Clim Dyn
DOI 10.1007/s00382-007-0351-9

Evaluation of proxy-based millennial reconstruction methods

Terry C. K. Lee · Francis W. Zwiers · Min Tsao

Received: 3 July 2007 / Accepted: 9 November 2007

Abstract A range of existing statistical approaches for reconstructing historical temperature variations from proxy data are compared using both climate model data and real-world paleoclimate proxy data. We also propose a new method for reconstruction that is based on a state-space time series model and Kalman filter algorithm. The state-space modelling approach and the recently developed RegEM method generally perform better than their competitors when reconstructing interannual variations in Northern Hemispheric mean surface air temperature. On the other hand, a variety of methods are seen to perform well when reconstructing surface air temperature variability on decadal time scales. An advantage of the new method is that it can incorporate additional, non-temperature, information into the reconstruction, such as the estimated response to external forcing, thereby permitting a simultaneous reconstruction and detection analysis as well as future projection. An application of these extensions is also demonstrated in the paper.

Keywords Kalman filter · State-space model · Temperature reconstruction

T. C. K. Lee · M. Tsao
Department of Mathematics and Statistics,
University of Victoria, Victoria, BC, Canada

F. W. Zwiers (✉)
Climate Research Division, Environment Canada,
4905 Dufferin Street, Toronto, ON M3H 5T4, Canada
e-mail: Francis.Zwiers@ec.gc.ca

1 Introduction

A number of studies have attempted to reconstruct hemispheric mean temperature for the past millennium from proxy climate indicators. However, the available reconstructions vary considerably. Different statistical methods are employed in these reconstructions and it therefore seems natural to ask how much of this discrepancy is caused by the variations in methods. The reliability of some of the reconstruction methods has been looked at in different studies (e.g. Zortia et al. 2003; von Storch et al. 2004; Bürger and Cubasch 2005; Esper et al. 2005; Mann et al. 2005; Zorita and von Storch 2005; Bürger et al. 2006; Juckes et al. 2006).

In this paper, we provide an empirical comparison between different reconstruction methods. Analyses are carried out using both pseudo-proxy data from climate models and real-world paleoclimate proxy data. We will also propose a method for reconstruction that is based on a state-space time series model and Kalman filter algorithm. As an option, this method allows one to simultaneously incorporate historical forcing information to reconstruct the unknown historical temperature, and to carry out detection analysis to quantify the influence of external forcing on historical temperature.

The remainder of this paper is organized as follows. The existing statistical procedures used for reconstruction are discussed in Sect. 2, together with the approach that utilizes a state-space time series model and Kalman filter algorithm. Following the general approach proposed by von Storch et al. (2004), comparisons between these methods are made with the help of both climate model data and real-world paleoclimate proxy data. The results are presented in Sect. 3. Concluding remarks are given in Sect. 4.

2 Reconstruction approaches

2.1 Existing approaches

All reconstruction approaches, in general, involve statistical procedures that map the proxy series onto the reconstructed temperature. The mapping involves first finding the relationship between the proxy data and the instrumental record during a calibration period when both records are available. This relationship is then applied to the proxy data to provide the reconstructed historical temperature.

Mann et al. (2005) classified the existing reconstruction approaches into two major categories: composite plus scale (CPS) methods and climate field reconstruction (CFR) methods. In the CPS approach, a number of proxy series is first combined together to form a composite record, which is then used to reconstruct the temporal evolution of mean temperature over some spatial domain, typically the hemispheric mean. In many studies, the reconstruction target is the northern hemispheric mean temperature because proxy data, derived mainly from trees, are mostly available from the northern hemisphere land masses. The composite record is typically formed by calculating a weighted average of all the available proxy series. The weights can either be uniform (e.g. Jones et al. 1998; Briffa et al. 2001; Esper et al. 2002) or can be determined using the correlation between the proxy series and the instrumental record during the calibration period (e.g. Hegerl et al. 2007). The correlation based weighting scheme has the advantage of minimizing the influence of potentially unreliable proxy series on the composite record. One recent study by Moberg et al. (2005) used a different averaging scheme, in which high-frequency and low-frequency composites were first formed individually using wavelet transformation. Each of these composites was formed using only proxy indicators that were thought to be able to capture variability in the corresponding frequency range. The two composites are then combined to form a single composite. This averaging scheme can be beneficial when the individual proxy series are known to capture variability at different timescales.

