in all CFRs (Fig. 2), but the magnitude and patterns of these biases are variable across methods and model experiments. CCA and ridge regression CFRs tend to be most consistently biased warm, while the two RegEM methods tend to generate CFRs that are more balanced by warm and cold biased regions. The GISS PPEs are an exception to the former rule, and to a lesser extent the IPSL results, in which CCA and ridge regression also tend to produce more balanced warm and cold biases in the CFRs. It was argued in Smerdon et al. (2011) that mean biases generally reflect the differences between the calibration and reconstruction interval means, which are also provided in Fig. 2. This observation is consistent with the results presented herein, although the tendency of the mean biases in the CFRs to reflect calibration-reconstruction interval mean differences is somewhat model dependent. For instance, there is a clear relationship between the patterns in the GISS model, while the relationship is much less evident in the BCC PPEs. Note that in some cases the mean bias patterns in the CFR appear to reflect a centered version of the calibration-reconstruction interval mean differences in the models, many of which present overall warmer conditions in the calibration interval relative to the reconstruction interval.

The globally averaged mean biases (Table 4) indicate that the RegEM-TTLS and M09 methods are generally least biased at smaller SNR levels, while CCA and ridge regression have similar or smaller average mean biases at higher SNRs; all methods nevertheless yield regional and aggregate CFRs with means different from the target. Notably, these biases are one statistic of the M09 CFRs that show a marked improvement over the RegEM-TTLS method, even though the overall errors in the M09 CFRs are not necessarily reduced relative to other methods (Table 5); this finding is consistent with Smerdon et al. (2011) who demonstrated the same result in PPEs derived from earlier model simulations. As we discuss in Sect. 3.6, the important implication of the mean bias patterns is the possibility that large biases exist in dynamically important regions, examples of which can be seen across multiple methods and model experiments in Fig. 2.

Table 4 Mean bias during the verification interval for all reconstructions and models in this study

SNR	BCC	CCSM	GISS	IPSL	MPI
CCA					
Inf	**0.037**	0.009 (−0.002)	**0.001**	0.026	**0.002**
1.0	0.056	0.036 (0.035)	**0.014**	0.054	0.008
0.5	0.117	0.099 (0.116)	0.051	0.142	0.036
0.25	0.240	0.239 (0.202)	0.113	0.360	0.127
Ridge regression					
Inf.	0.043	**0.007**	0.012	**0.023**	0.003
1.0	0.081	0.059	0.030	0.086	0.026
0.5	0.153	0.134	0.074	0.197	0.070
0.25	0.271	0.272	0.136	0.404	0.164
RegEM-TTLS					
Inf.	0.045	0.035	−0.005	0.041	−0.006
1.0	**0.054**	0.047	0.021	0.065	0.006
0.5	0.074	0.067	0.041	0.087	0.020
0.25	0.128	0.123	0.061	0.155	0.049
M09					
Inf.	0.075	0.027	0.023	0.029	−0.008
1.0	0.069	**0.022**	0.020	**0.035**	**−0.005**
0.5	**0.032**	**0.001**	**−0.002**	**0.040**	**−0.002**
0.25	**0.048**	**0.003**	**0.001**	**0.067**	**0.017**

Numbers given in bold correspond to the best performance within each model experiment for each noise level

Numbers shown in parenthesis are from Smerdon et al. (2011) for a CCA PPE that used the older CCSM1.4 LM simulation, the same spatial sampling convention used herein for the pseudoproxy network and target field, and a shorter calibration interval (1856–1980 C.E.)

3.3 Changes in reconstructed spectral power

Variance losses are expected for linear CFR methods that blend signal and error variances as a characteristic of formulation (e.g. von Storch et al. 2004). All derived CFRs suffer variance losses (Table 6), the patterns of which vary appreciably between methods and models (Fig. 3). Ridge regression and CCA display similar patterns within each model experiment (Fig. 3), although variance losses are larger for ridge regression. These two methods generally exhibit the well-behaved characteristic of preserving more variance in regions where correlation coefficients are largest, but this is not as clearly the case across all models as was reported in Smerdon et al. (2011). For instance, the pattern of standard deviation ratios in BCC is maximized through the tropics, and is less pronounced through the dense pseudoproxy sampling regions of North America and Europe. This is in contrast to the patterns of correlation coefficients, which are all maximized in the dense pseudoproxy sampling regions in the BCC experiments. The RegEM-TTLS and M09 CFRs exhibit the largest standard deviation ratios throughout all of the model experiments, but more consistently enhance variance in areas where correlation coefficients are small. This is particularly true through the tropics and subtropics across almost all of the models, indicating that much of the variance in the RegEM-TTLS and M09 experiments is associated with noise.

Table 5 Mean RMSE during the verification interval for all reconstructions and models in this study

SNR	BCC	CCSM	GISS	IPSL	MPI
CCA					
Inf	0.440	0.385 (0.394)	0.380	0.360	0.414
1.0	**0.457**	**0.425** (0.450)	0.417	**0.415**	0.437
0.5	**0.515**	**0.499** (0.532)	0.470	**0.487**	**0.490**
0.25	0.602	0.611 (0.639)	0.529	0.646	0.551
Ridge regression					
Inf.	**0.388**	**0.339**	**0.352**	**0.326**	**0.372**
1.0	**0.457**	0.427	**0.414**	0.416	**0.435**
0.5	0.524	0.506	**0.468**	0.506	**0.490**
0.25	0.611	0.624	0.533	0.664	0.564
RegEM-TTLS					
Inf.	0.518	0.470	0.459	0.474	0.480
1.0	0.528	0.513	0.475	0.487	0.495
0.5	0.545	0.535	0.499	0.507	0.506
0.25	0.581	0.622	0.521	0.564	**0.547**
M09					
Inf.	0.535	0.486	0.479	0.482	0.485
1.0	0.544	0.515	0.488	0.497	0.498
0.5	0.544	0.529	0.505	0.508	0.512
0.25	**0.573**	**0.589**	**0.519**	**0.558**	0.552

Numbers given in bold correspond to the best performance within each model experiment for each noise level

Numbers shown in parenthesis are from Smerdon et al. (2011) for a CCA PPE that used the older CCSM1.4 LM simulation, the same spatial sampling convention used herein for the pseudoproxy network and target field, and a shorter calibration interval (1856–1980 C.E.)

Table 6 Mean standard deviation ratio during the verification interval for all reconstructions and models in this study

SNR	BCC	CCSM	GISS	IPSL	MPI
CCA					
Inf	0.646	0.801 (0.781)	**0.757**	**0.808**	0.714
1.0	**0.556**	0.711 (0.597)	0.607	**0.688**	**0.627**
0.5	0.460	0.597 (0.451)	0.493	0.571	0.530
0.25	0.351	0.430 (0.311)	0.294	0.402	0.402
Ridge regression					
Inf.	**0.654**	**0.816**	0.739	0.791	**0.739**
1.0	0.494	0.663	0.554	0.635	0.572
0.5	0.413	0.550	0.431	0.531	0.477
0.25	0.328	0.395	0.267	0.409	0.360
RegEM-TTLS					
Inf.	0.495	0.761	0.631	0.671	0.639
1.0	0.541	**0.713**	0.677	0.622	0.585
0.5	**0.541**	**0.736**	**0.644**	**0.665**	**0.591**
0.25	**0.561**	**0.790**	**0.635**	**0.724**	**0.642**
M09					
Inf.	0.441	0.741	0.692	0.623	0.630
1.0	0.455	0.690	**0.688**	0.596	0.597
0.5	0.483	0.682	0.628	0.592	0.559
0.25	0.507	0.737	0.614	0.686	0.622

Numbers given in bold correspond to the best performance within each model experiment for each noise level

Numbers shown in parenthesis are from Smerdon et al. (2011) for a CCA PPE that used the older CCSM1.4 LM simulation, the same spatial sampling convention used herein for the pseudoproxy network and target field, and a shorter calibration interval (1856–1980 C.E.)

The standard deviation ratios in Fig. 3 do not indicate which part of the spectral domain preserves the targeted variance. We quantify the proportion of preserved high and low-frequency variance by plotting the standard deviation ratios between the CFRs and model truth in high and low-frequency domains split at the 20-year period using a ten-point butterworth filter (Figs. 4, 5, respectively). A comparison between Figs. 3, 4 and 5 indicates that all of the methods preserve or enhance low-frequency variability in regions where variance preservation is maximized (Fig. 3, e.g. the tropics and subtropics), while reducing high-frequency variability in those regions (see also Tables 6, 7). This uneven influence on variability in the CFRs changes the ratio of spectral power in the estimated climate field, such that the power spectra of the time series in the affected grid cells is reddened relative to the true model field. This is further illustrated in Fig. 6, which plots the difference between the scaling exponent, β, of the spectral density in the CFRs and target model field (e.g. Huybers and Curry 2006; Franke et al. 2013). Red regions of the maps in Fig. 6 are those areas where the reconstruction methodologies and/or the spectral sampling biases of the pseudoproxy network cause the derived CFR to be redder than the true model field; blue regions are the areas where the CFR is bluer than truth. These results indicate that the estimated spectral biases of a derived CFR are expressed in complicated spatial patterns and that a simple rule-of-thumb characterization about how reconstruction methodologies influence the spectral characteristics of a climate field may be difficult. For instance, Franke et al. (2013) use previous literature to argue that reconstruction methods bias reconstruction spectra blue, but such a characterization is clearly not the case for the global CFRs considered herein. Specifically, the widespread reddening of the tropics in the derived CFRs argue that methodologies or spatial sampling can in fact redden the reconstructed field, thereby enhancing the red biases that Franke et al. (2013) argue to be inherent in tree-ring proxies. In this context, it should be noted that all of the pseudoproxies in this study have used only Gaussian white-noise perturbations. The overall impact on the spectral power of each pseudoproxy should therefore be minimal, but any effect would tend to make the pseudoproxies more blue. The red biases evident in Fig. 6 therefore must be a product of either the CFR methodology or spatial sampling biases, and not because of any biases inherent in the pseudoproxies themselves, i.e. the kinds of biases that were most directly investigated by Franke et al. (2013).