Once the composite is formed, it is then calibrated to produce the reconstructed temperature. One calibration approach, the forward regression approach, utilizes the ordinary least squares regression model which assumes that, at time t, for $t = n + 1, n + 2, \ldots, n + m$,

$$\mathbf{T}_t = \alpha \mathbf{P}_t + \varepsilon_t \tag{1}$$

where $n + m$ and m are the length of the composite proxy record and instrumental record respectively (m is in general much smaller than n), $\mathbf{T}$ is the mean hemispheric instrumental record, $\mathbf{P}$ is the composite proxy record and ε_t represents error that results from incomplete spatial sampling in the instrumental record. The coefficient α scales $\mathbf{P}_t$ to $\mathbf{T}_t$ and is estimated by ordinary least squares regression which minimizes the residual sum of squares between $\alpha\mathbf{P}_t$ and $\mathbf{T}_t$ during the calibration period. By assuming that Eq. (1) also holds at times prior to the calibration period, the unknown historical hemispheric mean temperature can then be reconstructed by scaling the pre-calibration period composite record $\mathbf{P}_t$ by α, for $t = 1, 2, \ldots, n$. The estimate of α, denoted by $\hat{\alpha}$, is in general negatively biased because of the measurement error and non-temperature variability inherit in the composite proxy series (see, e.g. Fuller 1987; Allen and Stott 2003). This under-estimation will result in a loss of variance in the reconstructed series because $\mathrm{Var}(\hat{\alpha}\mathbf{P}) < \mathrm{Var}(\alpha\mathbf{P})$ for $0 < \hat{\alpha} < \alpha$.

One way to avoid this underestimation in α is to use the total least squares (TLS) method to estimate α (see Allen and Stott 2003 for an application of this technique in climate change detection analysis). In this case the statistical relationship between $\mathbf{P}_t$ and $\mathbf{T}_t$ is defined by

$$\mathbf{T}_t = \alpha(\mathbf{P}_t - \eta_t) + \varepsilon_t \tag{2}$$

where the additional term η_t represents non-temperature variability in the composite proxy series. By explicitly incorporating η_t into the regression model, the underlying value of α, in theory, can be better estimated. Hegerl et al. (2007) applied the total least squares method in their reconstruction. To derive the best guess estimate of α, one needs to know the ratio of $\mathrm{Var}(\eta_t)$ to $\mathrm{Var}(\varepsilon_t)$. Since this ratio is unknown in real-world applications, it is estimated using climate model simulations. Hegerl et al. (2007) find that their reconstruction is insensitive to the precise choice of the $\mathrm{Var}(\eta_t)$ to $\mathrm{Var}(\varepsilon_t)$ ratio.

Another CPS reconstruction method, which is termed the variance matching approach, is favored by Jones et al. (1998) and others. In this approach, the pre-calibration period $\mathbf{P}_t$ is scaled by a parameter β, where β is determined so that, during the calibration period, $\mathrm{Var}(\mathbf{T}) = \mathrm{Var}(\beta\ \mathbf{P})$.

A variant of the forward regression model (Eq. 1) is the inverse regression model (see Brown 1993, for an introduction; Coehlo et al. 2004, for an example)

$$\mathbf{P}_t = \zeta \mathbf{T}_t + \eta_t \tag{3}$$

where η_t is defined similarly as in Eq. (2) and ζ is estimated by minimizing the residual sum of squares between $\zeta\mathbf{T}_t$ and $\mathbf{P}_t$ during the calibration period. The reconstructed hemispheric temperature is then estimated by $\mathbf{P}_t/\zeta$. Different implementations of the inverse regression method have been used in the paleoclimate reconstruction literature. In Mann et al. (1998), the regression is done between the principal components of the instrumental record and the individual proxy indictor (see latter part of this section for details), some of which were obtained by principle

component analysis from a set of proxy series. In Juckes et al. (2006), the inverse regression is done between each individual proxy series and the Northern Hemisphere annual mean temperature. A weighted average of the individual proxy series, weights determined by the regression coefficients, is then used as the reconstructed series.