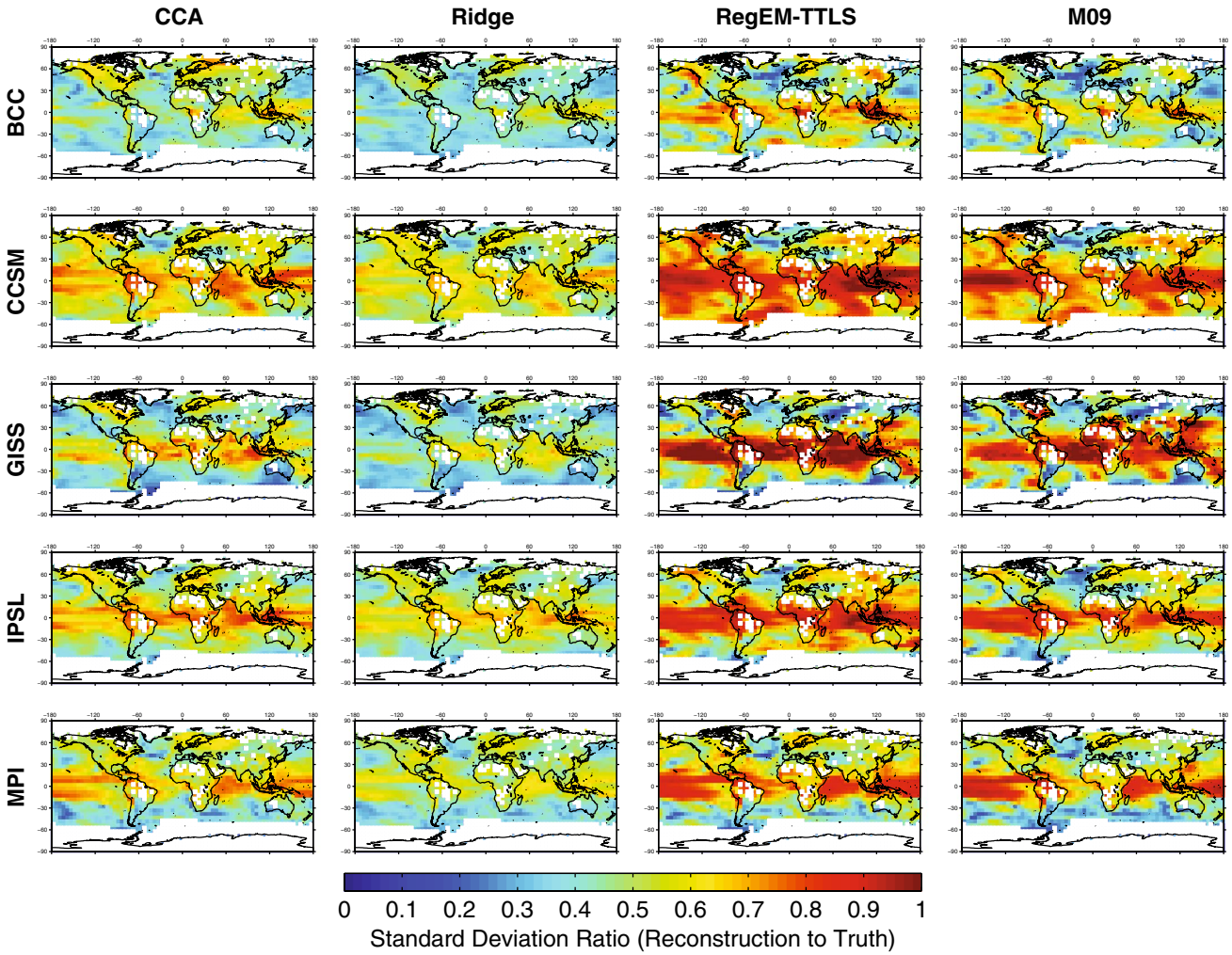

Fig. 3 Same as in Fig. 1, but for the ratio between the sample standard deviations estimated in the reconstruction interval for each CFR and the true model fields

3.4 RMSE

Total CFR RMSE tends to be a trade off between the preserved signal variance and biases (Wang et al. 2014). Despite the relative performance across the evaluated methods in preserved variance, standard deviation ratios and biases, the total error across CFRs within a given model are all quite similar (Table 5), as are the spatial patterns (Fig. 7). Additionally, the mean RMSE for a given method and SNR across all of the model experiments is typically very similar in magnitude, and by consequence the relative increase in RMSE with pseudoproxy noise is also consistent across models (Table 5). Despite the relative consistency in RMSE, the CCA and ridge regression methods generate CFRs with the smallest total errors at low and intermediate noise levels, while the two RegEM methods yield the smallest total error for the highest noise case (SNR = 0.25). These results indicate that no single CFR method can be considered to have universally beneficial characteristics, and more generally that the choice of CFR method will involve a trade off in respective errors.

3.5 Global Mean Index performance

The timeseries of area-weighted global mean temperature for the reconstruction methodologies are plotted in Fig. 8. While each method is largely successful at reconstructing global temperature at low noise levels (mean correlations across all models and methods of ~0.9 for the infinite and 1.0 SNRs in Table 8), there are inter-model and methodological differences particularly at high noise levels. For instance, global temperature appears more difficult to reconstruct in models with weak global temperature variability such as BCC and to a lesser degree GISS and IPSL (right panels of Fig. 8). At high noise levels, however, the RegEM-based CFR methodologies in all models capture

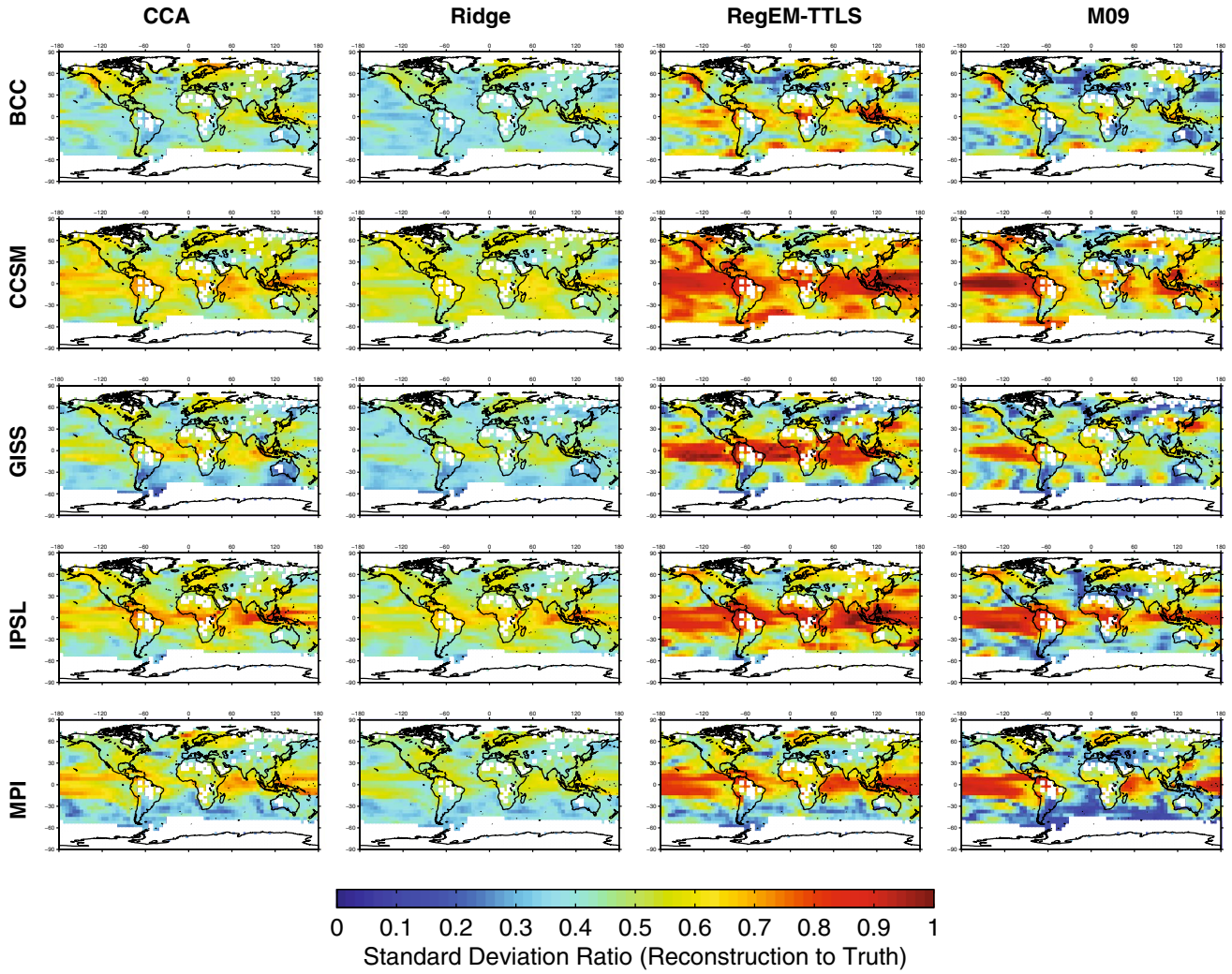

Fig. 4 Same as in Fig. 1, but for the ratio between the high-frequency (<20-year periods) sample standard deviations estimated in the reconstruction interval for each CFR and the true model fields. The spectral separation of the high-frequency domain in each of the model fields was done using a ten-point butterworth filter

better the timing and magnitude of global temperature variability, with the M09 framework being particularly successful (Fig. 8; Table 8). These results suggest that while no single reconstruction methodology is best able to capture the characteristics of the full climate field, the RegEM-based methods are most skillful at reconstructing the features of global temperature indices at realistic noise levels.

3.6 MCA and LIA mean differences in the tropical Pacific

Herein we evaluate the implications of regional errors in the derived CFRs in the context of a prominent dynamical interpretation of the M09 CFR: a cold tropical Pacific during the MCA (950–1250) relative to the LIA (1400–1700), which is evident in the full temperature field reconstruction and in the mean Niño3 (2.5°S–2.5°N, 92.5°W–147.5°W) index extracted from the CFR. This result has buoyed hypotheses that a La Niña-like (or anomalously cool) tropical Pacific was the dominant driver of megadroughts in the southwest of North America, all of which occurred during the MCA (e.g. Cook et al. 2007; Herweijer et al. 2007). It is important to note that the current generation of GCMs does not simulate a forced radiative response in the tropical Pacific during the MCA and differences between MCA and LIA periods in LM simulations do not match the pattern in the M09 CFR (e.g. M09; González-Rouco et al. 2011). It is difficult to determine, however, whether this disagreement between models and the reconstruction is because models do not capture a forced response in the tropical Pacific, the forcing used to drive the model simulations is unrealistic, the MCA La-Niña conditions suggested by

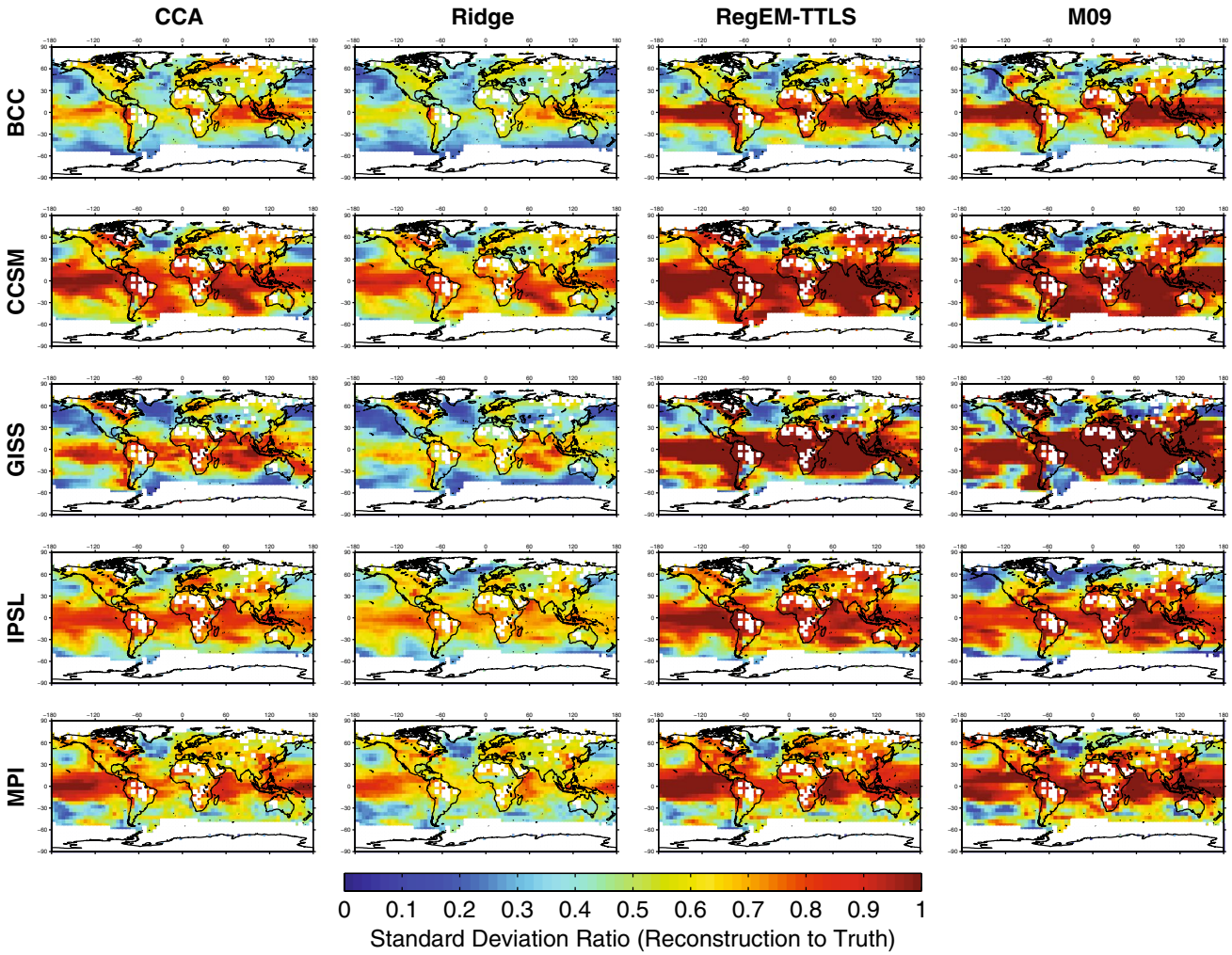

Fig. 5 Same as in Fig. 1, but for the ratio between the low-frequency (>20-year periods) sample standard deviations estimated in the reconstruction interval for each CFR and the true model fields. The spectral separation of the low-frequency domain in each of the model fields was done using a ten-point butterworth filter

Table 7 Mean low (high)-frequency standard deviation ratios during the verification interval for all reconstructions and models using pseudoproxies with an SNR of 0.5

Method	BCC	CCSM	GISS	IPSL	MPI
CCA	0.521 (0.453)	0.714 (0.556)	0.572 (0.462)	0.643 (0.549)	0.642 (0.492)
Ridge	0.451 (0.411)	0.637 (0.523)	0.476 (0.414)	0.577 (0.520)	0.560 (0.451)
RegEM-TTLS	0.636 (0.527)	0.858 (0.694)	0.812 (0.573)	0.738 (0.645)	0.710 (0.551)
M09	0.645 (0.443)	0.840 (0.604)	0.885 (0.476)	0.691 (0.536)	0.726 (0.473)

the reconstruction are in error, or the La-Niña conditions during the MCA were the product of internal variability. In the case of the latter, it has been shown that some models simulate megadroughts that are consistently associated with cold-states in the tropical Pacific arising from internal variability (Coats et al. 2013a, 2015a).

Despite the ambiguity of the causes of megadroughts in the past, the present pseudoproxy results can be used to evaluate the robustness of MCA-LIA mean differences from the perspective of whether or not spatial errors in derived CFRs may increase or decrease the probability of estimating MCA-LIA differences that are of the same

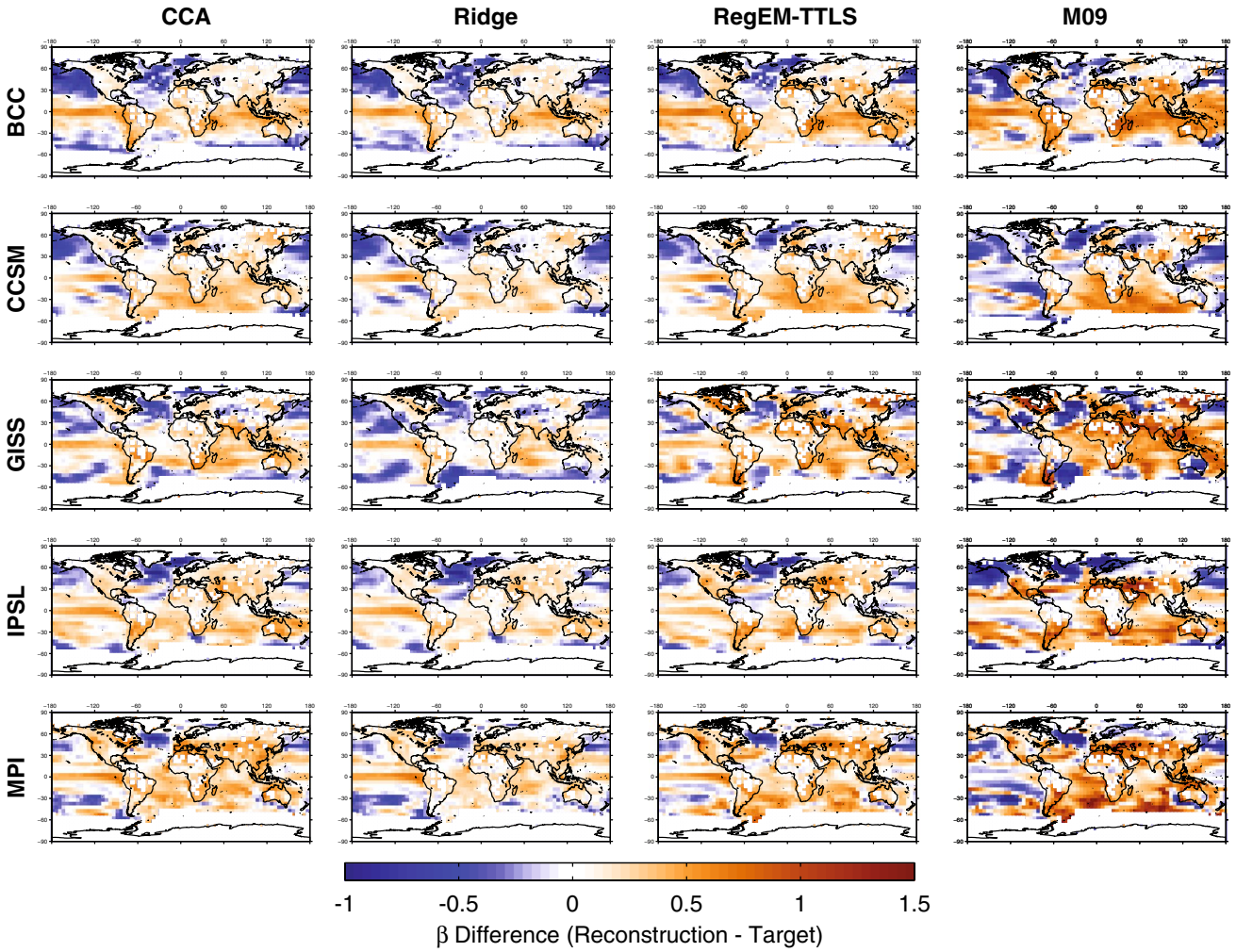

Fig. 6 Same as in Fig. 1, but for the difference between the scaling exponent, β, of the spectral density values estimated in the reconstructions and target model fields

magnitude as estimated by M09. Toward such ends, we ask three questions in our subsequent analyses: (1) Do models simulate paired 300-year periods (the length of the MCA and LIA defined in M09) with temperature differences in the tropical Pacific that are as large as those between the MCA and LIA in the M09 reconstructed Niño3 SST index? (2) If these periods exist in the simulations, are they altered significantly by the errors introduced by the CFR methodologies? and (3) Are the spatial patterns of the full temperature field for these periods indicative of a La Niña-like tropical Pacific?