Note that the instrumental record $\mathbf{T}_t$ is subject to sampling uncertainty (see Jones et al. 1997 for an estimate of this uncertainty) and neglecting such uncertainty when estimating ζ in inverse regression will result in an estimate that is negatively biased. However, such bias would be smaller than the bias incurred by using forward regression because the non-temperature variability inherent in the composite record is usually larger than the sampling uncertainty in the instrumental record. On the other hand, the total least squares approach used by Hegerl et al. (2007), in theory, should provide a more accurate estimate of the regression coefficient when both $\mathbf{P}_t$ and $\mathbf{T}_t$ are noise contaminated. However, this only holds if the ratio of $\mathrm{Var}(\eta_t)$ to $\mathrm{Var}(\varepsilon_t)$, which is required to derive the estimate of α, is known. This ratio is unknown in real-world applications and is estimated using a limited number of climate model simulations. The bias from such estimation is hard to determine and thus its impact on the reconstruction is unclear. For this reason, it is not obvious whether inverse regression or total least squares regression will result in smaller bias.

In the case of the CFR approach, the proxy series are used to reconstruct both the underlying temporal and spatial patterns of historical temperature. The Mann et al. (1998) method (often referred to as the MBH method) is an example of a technique that uses the climate field reconstruction approach. The MBH method brings together techniques used in principal component analysis and regression. The instrumental record is first decomposed into its spatial and temporal parts through principal component analysis and only a subset of the components are retained. Next, the relationship between the subset of temporal principal components (PCs) and the ith proxy series ($i = 1, 2, ..., p$) during the calibration period is obtained by inverse regression which estimates the coefficients $\beta_j^{(i)}$ ($j = 1, 2, ..., N_{\mathrm{eof}}$) in

$$\begin{bmatrix} \mathbf{P}_{n+1}^{(i)} \\ \mathbf{P}_{n+2}^{(i)} \\ \vdots \\ \vdots \\ \mathbf{P}_{n+m}^{(i)} \end{bmatrix} = \mathbf{U} \begin{bmatrix} \beta_1^{(i)} \\ \beta_2^{(i)} \\ \vdots \\ \beta_{N_{eof}}^{(i)} \end{bmatrix} + \delta^{(i)}$$

where N_{eof} is the number of EOFs retained, $\mathbf{P}_t^{(i)}$ is the ith proxy data at time t , $\mathbf{U}$ is the matrix of temporal PCs ($m \times N_{eof}$) of the instrumental record and $\delta^{(i)}$ is a noise series ($m \times 1$). This procedure is repeated for each proxy series and this yields a matrix of coefficients, denoted by $\mathbf{G}$ ($p \times N_{eof}$), which is defined as,

$$\mathbf{G} = \begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \cdots & \beta_{N_{eof}}^{(1)} \\ \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_{N_{eof}}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1^{(p)} & \beta_2^{(p)} & \cdots & \beta_{N_{eof}}^{(p)} \end{bmatrix}$$

The pre-calibration period temporal PCs at time t, denoted by $\mathbf{Z}_t$ ($N_{eof} \times 1$), are then reconstructed through least squares regression using

$$\begin{bmatrix} \mathbf{P}_t^{(1)} \\ \mathbf{P}_t^{(2)} \\ \vdots \\ \mathbf{P}_t^{(p)} \end{bmatrix} = \mathbf{G}\mathbf{Z}_t + \kappa_t$$

where κ_t is a noise series ($p \times 1$). The sequence of reconstructed temporal PCs, $\hat{\mathbf{Z}}_t$, is then scaled to have the same variance as the instrumental temporal PCs over the calibration period and subsequently re-combined with the calibration period spatial PCs to provide an estimate of the unknown local temperature. A hemispheric mean reconstruction is then formed by the appropriate spatial average of the reconstructed local temperatures. The regression procedures above are both inverse regression. However, unlike the CPS inverse regression, the individual proxy series are used in the regression process. Also, there are N_{eof} coefficients to estimate in each regression, compared to only one coefficient in the CPS case.