Figure 9 plots the MCA-LIA mean differences in the true model simulations and the CFRs derived in each PPE across all methods and models. The true model differences further confirm what has been demonstrated elsewhere (e.g. M09; González-Rouco et al. 2011), namely that the employed GCMs do not simulate a colder tropical Pacific during the MCA, relative to the LIA. Similarly, the derived CFRs do not generate estimates of MCA tropical Pacific temperature that are colder than the LIA. More subtle differences are noted, however, in the patterns of MCA-LIA mean differences in all of the CFRs relative to the model truths. The GISS-based PPEs, for instance, all generate CFRs with a pronounced cold region of the northern Pacific that is not present in the true model field, while underestimating the warm differences in the northern Atlantic. One caveat for the GISS results is that the MCA is the period most affected by the drift in the simulation (Bothe et al. 2013), and the MCA-LIA difference patterns therefore may be a product of the drift. It is less likely, however, that the drift is relevant for dissimilarities between the true model MCA-LIA difference pattern and those in the CFRs, in light of the fact that each of the CFR methods reproduce the early period of the GISS simulation in the global mean (Fig. 8). Regardless of the particular details

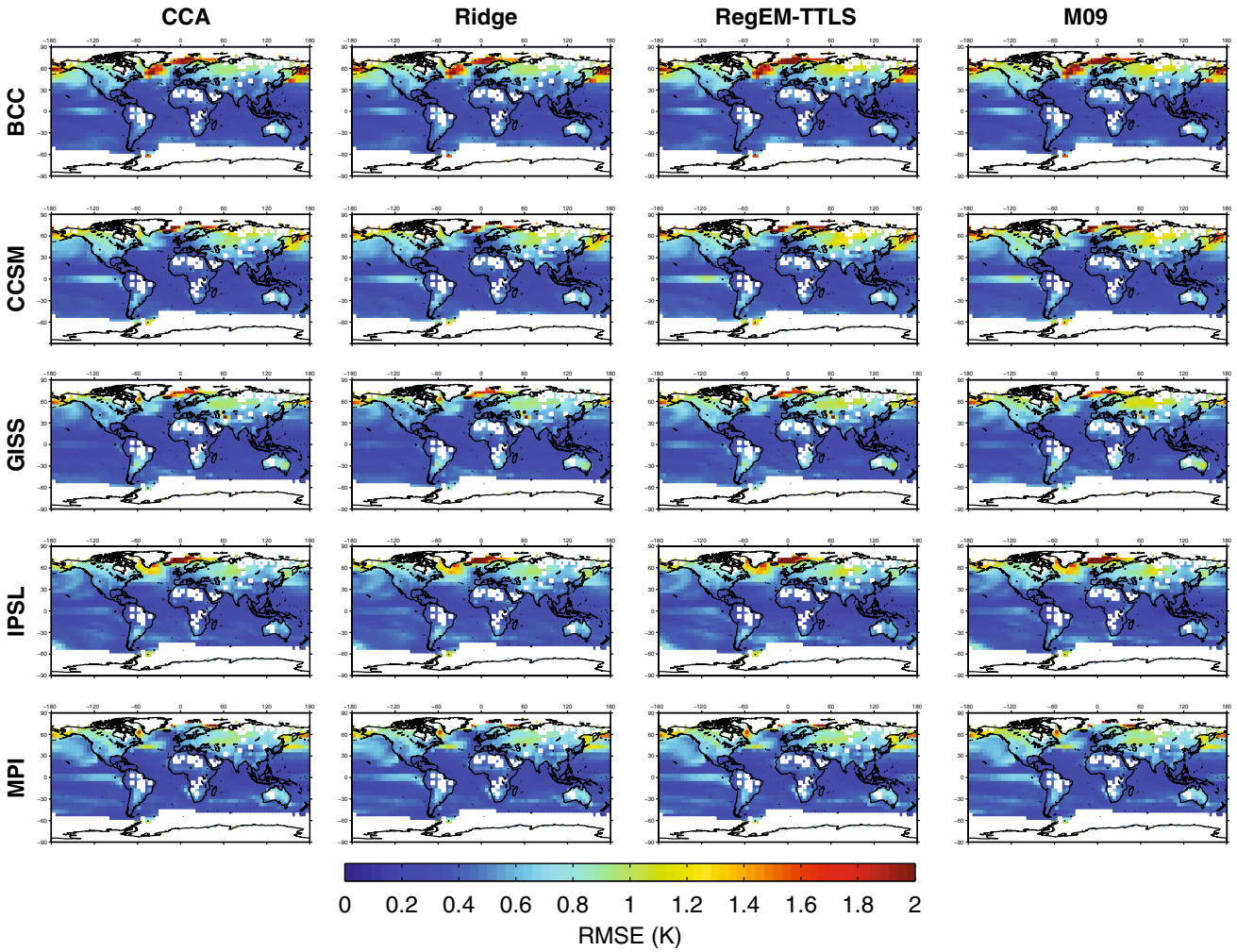

Fig. 7 Same as in Fig. 1, but for the RMSE for each CFR, relative to the true model fields

of each simulation, the collective implication of Fig. 9 is that errors in derived CFRs can alter the estimated mean differences between various temporal periods and should be interpreted cautiously.

To focus more specifically on the tropical Pacific, we evaluate the difference between the MCA and LIA in the Niño3 SST index derived from M09, the value of which is 0.21 °C. For comparison, we randomly choose two independent 300-year periods between 850 and 1850 C.E., repeated 1000 times, from both the true model simulations and the associated M09 pseudoproxy CFRs from the SNR = 0.5 experiments. Figure 10 plots the fraction of the time that the squared difference between the two randomly drawn periods in the mean Niño3 SST indices is as large as the difference in M09 (differences are squared to avoid the arbitrary distinction of which period is warmer or colder in this experiment). Only the IPSL model contains any paired periods in which the true simulation has a mean Niño3 temperature difference that is equal or greater than the MCA-LIA difference in M09. More remarkably, however, the CCSM and GISS experiments both yield pseudoproxy CFRs using the M09 method that generate periods in which the differences in the Niño3 SST index are above the value of the MCA-LIA difference in M09, even when the true model simulations contain no such periods. For the IPSL model, the number of these periods in the CFR is reduced relative to the true model field, while the BCC and MPI experiments include no periods with differences that exceeded the M09 MCA-LIA value in either the true model simulations or the M09 pseudoproxy CFRs.

Figure 11 shows the squared difference in the mean Niño3 SST index from the true model output for all randomly selected 300-year pairs in which the CFR derived from the M09 method has a squared difference of at least $0.044°C^2$ (the actual value of the MCA-LIA squared difference from the M09 reconstructed Niño3 SST index). The figure clearly illustrates that the squared difference in the true model output from the CCSM and GISS

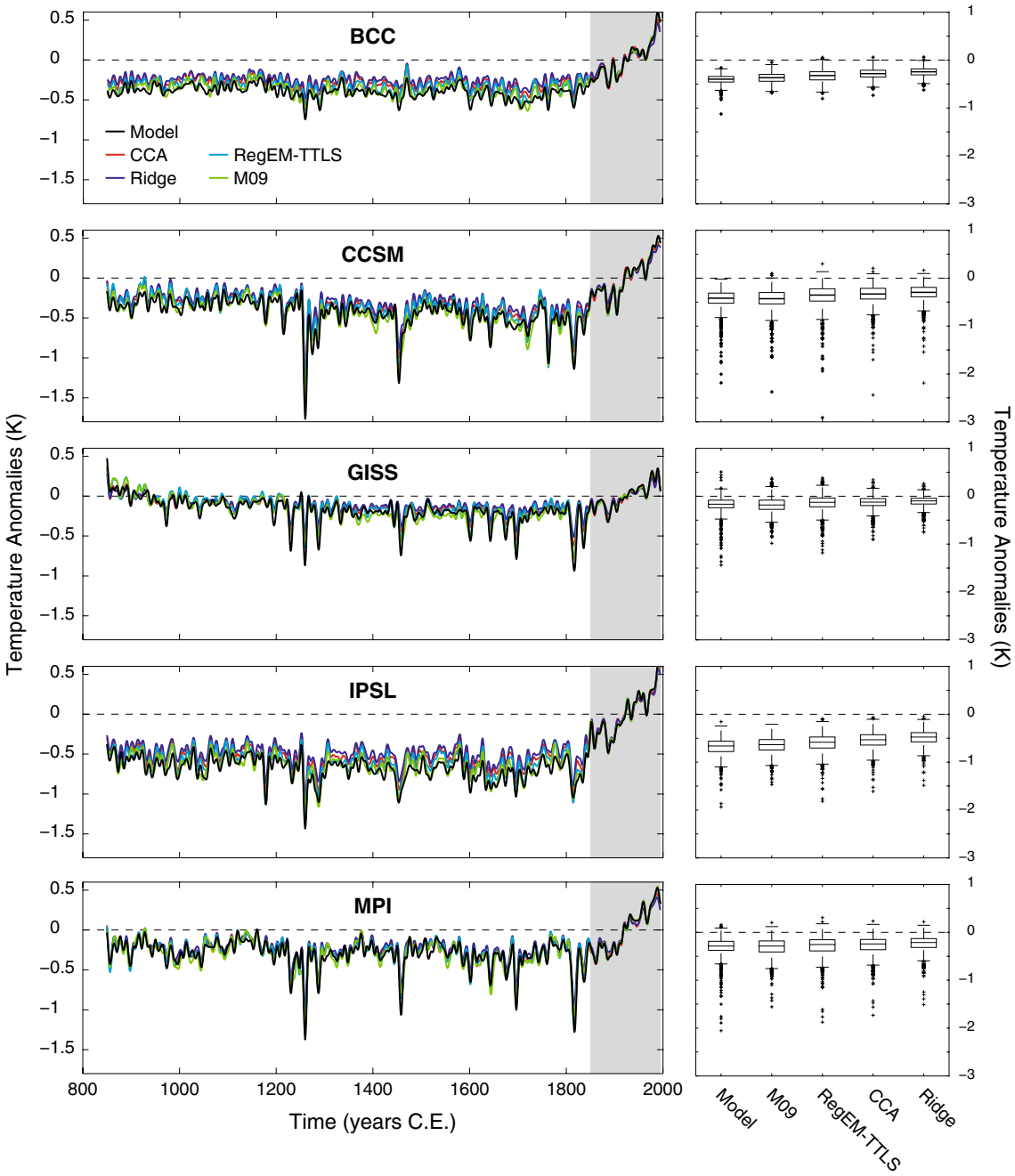

Fig. 8 (*left*) Area-weighted global mean time series for the four CFR methods and the associated true model fields. Time series have been smoothed using a decadal low-pass filter (ten-point butterworth). *Grey shaded areas* represent the calibration interval (1850–1995 C.E.) (*right*) *Box plots* associated with each of the global mean time series shown on the *left*. These plots were calculated from the distribution of the individual annual means in each global time series during the reconstruction interval

simulations is smaller than the squared difference in the associated M09 pseudoproxy CFR in every case. This is particularly stark in the GISS simulation, in which the value of the true model output doesn't reach half the value of even the smallest difference in the associated M09 pseudoproxy CFR. The IPSL simulation exhibits the opposite behavior, with the squared difference being larger for the true model output than the M09 pseudoproxy CFR. Importantly, this result indicates that the M09 methodology can itself produce large temperature differences in the tropical Pacific between 300-year periods. While the direction of this behavior is model dependent, these results indicate the potential for CFRs to produce temperature differences in the tropical Pacific that are in excess of the model truth.

Table 8 Northern Hemisphere mean correlation during the verification interval for all reconstructions and models in this study

SNR	BCC	CCSM	GISS	IPSL	MPI
CCA					
Inf	0.916	0.981 (0.972)	0.982	0.976	0.973
1.0	**0.867**	**0.960** (0.940)	**0.951**	**0.938**	**0.947**
0.5	0.741	0.906 (0.867)	0.877	0.851	0.878
0.25	0.461	0.726 (0.638)	0.640	0.606	0.695
Ridge regression					
Inf.	**0.928**	**0.986**	**0.983**	**0.982**	**0.978**
1.0	0.863	0.958	0.948	**0.938**	0.943
0.5	**0.742**	0.903	0.879	0.852	0.879
0.25	0.493	0.735	0.681	0.628	0.711
RegEM-TTLS					
Inf.	0.849	0.968	0.946	0.947	0.955
1.0	0.808	0.942	0.922	0.918	0.928
0.5	0.695	**0.908**	**0.887**	**0.862**	**0.886**
0.25	0.506	0.793	0.767	0.691	0.755
M09					
Inf.	0.698	0.950	0.930	0.934	0.940
1.0	0.687	0.916	0.903	0.890	0.911
0.5	0.632	0.905	0.878	0.841	0.855
0.25	**0.560**	**0.818**	**0.827**	**0.698**	**0.756**

Numbers given in bold correspond to the best performance within each model experiment for each noise level

Numbers shown in parenthesis are from Smerdon et al. (2011) for a CCA PPE that used the older CCSM1.4 LM simulation, the same spatial sampling convention used herein for the pseudoproxy network and target field, and a shorter calibration interval (1856–1980 C.E.)

Figure 12 shows the composite of the difference in the average temperature field for all paired 300-year periods in which the M09 pseudoproxy CFRs produced a squared difference in the Niño3 SST index that is as large as the MCA-LIA value in M09. To maintain consistency with M09, the temperature difference was calculated by subtracting the average reconstructed grid point temperature for the 300-year period with the warmer Niño3 index mean from the period with the colder Niño3 index mean. While M09 suggests a La Niña-like tropical Pacific and positive temperatures (relative to the LIA) outside of the tropical Pacific, none of the models reproduce these characteristics in the true difference composites shown in Fig. 12. This is also the case for the M09 pseudoproxy CFRs, with the exception of GISS, which does have a warm extratropical Pacific in the pseudoproxy CFR difference composite, despite the fact that the feature is not present in the true model output. For the tropics, the pseudoproxy CFR difference composite is weakly La Niña-like for every model, but this is not the case for the actual CCSM and GISS output and the magnitude is less pronounced in the true model output for IPSL. These results indicate that the M09 CFR method can itself produce a La Niña-like spatial footprint, when in fact the Niño3 temperature difference is the consequence of globally cooler temperatures that are not La Niña-like in character.

Collectively, these results further imply what was suggested by Wang et al. (2014): the M09 CFR method and associated proxy sampling distribution, while likely successful at capturing global and hemispheric temperature variability, generates errors on regional spatial scales that complicate interpretations of the actual M09 CFR. Specifically, the M09 CFR method has been shown herein to produce unrealistically large centennial-scale tropical Pacific temperature changes and a spatial footprint of temperature variability that is erroneously La Niña-like in character in some models. While these results are indeed synthetic and model dependent, they suggest that regionally and dynamically specific features of current CFRs must be interpreted carefully and definitive statements about MCA-LIA dynamical conditions derived from these CFRs are currently difficult to justify.

4 Conclusions

The spatial performance of four CFR methods identifies some limits on the ability of currently employed multivariate linear techniques to extract information from sparse and noisy observations. No single method produced CFRs with universally advantageous characteristics, and it is difficult to advocate for the singular benefit of one method over another. Hybrid CFR

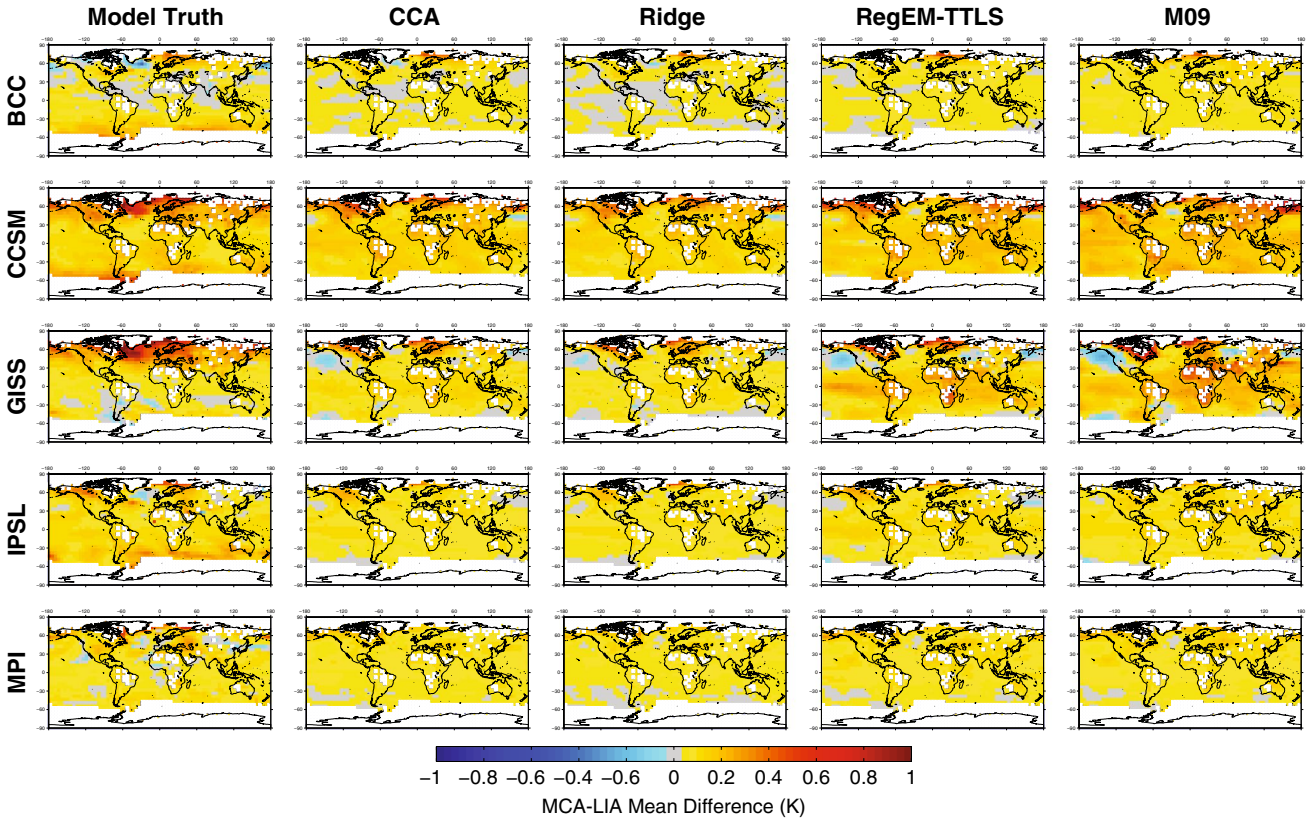

Fig. 9 Same as in Fig. 1, but for the difference between the MCA and LIA periods as defined in M09 (950–1250 and 1400–1700 C.E., respectively). In addition to the reconstructed differences, the true model differences are shown in the *far left column*

methodologies also do not appear to circumvent the noted deficiencies, despite the fact that these methods are the most successful in reproducing global and hemispheric mean temperature indices. Global means are nevertheless insufficient for evaluating the spatial performance of CFR methods given that skillful global mean reconstructions do not generally track success in various spatial assessment metrics. Smerdon et al. (2011) also noted that correlation coefficients are largest in regions of densest pseudoproxy sampling, an observation that is supported by our results, but specific model dependencies indicate the importance of evaluating CFR methods with multiple model-based PPEs. Specifically, comparison between the spatiotemporal characteristics in employed model fields and observed climate fields appears essential for determining the applicability of PPEs to assessments of real-world CFRs. A comprehensive PPE using multiple last millennium simulations (as has been completed herein) combined with instrumental and paleoclimate information on the true climate system is therefore necessary to ascertain the likelihood of the model-specific spatial errors occurring in actual CFRs.