Another CFR method uses the regularized expectation maximization (RegEM) algorithm (Schneider 2001) and is advocated by Rutherford et al. (2005) and others. Let $\mathbf{T}_t^{(j)}$ be the local temperature at time t at the jth location ($j = 1, 2, ..., q$) and let $\mathbf{X}$ be a $(n + m) \times (q + p)$ matrix with missing data which is defined as

$$\mathbf{X} = \begin{bmatrix} \mathbf{T}_1^{(1)} & \mathbf{T}_1^{(2)} & \cdots & \mathbf{T}_1^{(q)} & \mathbf{P}_1^{(1)} & \mathbf{P}_1^{(2)} & \cdots & \mathbf{P}_1^{(p)} \\ \mathbf{T}_2^{(1)} & \mathbf{T}_2^{(2)} & \cdots & \mathbf{T}_2^{(q)} & \mathbf{P}_2^{(1)} & \mathbf{P}_2^{(2)} & \cdots & \mathbf{P}_2^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{T}_{n+m}^{(1)} & \mathbf{T}_{n+m}^{(2)} & \cdots & \mathbf{T}_{n+m}^{(q)} & \mathbf{P}_{n+m}^{(1)} & \mathbf{P}_{n+m}^{(2)} & \cdots & \mathbf{P}_{n+m}^{(p)} \end{bmatrix}$$

$$\stackrel{def}{=} \begin{bmatrix} \mathbf{T}_{(1)} & \mathbf{P}_{(1)} \\ \mathbf{T}_{(2)} & \mathbf{P}_{(2)} \\ \vdots & \vdots \\ \mathbf{T}_{(n+m)} & \mathbf{P}_{(n+m)} \end{bmatrix}$$

with

$$\mathrm{E}[\mathbf{T}_{(t)} \quad \mathbf{P}_{(t)}] = \mu = [\mu_T \quad \mu_P], \quad \text{for } t = 1, 2, \ldots, n+m$$
$$\mathrm{Var}(\mathbf{X}) = \Sigma = \begin{bmatrix} \Sigma_{TT} & \Sigma_{TP} \\ \Sigma_{PT} & \Sigma_{PP} \end{bmatrix}$$

where μ_T and μ_P have length q and p, respectively, $\mathbf{T}_{(t)} = [\mathbf{T}_t^{(1)} \quad \mathbf{T}_t^{(2)} \ \ldots \ \mathbf{T}_t^{(q)}]$ and $\mathbf{P}_{(t)} = [\mathbf{P}_t^{(1)} \quad \mathbf{P}_t^{(2)} \ \ldots \ \mathbf{P}_t^{(p)}]$. The $(q + p) \times (q + p)$ matrix Σ is partitioned according to $\mathbf{T}$ and $\mathbf{P}$ so that Σ_{TT}, Σ_{PP} and $\Sigma_{TP} = \Sigma_{PT}^{\mathrm{T}}$ are $q \times q$, $p \times p$ and $q \times p$ matrices, respectively. In the matrix $\mathbf{X}$, $\mathbf{T}_{(t)}$ is unknown for $t = 1, 2, \ldots, n$. It is assumed that each record $\mathbf{T}_{(t)}$ can be represented by a linear model of the form

$$\mathbf{T}_{(t)} = \mu_T + (\mathbf{P}_{(t)} - \mu_P)\mathbf{B} + \epsilon_{(t)}.$$

Here, $\varepsilon_{(t)}$ is the residual and $\mathbf{B}$ is a $p \times q$ matrix of regression coefficients. The reconstructed temperature $\hat{\mathbf{T}}_{(s)}$ at time s ($s = 1,2, \ldots, n$), is thus defined as $\mu_T + (\mathbf{P}_{(s)} - \mu_P)\,\mathbf{B}$.

To calculate $\hat{\mathbf{T}}_{(s)}$, estimates of $\mathbf{B}$ and μ are required. If $n + m \geq p + 1$, the expectation maximization (EM) algorithm provides a way to iteratively estimate these parameters (see, e.g. Schneider 2001). First, initialize the unknown $\hat{\mathbf{T}}_{(t)}$ with some values and obtain an estimate of the mean and covariance of matrix $\mathbf{X}$ (see Schneider 2001, for the formulas). Then, estimate $\mathbf{B}$ using the maximum likelihood estimate $\hat{\mathbf{B}} = \hat{\Sigma}_{PP}^{-1}\hat{\Sigma}_{PT}$, where the $\hat{\Sigma}_{PP}$ and $\hat{\Sigma}_{PT}$ denote the partitioned covariance matrix estimate. If $n + m < p + 1$, the estimate of Σ_{PP} is singular and the coefficient $\mathbf{B}$ is not defined. Next, impute the missing $\hat{\mathbf{T}}_{(t)}$ by $\hat{\mu}_T + (\mathbf{P}_{(t)} - \hat{\mu}_P)\hat{\mathbf{B}}$ and update the $\mathbf{X}$ matrix. Re-estimate μ, Σ and $\mathbf{B}$ using the updated matrix and then recalculate the missing temperatures. This process is continued until the change in imputed values become sufficiently small.