There are nevertheless general conclusions that can be made from the PPEs derived herein about the likely errors inherent to CFRs. Firstly, the tested CFR methods can alter the ratio of spectral power at all locations in the

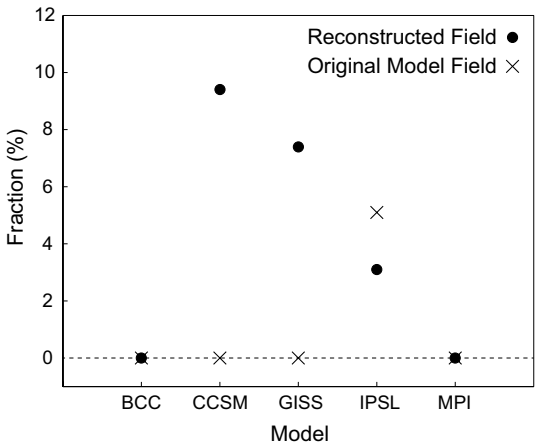

Fig. 10 Fraction of 1000 randomly selected 300-year pairs with squared difference in the mean Niño3 index ($°C^2$) larger than the MCA and LIA squared difference in the Niño3 index from M09. The *circles* are based on the Niño3 index taken from the M09 pseudoproxy CFR from each model and the *crosses* are based on the true model Niño3 index

field. While the patterns of these biases are largely model dependent, large negative and positive biases in the ratio of spectral power are both possible. This contradicts the

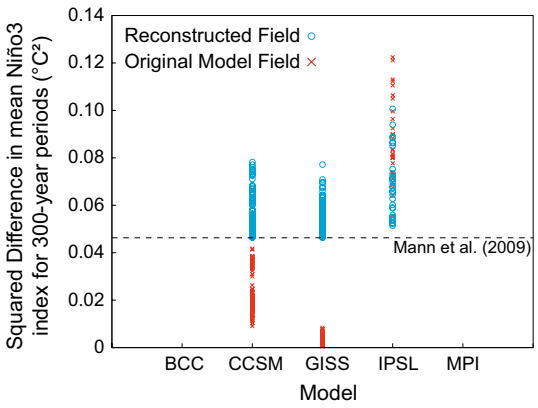

Fig. 11 Squared difference in the mean Niño3 SST index between 300-year periods for the true model output (*cross symbols*) and the M09 CFR (*circle symbols*) during each set of 300-year periods where the M09 pseudoproxy CFR value is as large as the MCA-LIA squared difference (*dashed line*) in M09

conclusions of Franke et al. (2013) regarding the likelihood that blue spectral biases are imposed specifically by CFR methods and suggests that they can in fact redden climate signals, independent of any biases inherent to the employed proxy network. This is especially important to emphasize because it raises the possibility that strong, natural low-frequency variability in the Pacific (e.g., Emile-Geay et al. 2013a, b; Ault et al. 2013a) arises solely as an artifact of reconstruction methodology, and does not reflect true variations in decadal to centennial variability in the climate system. This possibility needs to be investigated further because it would consequently provide more confidence in the performance of current LM climate model simulations and hence allow more confidence in model projections of climate change in the future.

Secondly, all of the methods tested herein yield regional and aggregate CFRs with means different from the target field, often in regions of considerable dynamical importance. The clear implication of these results is that regionally specific features of CFRs must to be interpreted carefully, especially with regard to their dynamic implications (e.g. Mann et al. 2009a, b; Rahmstorf et al. 2015). More explicitly, however, the impact of biases in CFRs was tested herein with regard to a prominent dynamical conclusion derived from the M09 CFR associated with a La Niña-like tropical Pacific during the MCA. For some models, we have demonstrated the potential for the M09 CFR methodology to alter the magnitude and pattern of temperature differences in the tropical Pacific within PPEs. These impacts can increase the likelihood that temperature differences between 300-year periods in the Niño3 index will be on the order of the differences observed in the actual M09 CFR, even in models that do not inherently simulate such differences. These differences are nevertheless very unlikely in the native model simulations, and still relatively unlikely in the biased CFRs derived from the M09 method. Our PPE observations are, however, cause for caution and indicate that

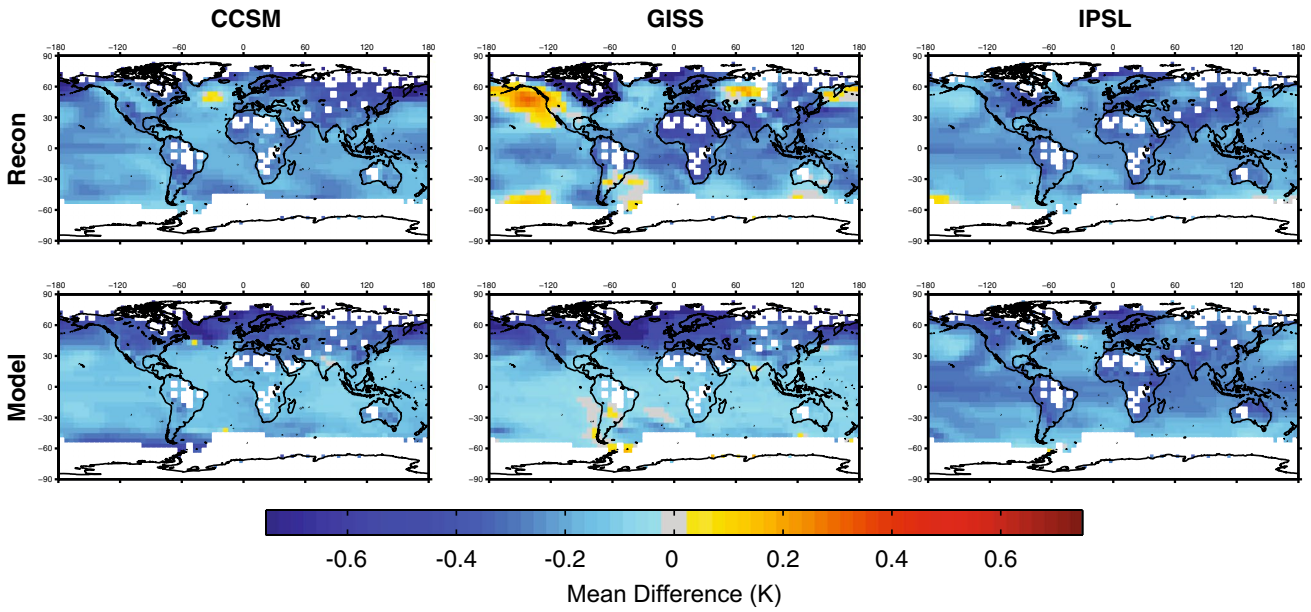

Fig. 12 Mean field temperature difference composite for periods in which the squared difference in the mean Niño3 SST index between two 300-year periods for the M09 pseudoproxy CFR is larger than the MCA-LIA squared difference (*dashed line*) in M09. The CFR composites are shown in the *top panels* and the true model composites are shown in the *bottom panels*

more uncertainty assessments are necessary before definitive dynamic statements can be derived from the current generation of global CFRs and their characterization of MCA-LIA differences.

Thirdly, it should be underscored that all of the results presented herein are based on PPE designs that are arguably best-case scenarios. We have used white noise perturbations to temperature-only pseudoproxies that span the entire reconstruction interval. We also have used a calibration interval that matches the period used by M09, but invariably would need to be shortened for most methods that reserve a portion of the instrumental data to calculate cross-validation statistics. PPEs that have adopted more complicated and realistic frameworks, such as multivariate pseudoproxies (Evans et al. 2014), heterogeneously declining sampling density in pseudoproxy networks (Wang et al. 2014), and pseudoproxy noise perturbations with frequency-dependent spectral densities (e.g. von Storch et al. 2004), have all shown additional skill reductions relative to the simpler PPE design used herein. One exception has been noted by Wang et al. (2014) who found that a few pseudoproxies with high SNRs, in an otherwise noisy network, could improve skill throughout much of the reconstructed field; the application of uniform SNRs within our pseudoproxy networks therefore may be one element of our design that is overly pessimistic. On balance, however, our PPE design likely lends itself to the most optimistic assessment of CFR performance. For instance, pseudoproxies constructed with noise perturbations that have frequency-dependent spectral densities, such as red noise models, are likely to worsen and complicate the patterns of spectral biases characterized in Figs. 4, 5 and 6. It therefore is imperative to further test our results, particularly in a multi-model context, for their dependence on these various PPE design choices, despite the fact that our findings, even under a best-case scenario, already indicate important spatial errors in CFRs derived with the evaluated methods.

In conclusion, we have shown the importance of multi-model evaluations of CFR spatial skill, which ultimately has implications for how PPEs are interpreted in real-world contexts. The now large and growing ensemble of multi-century (and longer) forced and control simulations from models with diverse dynamic characteristics allows for a wide range of PPE assessments of CFRs and their uncertainties. In particular, one approach will involve defining a regional or dynamically specific hypothesis about pre-instrumental climate based on a real-world CFR, the robustness of which could then be evaluated using multiple LM simulations in a PPE framework to assess the potential impact of model-based features and proxy characteristics. Important model features include teleconnection stationarity, forcing sensitivity, forcing uncertainty, the magnitudes of forced and internal variability, and the simulated expressions of atmospheric variability. When coupled with more traditional methodological evaluations and dependencies on proxy features such as inherent noise, network distributions and temporal density, the expanded collection of LM simulations will therefore enhance the ability of PPEs to test the impacts of dynamical uncertainties, and their connections to other reconstruction uncertainties, on hypotheses that arise from real-world CFRs.

Acknowledgments We are grateful for the helpful comments from the reviewer of our manuscript. Supported in part by NOAA grants NA10OAR4320137 and NA11OAR4310166. Supplementary data can be accessed at http://www.ldeo.columbia.edu/~jsmerdon/2015_clidyn_smerdonetal_supplement.html. LDEO contribution #7903.

References

Ammann CM, Joos F, Schimel DS, Otto-Bliesner BL, Tomas RA (2007) Solar influence on climate during the past millennium: results from transient simulations with the NCAR Climate System Model. Proc Nat Acad Sci USA 104:3713–3718

Anchukaitis KJ, Evans MN, Kaplan A, Vaganov EA, Hughes MK, Grissino-Mayer HD, Cane MA (2006) Forward modeling of regional-scale tree-ring patterns in the southeastern United States and the recent emergence of summer drought stress. Geophys Res Lett 33(4):L04705. doi:10.1029/2005GL025050

Anchukaitis KJ, Buckley BM, Cook ER, Cook BI, D'Arrigo RD, Ammann CM (2010) Influence of volcanic eruptions on the climate of the Asian monsoon region. Geophys Res Lett 37:L22703. doi:10.1029/2010GL044843

Anchukaitis KJ, D'Arrigo RD, Andreu-Hayles L, Frank D, Verstege A, Buckley BM, Curtis A, Jacoby GC, Cook ER (2013) Tree-ring reconstructed summer temperatures from northwestern North America during the last nine centuries. J Clim 26(10):3001–3012. doi:10.1175/JCLI-D-11-00139.1

Annan JD, Hargreaves JC (2012) Identification of climatic state with limited proxy data. Clim Past 8:1141–1151. doi:10.5194/cp-8-1141-2012

Ault TR, Deser C, Newman M, Emile-Geay J (2013a) Characterizing decadal to centennial variability in the equatorial Pacific during the last millennium. Geophys Res Lett 40:3450–3456. doi:10.1002/grl.50647

Ault TR, Cole JE, Overpeck JT, Pederson GT, St. George S, Otto-Bliesner B, Woodhouse CA, Deser C (2013b) The continuum of hydroclimate variability in Western North America during the last millennium. J Clim 26:5863–5878. doi:10.1175/JCLI-D-11-00732.1

Berdahl M, Robock A (2013) Northern Hemispheric cryosphere response to volcanic eruptions in the Paleoclimate Modeling Intercomparison Project 3 last millennium simulations. J Geophys Res Atmos 118:12359–12370. doi:10.1002/2013JD019914

Bothe O, Jungclaus JH, Zanchettin D (2013) Consistency of the multi-model CMIP5/PMIP3-past1000 ensemble. Clim. Past 9:2471–2487. doi:10.5194/cp-9-2471-2013

Bothe O, Evans M, Fernández Donado F, Garcia Bustamante E, Gergis J, Gonzalez-Rouco JF, Goosse H, Hegerl G, Hind A, Jungclaus J, Kaufman D, Lehner F, McKay N, Moberg A, Raible CC, Schurer A, Shi F, Smerdon JE, von Gunten L, Wagner S, Warren E, Widmann M, Yiou P, Zorita E (2015) Continental-scale temperature variability in PMIP3 simulations and PAGES 2k regional temperature reconstructions over the past millennium. Clim Past (in review)

Briffa K, Schweingruber F, Jones PD, Osborn T (1998) Reduced sensitivity of recent tree growth to temperature at high northern latitudes. Nature 391:678–682

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

Christiansen B, Ljungqvist FC (2012) The extra-tropical Northern Hemisphere temperature in the last two millennia: reconstructions of low-frequency variability. Clim Past 8:765–786. doi:10.5194/cp-8-765-2012

Christiansen B, Schmith T, Thejll P (2009) A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. J Clim 22(4):951–976

Christiansen B, Schmith T, Thejll P (2010) Reply. J Clim 23(10):2839–2844. doi:10.1175/2010JCLI3281.1

Coats S, Smerdon JE, Seager R, Cook BI, González-Rouco JF (2013a) Megadroughts in Southwestern North America in ECHO-G millennial simulations and their comparison to proxy drought reconstructions. J Clim 26:7635–7649. doi:10.1175/JCLI-D-12-00603.1

Coats S, Smerdon JE, Cook BI, Seager R (2013b) Stationarity of the tropical Pacific teleconnection to North America in CMIP5/PMIP3 model simulations. Geophys Res Lett 40:1–6. doi:10.1002/grl.50938

Coats S, Smerdon JE, Cook BI, Seager R (2015a) Are simulated megadroughts in the North American Southwest forced? J Clim 28:124–142. doi:10.1175/JCLI-D-14-00071.1

Coats S, Cook BI, Smerdon JE, Seager R (2015b) North American Pan-Continental droughts in model simulations of the last millennium. J Clim (in press)

Cook E, Krusic P (2004) The North American Drought Atlas. NOAA Paleoclimatology, Boulder

Cook ER, Seager R, Cane MA, Stahle DW (2007) North American drought: reconstructions, causes, and consequences. Earth Sci Rev 81(1–2):93–134

Cook ER, Anchukaitis KJ, Buckley BM, D'Arrigo RD, Jacoby GC, Wright WE (2010) Asian monsoon failure and megadrought during the last millennium. Science 328:486–489

Dannenberg MP, Wise EK (2013) Performance of climate field reconstruction methods over multiple seasons and climate variables. J Geophys Res Atmos 118:9595–9610. doi:10.1002/jgrd.50765

Emile-Geay J, Cobb KM, Mann ME, Wittenberg AT (2013a) Estimating central equatorial Pacific SST variability over the past millennium. Part I: methodology and validation. J Clim 26:2302–2328. doi:10.1175/JCLI-D-11-00510.1

Emile-Geay J, Cobb KM, Mann ME, Wittenberg AT (2013b) Estimating central equatorial Pacific SST variability over the past millennium. Part II: reconstructions and implications. J Clim 26:2329–2352. doi:10.1175/JCLI-D-11-00511.1

Esper J, Wilson RJS, Frank DC, Moberg A, Wanner A, Luterbacher J (2005) Climate: past ranges and future changes. Quat Sci Rev 24:2164–2166

Evans MN, Kaplan A, Cane MA (2002) Pacific sea surface temperature field reconstruction from coral δ18O data using reduced space objective analysis. Paleoceanography 17:1007. doi:10.1029/2000PA000590

Evans MN, Smerdon JE, Kaplan A, Tolwinski-Ward SE, Gonzalez-Rouco JF (2014) Climate field reconstruction uncertainty arising from multivariate and nonlinear properties of predictors. Geophys Res Lett 41(24):9127–9134. doi:10.1002/2014GL062063

Fernández-Donado L, González-Rouco JF, Raible CC, Ammann CM, Barriopedro D, García-Bustamante E, Jungclaus JH, Lorenz SJ, Luterbacher J, Phipps SJ, Servonnat J, Swingedouw D, Tett SFB, Wagner S, Yiou P, Zorita E (2013) Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium. Clim Past 9:393–421. doi:10.5194/cp-9-393-2013

Franke J, Frank D, Raible CC, Esper J, Brönnimann S (2013) Spectral biases in tree-ring climate proxies. Nat Clim Ch. 3:360–364

Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2):215–223

González-Rouco F, von Storch H, Zorita E (2003) Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. Geophys Res Lett 30(21):2116. doi:10.1029/2003GL018264

González-Rouco JF, Beltrami H, Zorita E, von Storch H (2006) Simulation and inversion of borehole temperature profiles in surrogate climates: spatial distribution and surface coupling. Geophys Res Lett 33:L01703

González-Rouco JF, Fernández-Donado L, Raible CC, Barriopedro D, Garcia-Herrera R, Luterbacher J, Jungclaus J, Swingedow D, Servonat J, Tett S, Brohan P, Zorita E, Wagner S, Ammann C (2011) Medieval climate anomaly to little ice age transition as simulated by current climate models. Pages Newsl 19(1):7–10

Goosse H, Crespin E, de Montety A, Mann ME, Renssen H, Timmermann A (2010) Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation. J Geophys Res 115:D09108. doi:10.1029/2009JD012737

Goosse H, Crespin E, Dubinkina S, Loutre M-F, Mann ME, Renssen H, Sallaz-Damaz Y, Shindell DT (2012) The role of forcing and internal dynamics in explaining the "Medieval Climate Anomaly". Clim Dyn 39:2847–2866

Guillot D, Rajaratnam B, Emile-Geay J (2015) Statistical paleoclimate reconstructions via Markov random fields. Ann Appl Stat (in press), arXiv:1309.6702

Hegerl GC, Crowley TJ, Allen M, Hyde WT, Pollack HN, Smerdon J, Zorita E (2007) Detection of human influence on a new, validated 1500-year climate reconstruction. J Clim 20:650–666