In a typical CFR application, $n + m$ is greater than $p + 1$. However, the estimate of Σ is rank deficient because $n + m < p + q + 1$, which can result in a poor estimate of the coefficient $\mathbf{B}$. Hence, the RegEM algorithm is used instead. The RegEM algorithm consists of the same steps as the EM algorithm, except the maximum likelihood estimate $\hat{\mathbf{B}}$ is replaced with a *regularized* estimate, which is obtained by a regularized regression procedure. Different regularization scheme have been used. In Rutherford et al. (2005) and Mann et al. (2005), the ridge regression scheme is used to estimate the coefficient matrix $\mathbf{B}$. Readers are referred to Schneider (2001) for more details of the ridge regression scheme. In Mann et al. (2007), truncated total least squares (TTLS) (Fierro et al. 1997) is used to estimate $\mathbf{B}$.

2.2 State-space model and Kalman filter algorithm

In addition to the methods described above, we propose a new method to be used for reconstructing historical temperature that is based on a state-space time series model and the Kalman filter algorithm. (Kalman 1960; see Harvey 1989 or Durbin and Koopman 2001, for an introduction). The proposed technique is essentially a further variant on inverse regression. A *state-space time series model* is a system that consists of two equations: the *observation equation* and the *state equation*. In the context of paleoclimate reconstruction, the observation equation describes the relationship between the proxy series and the unknown hemispheric mean temperature. On the other hand, the state equation models the dynamics of the unknown hemispheric mean temperature. Based on the state-space model, one can estimate the unknown hemispheric mean temperature using the *Kalman filter algorithm*, a Bayesian updating scheme that estimates the unobserved hemispheric mean temperature based on the proxy data. The estimates from the Kalman filter algorithm are the best linear estimator of the unobserved process in the sense of minimum mean square error. The Kalman filter algorithm can also be extended to provide a forecast of the hemispheric mean temperature.

To describe the state-space representation of the hemispheric mean temperature, we must first introduce some notation and make certain assumptions. Thus, let $\mathbf{GS}_t$, $\mathbf{VOL}_t$ and $\mathbf{SOL}_t$ be the climate model estimated responses to greenhouse gas and sulphate aerosol forcing combined (GS), volcanic forcing (VOL) and solar forcing (SOL) at time t, respectively. We can then think of the hemispheric mean temperature $\mathbf{T}_t$ as being the sum of the response to external forcing plus the effect of internal variability. With these assumptions, a reasonable statistical model for $\mathbf{T}_t$ ($t = 1, 2, \ldots, n + m$) might be

$$\begin{aligned} \mathbf{T}_t &= \tau + \delta_{\mathbf{GS}}\mathbf{GS}_t + \delta_{\mathbf{VOL}}\mathbf{VOL}_t + \delta_{\mathbf{SOL}}\mathbf{SOL}_t + \mathbf{Z}_t \\ &= \Upsilon\mathbf{X}_t + \mathbf{Z}_t \end{aligned} \tag{4}$$

where $\Upsilon = [\tau \quad \delta_{\mathbf{GS}} \quad \delta_{\mathbf{VOL}} \quad \delta_{\mathbf{SOL}}]$, $\mathbf{X}_t = [1 \quad \mathbf{GS}_t \quad \mathbf{VOL}_t \quad \mathbf{SOL}_t]^{\mathrm{T}}$, τ is the mean state of the climate system, $\delta_{\mathbf{GS}}$,$\delta_{\mathbf{VOL}}$ and $\delta_{\mathbf{SOL}}$ are scaling factors that account for error in the magnitude of the estimated response to the specified external forcing and $\mathbf{Z}_t$ represents random variations resulting from internal variability. We further assumed that $\mathbf{Z}_t$ is an AR(1) process with lag-one autocorrelation ϕ. Thus,

$$\mathbf{Z}_t = \sum_{j=0}^{\infty} \phi^j v_{t-j}, \tag{5}$$

where the v_t's are white noise with mean 0 and variance Q.