Herweijer C, Seager R, Cook ER (2007) North American droughts of the last millennium form a gridded network of tree-ring data. J Clim 20:1353–1376

Hind A, Moberg A (2013) Past millennial solar forcing magnitude. A statistical hemispheric-scale climate model versus proxy data comparison. Clim Dyn 41:2527–2537. doi:10.1007/s00382-012-1526-6

Hind A, Moberg A, Sundberg R (2012) Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium—Part 2: a pseudo-proxy study addressing the amplitude of solar forcing. Clim Past 8:1355–1365. doi:10.5194/cp-8-1355-2012

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12:55–67

Huybers P, Curry W (2006) Links between annual, Milankovitch, and continuum temperature variability. Nature 441:329–332. doi:10.1038/nature04745

Jacoby GC, D'Arrigo R (1995) Tree-ring width and density evidence of climatic and potential forest change in Alaska. Global Biogeochem Cycles 9:227–234

Lee TCK, Zwiers FW, Tsao M (2007) Evaluation of proxy-based millennial reconstruction methods. Clim Dyn 31:263–281

Lehner F, Raible CC, Stocker TF (2012) Testing the robustness of a precipitation proxy-based North Atlantic oscillation reconstruction. Quat Sci Rev 45:85–94

Lewis SC, LeGrande AN (2015) Stability of ENSO and its tropical Pacific teleconnections over the last millennium. Clim Past Discuss 11:1579–1613. doi:10.5194/cpd-11-1579-2015

Li B, Smerdon JE (2012) Defining spatial assessment metrics for evaluation of paleoclimatic field reconstructions of the Common Era. Environmetrics 23(5):394–406

Luterbacher J, Xoplaki E, Dietrich D, Rickli R, Jacobeit J, Beck C, Gyalistras D, Schmutz C, Wanner H (2002) Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500. Clim Dyn 18:545–561

Luterbacher J, Dietrich D, Xoplaki E, Grosjean M, Wanner H (2004) European seasonal and annual temperature variability, trends and extremes since 1550. Science 303:1499–1503

Mann ME, Bradley RS, Hughes MK (1998) Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392:779–787

Mann ME, Rutherford S, Wahl E, Ammann C (2005) Testing the fidelity of methods used in proxy-based reconstructions of past climate. J Clim 18:4097–4107

Mann ME, Rutherford S, Wahl E, Ammann C (2007) Robustness of proxy-based climate field reconstruction methods. J Geophys Res 112:D12109

Mann ME, Zhang Z, Hughes MK, Bradley RS, Miller SK, Rutherford S, Ni F (2008) Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proc Nat Acad Sci USA 105(36):13252–13257

Mann ME, Zhang Z, Rutherford S, Bradley RS, Hughes MK, Shindell D, Ammann C, Faluvegi G, Ni F (2009a) Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. Science 326(5957):1256–1260

Mann ME, Woodruff JD, Donnelly JP, Zhang Z (2009b) Atlantic hurricanes and climate over the past 1,500 years. Nature 460:880–883

Masson-Delmotte V, Schulz M, Abe-Ouchi A, Beer J, Ganopolski A, González Rouco JF, Jansen E, Lambeck K, Luterbacher J, Naish T, Osborn T, Otto-Bliesner B, Quinn T, Ramesh R, Rojas M, Shao X, Timmermann A (2013) Information from paleoclimate archives. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge and New York

Moberg A, Sonechkin D, Holmgren K, Datsenko N, Karlen W (2005) Highly variable Northern Hemisphere temperatures reconstructed from low-and high-resolution proxy data. Nature 433:613–617

Neukom R, Luterbacher J, Villalba R, KüttelM Frank D, Jones PD, Gosjean M, Esper J, Lopez L, Wanner H (2010) Multi-centennial summer and winter precipitation variability in southern South America. Geophys Res Lett 37:L14708. doi:10.1029/2010GL043680

Otto-Bliesner BL, Brady EC, Fasullo J, Jahn A, Landrum L, Stevenson S, Rosenbloom N, Mai A, Strand G (2015), Climate variability and change since 850 C.E.: an ensemble approach with the community earth system model, Bull Am Meteorol Soc, (in review)

PAGES 2k Consortium (2013) Continental-scale temperature variability over the last two millennia. Nat Geosci 6:339–346. doi:10.1038/ngeo1797

Pauling A, Luterbacher J, Wanner H (2003) Evaluation of proxies for European and North Atlantic temperature field reconstructions. Geophys Res Lett 30:1787. doi:10.1029/2003GL017589

Phipps SJ, McGregor HV, Gergis J, Gallant AJE, Neukom R, Stevenson S, Ackerley D, Brown JR, Fischer MJ, van Ommen TD (2013) Paleoclimate data-model comparison and the role of climate forcings over the past 1500 years. J Clim 26:6915–6936

Rahmstorf S, Box JE, Feulner G, Mann ME, Robinson A, Rutherford S, Schaffernicht EJ (2015) Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. Nature Clim. Ch. 5:475–480

Riedwyl N, Kuttel M, Luterbacher J, Wanner H (2009) Comparison of climate field reconstruction techniques: application to Europe. Clim Dyn 32:381–395

Rutherford S, Mann ME, Delworth TL, Stouffer RJ (2003) Climate field reconstruction under stationary and nonstationary forcing. J Clim 16:462–479

Rutherford S, Mann ME, Osborn TJ, Bradley RS, Briffa KR, Hughes MK, Jones PD (2005) Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to methodology, predictor network, target season, and target domain. J Clim 18:2308–2329

Rutherford SD, Mann ME, Ammann CM, Wahl ER (2010) Comments on: "A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness" by Christiansen, Schmith and Thejll. J Clim 23(10):2832–2838. doi:10.1175/2010JCLI3146.1

Schmidt GA, Jungclaus JH, Ammann CM, Bard E, Braconnot P, Crowley TJ, Delaygue G, Joos F, Krivova NA, Muscheler R, Otto-Bliesner BL, Pongratz J, Shindell DT, Solanki SK, Steinhilber F, Vieira LEA (2011) Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). Geosci Model Dev 4:33–45. doi:10.5194/gmd-4-33-2011

Schmidt GA, Annan JD, Bartlein PJ, Cook BI, Guilyardi E, Hargreaves JC, Harrison SP, Kageyama M, LeGrande AN, Konecky B, Lovejoy S, Mann ME, Masson-Delmotte V, Risi C, Thompson D, Timmermann A, Tremblay L-B, Yiou P (2014) Using palaeoclimate comparisons to constrain future projections in CMIP5. Clim Past 10:221–250. doi:10.5194/cp-10-221-2014

Schneider T (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J Clim 14:853–887

Seager R, Burgman R, Kushnir Y, Clement A, Cook ER, Naik N, Miller J (2008) Tropical Pacific forcing of North American medieval megadroughts: testing the Concept with an atmosphere model forced by coral-reconstructed SSTs. J Clim 21(23):6175–6190. doi:10.1175/2008JCLI2170.1

Smerdon JE (2012) Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. WIREs Clim Change 3:63–77. doi:10.1002/wcc.149

Smerdon JE, Kaplan A (2007) Comments on "Testing the fidelity of methods used in proxy-based reconstructions of past climate": the role of the standardization interval. J Clim 20:5666–5670

Smerdon JE, Kaplan A, Chang D (2008a) On the origin of the standardization sensitivity in RegEM climate field reconstructions. J Clim 21(24):6710–6723

Smerdon JE, González-Rouco JF, Zorita E (2008b) Comment on "Robustness of proxy-based climate field reconstruction methods" by Michael E. Mann et al. J Geophys Res 113:D18106. doi:10.1029/2007JD009542

Smerdon JE, Kaplan A, Chang D, Evans MN (2010a) A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. J Clim 23:4856–4880

Smerdon JE, Kaplan A, Amrhein DE (2010b) Erroneous model field representations in multiple pseudoproxy studies: corrections and Implications. J Clim 23:5548–5554

Smerdon JE, Kaplan A, Zorita E, González-Rouco JF, Evans MN (2011) Spatial performance of four climate field reconstruction methods targeting the Common Era. Geophys Res Lett 38:L11705. doi:10.1029/2011GL047372

Steiger NJ, Hakim GJ, Steig EJ, Battisti DS, Roe GH (2014) Assimilation of time-averaged pseudoproxies for climate reconstruction. J Clim 27:426–441

St. George S, Meko DM, Cook ER (2010) The seasonality of precipitation signals embedded within the North American Drought Atlas. Holocene 20:983–988

Sundberg R, Moberg A, Hind A (2012) Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium—Part 1: theory. Clim Past 8:1339–1353. doi:10.5194/cp-8-1339-2012

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498. doi:10.1175/BAMS-D-11-00094.1

Tingley MP, Huybers P (2010) A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: development and applications to paleoclimate reconstruction problems. J Clim 23:2759–2781

Tingley MP, Craigmile PF, Haran M, Li B, Mannshardt-Shamseldin E, Rajaratnam B (2012) Piecing togetherthe past: statistical insights into paleoclimatic reconstructions. Quart Sci Rev 35:1–22

von Storch H, Zorita E, Jones JM, Dimitriev Y, González-Rouco F, Tett SFB (2004) Reconstructing past climate from noisy data. Science 306:679–682

von Storch H, Zorita E, Jones JM, González-Rouco F, Tett SFB (2006) Response to comment on "Reconstructing past climate from noisy data,". Science 312:529c

Wang J, Emile-Geay J, Guillot D, Smerdon JE, Rajaratnam B (2014) Evaluating climate field reconstruction techniques using improved emulations of real-world conditions. Clim Past 10:1–19. doi:10.5194/cp-10-1-2014

Werner JP, Luterbacher J, Smerdon JE (2013) A pseudoproxy evaluation of bayesian hierarchical modeling and canonical correlation analysis for climate field reconstructions over Europe. J Clim 26(3):851–867