With these assumptions, our state-space representation of the hemispheric mean temperature $\mathbf{T}_t$ ($t = 1, 2, \ldots, n + m$) is given by the following system of equations:

$$\begin{aligned} \mathbf{T}_t &= \phi\mathbf{T}_{t-1} + \Upsilon\mathbf{F}_t + v_t \\ \mathbf{P}_t &= \zeta\mathbf{T}_t + \eta_t \end{aligned} \tag{6}$$

where $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$, η_t is white noise with mean zero and variance R. The first equation is the state equation

which determines how the unknown hemispheric mean temperature evolves over time. It is obtained by considering the difference between $\mathbf{T}_t$ and $\phi\mathbf{T}_{t-1}$ using Eq. (4). This equation assumed that $\mathbf{T}_t$ follows an autoregressive process of order 1, i.e. AR(1), with exogenous variables $\mathbf{F}_t$ and additive white noise v_t. The state equation in effect states that the rate of change in temperature depends upon the response to forcing which is governed by the forcing coefficients δ_{GS}, δ_{VOL} and δ_{SOL}, and upon natural internal variability. Alternatively, one can also define a state equation that does not contain the response to forcings. This can be achieved by setting $\mathbf{X}_t = 1$ and $\Upsilon = \tau$. The second equation in Eq. (6), which is the same equation as used in inverse regression, is the observation equation. This equation simply assumes that the relationship between the composite proxy series and hemispheric temperature that holds during the calibration period in Eq. (3) also holds in the pre-calibration period.

The problem of estimating the pre-calibration period $\mathbf{T}_t$ in a state-space model can be approached by using the Kalman filter and smoother algorithm. For notational purposes, let $\mathbf{T}_t^s$ represent the estimate of $\mathbf{T}$ at time t given $\mathbf{P}_1$, $\mathbf{P}_2$, ..., $\mathbf{P}_s$ and $\mathbf{F}_1$, $\mathbf{F}_2$, ..., $\mathbf{F}_s$, where $s \le t$. The Kalman filter estimates of $\mathbf{T}_t$ for $t = 1, 2, ..., n + m$, denoted by $\mathbf{T}_t^t$, are defined by the following set of recursive equations (Kalman 1960; see also Shumway and Stoffer 2000, p. 313),

$$\begin{aligned} \mathbf{T}_t^{t-1} &= \phi\mathbf{T}_{t-1}^{t-1} + \Upsilon\mathbf{F}_t \\ \mathbf{T}_t^t &= \mathbf{T}_t^{t-1} + K_t(\mathbf{P}_t - \zeta\mathbf{T}_t^{t-1}) \end{aligned} \tag{7}$$

where

$$\begin{aligned} K_t &= \zeta S_t^{t-1}(\zeta^2 S_t^{t-1} + R)^{-1} \\ S_t^{t-1} &= \phi^2 S_{t-1}^{t-1} + Q \\ S_t^t &= S_t^{t-1}(1 - \zeta K_t). \end{aligned}$$

Appropriate initial conditions for this recursion are $\mathbf{T}_0^0 = \mu$ and $S_0^0 = \Sigma$, where μ and Σ are constants. The Kalman prediction $\mathbf{T}_t^{t-1}$ is an estimate of $\mathbf{T}_t$ based on all the information we have at time $t - 1$, that is, our prior knowledge of $\mathbf{T}_t$ before observing $\mathbf{P}_t$. Such prior knowledge is expressed through our Kalman filter estimate $\mathbf{T}_{t-1}^{t-1}$. After observing $\mathbf{P}_t$, our knowledge of $\mathbf{T}_t$ is updated by calculating the Kalman filter estimate, $\mathbf{T}_t^t$. Once all of the Kalman filter estimates are found, we can then further update our estimates of $\mathbf{T}_t$ based on the entire data set $\{\mathbf{P}_t, \mathbf{F}_t; t = 1, 2, ..., n\}$ using the Kalman smoother algorithm. It should be noted that even though $\mathbf{T}_t$ is known for $t = n + 1, ..., n + m$, we will still provide filter estimates for them. This will enable us to utilize all of the available data to estimate the unknown $\mathbf{T}_t$ when using the Kalman smoother. With initial condition $\mathbf{T}_{n+m}^{n+m}$ obtained from the Kalman filter algorithm (Eq. 7), the Kalman smoother estimates $\mathbf{T}_t^{n+m}$ of $\mathbf{T}_t$, for $t = n + m - 1$, $n + m - 2, ..., 0$, are

$$\mathbf{T}_t^{n+m} = \mathbf{T}_t^t + J_t(\mathbf{T}_{t+1}^{n+m} - \mathbf{T}_{t+1}^t), \tag{8}$$

where $J_t = \phi S_t^t / S_{t+1}^t$.

A difficulty in using the state-space time series model is that the parameters, $\{\zeta, R, \phi, \Upsilon, Q, \mu, \Sigma\}$, are unknown and need to be estimated. An optimization algorithm such as Newton–Raphson scoring and EM algorithm can be used to estimate them numerically (see Appendix for details; see Shumway and Stoffer 2000 for examples). Note that in the optimization algorithm for state-space model, it is assumed that the exogenous variable $\mathbf{F}_t$ is known. If ϕ is unknown, the exogenous variables therefore cannot contain the parameter ϕ. Thus, in our state-space representation of the hemispheric mean temperature, one needs to fix the parameter ϕ that is involved in $\mathbf{F}_t = \mathbf{X}_t - \phi\mathbf{X}_{t-1}$ before estimating the other parameters. More details on the choice of ϕ are given in the next section. Note that there is no need to fix the parameter ϕ in front of $\mathbf{T}_{t-1}$ in Eq. (6) in order to use the optimization algorithm. Thus, this parameter can be estimated together with the other parameters.

The state-space model approach can also be extended to produce forecasts. Provided that $\mathbf{F}_t$ is known for the future time period, one can use the state equation to generate a forecast recursively. Recall that the state equation at time t is given by $\mathbf{T}_t = \phi\mathbf{T}_{t-1} + \Upsilon\mathbf{F}_t + v_t$. By forecasting the error term v_t as zero, the forecast of $\mathbf{T}_{n+m+s}$ ($s = 1, 2, ...$) can be given by

$$\tilde{\mathbf{T}}_{n+m+s} = \phi\tilde{\mathbf{T}}_{n+m+s-1} + \Upsilon\mathbf{F}_{n+m+s} \tag{9}$$

with $\tilde{\mathbf{T}}_{n+m} = \mathbf{T}_{n+m}$ being the known instrumental record at time $n + m$. Forecasts can also be made in the case when the response to external forcing is not included in Eq. (6), i.e. when $\mathbf{X}_t = 1$ and $\Upsilon = \tau$. However, such forecasts will likely not be very useful because they do not take the impact of forcings into account and thus quickly revert to a forecast of the mean τ.

An advantage of using the state-space model in Eq. (6) is that it provides the flexibility to incorporate forcing response information into the estimation of the unknown temperature. This is achieved in a two step process that first determines the impact of each forcing on the unknown temperature through the estimation of the forcing coefficients. The estimation process uses the information available in both the proxy series and the calibration data. From Eq. (6), it is obvious that if the forcing coefficient is significantly different from zero, one can claim that the corresponding forcing has a significant impact on the hemispheric mean temperature. Such an assessment can be made using the confidence bound of each coefficient obtained during the estimation step (see the Appendix for details). Once the coefficients are estimated, the forcing information is then incorporated into the final estimate of the unknown temperature using the Kalman filter and

smoother algorithm. In another words, the use of the state-space model allows one to simultaneously reconstruct the unknown hemispheric mean temperature and conduct a detection assessment of the importance of the response to GS, VOL and SOL forcing on hemispheric temperature. Furthermore, the use of the state-space model allows one to provide projections of future climate which are based on parameters that are estimated using the proxy records and past observations.

3 Results

3.1 Analysis with climate model simulations

The performance of a particular reconstruction method depends on many factors, such as the statistical methods used, the choice of proxies, the quality of the proxy records and the target season or latitude band. It is, in general, difficult to compare the performance of different reconstruction methods using the past instrumental temperature record because the instrumental period is simply too short to calibrate such techniques and reliably assess their performance. Climate model simulations, however, can provide a test bed for assessing the reliability of these methods as first introduced by Zorita et al. (2003; see also von Storch et al. 2004; Mann et al. 2005; Zorita and von Storch 2005 and others). In this paper, we will follow the methodology proposed by von Storch et al. (2004). The idea is to generate pseudo-proxy records by sampling a selection of simulated grid-box temperatures from the climate model and degrading them with additive noise. The reconstruction method is then applied to these pseudo-proxy records and the resulting reconstruction is validated against the known simulated hemispheric mean temperature record. To reflect the differences in the quality and properties of real-world proxy data, different colours of noise (e.g. red rather than white) and varying amplitudes of the noise variance have been used in different studies.

In this section, we followed these procedures for testing the reconstruction methods that are described in Sect. 2. We have conducted a suite of experiments that explore the sensitivity of each method to (1) the climate model simulation used, (2) the length of the calibration period and (3) the amount of noise introduced into the pseudo-proxy series. The two simulations used are the GKSS ECHO-G simulation (von Storch et al. 2004) and a simulation from the NCAR CSM 1.4 model (Ammann et al. 2007). Both simulations were forced with reconstructions of solar, volcanic and greenhouse gas forcing. The CSM 1.4 run was also forced with a reconstruction of aerosol forcing over the millennium. We used the ECHO-G simulation in its original form, rather than as adjusted (Osborn et al. 2006). For both simulations, only output between years 1000 and 1990 is used. To represent the varying calibration intervals that were used in actual reconstructions, we have used two calibration periods in our analysis (1880–1960 and 1860–1970). These calibration periods are similar to that used in Hegerl et al. (2007) and Moberg et al. (2005). In our experiments, pseudo-proxy records were formed by degrading the grid box temperatures with additive noise. The amount of noise introduced into the pseudo-proxy series is expressed in terms of the signal-to-noise ratio (SNR), which is defined as $\sqrt{\mathrm{Var}(X)/\mathrm{Var}(N)}$, where X is the grid box temperature series and N is the additive noise series. For all methods, experiments with SNR = 0.5 and 1 were performed.

First, we obtain a pseudo northern hemisphere instrumental record from the simulation. CFR methods use a set of continuous grid-box temperatures for analysis; we therefore fixed the spatial coverage of the entire pseudo-instrumental record at that of the instrumental network in 1920, as represented in the HadCRUT3 data set (Brohan et al. 2006). The choice of the year 1920 is arbitrary, but it corresponds roughly to the mid points of the calibration periods that were used in our analysis. For the other methods, the pseudo NH mean instrumental record is calculated from the appropriate areal average of the same grid-box temperatures used above. This ensures that all methods are provided with the same information for calibration.

Next, we define two pseudo-proxy networks of 15 and 100 randomly selected model grid boxes. These networks were sampled from the 321 NH grid boxes that are co-located with actual tree ring data found in the International Tree Ring Data Base (http://www.ncdc.noaa.gov/paleo/treering.html). These networks of grid-box temperature were converted to pseudo-proxy records by degrading the grid-box temperatures with added noise to mimic the measurement error that is inherent in the proxy records. The resulting pseudo-proxy series were then standardized relative to the calibration period. Since the RegEM method is computationally intensive, we will only apply this method to the larger network of the two. The size of this network is similar to networks that are used in the real-world application of this technique. On the other hand, the CPS methods and state-space model approach are less computationally intensive, and these methods were tested using both networks.

We used uniformly weighted spatial averages to form the composite record in our experiments; composites that are formed by wavelet transformation (Moberg et al. 2005) or correlation based weighted averages (Hegerl et al. 2007) are not considered here. Since the SNR and colour of the noise in each pseudo-proxy series is the same, the different weighting scheme should not substantially impact the resulting composite record.

For the total least squares method, $\mathrm{Var}(\varepsilon_t)$ is estimated by calculating the variability of the difference between the pseudo-instrumental record and the climate model simulated NH temperature over the whole reconstruction period. Since the pre-noise contaminated composite record is known in our experiment, $\mathrm{Var}(\eta_t)$ is estimate by calculating the variability of the difference between the pre-noise contaminated and noise contaminated composite pseudo-proxy record, instead of following the procedure described in Hegerl et al. (2007). This is a rather idealized situation in the sense that we have used information that is not available in real-world application to estimate the two variances.

For the MBH method, we retained the 10 largest EOFs instead of using the selection rule that was described in Mann et al. (1998). We also repeated our analysis by retaining the 5 or 15 largest EOFs and the results are almost identical to that obtained with the 10 largest EOFs. Thus, these results will not be shown. Also, no detrending was done to the data prior to calibration (see Wahl et al. 2006; von Storch et al. 2006 for further discussions on the use of detrended data).

For the RegEM method, the hybrid non-stepwise approach is used to reconstruct the grid box temperatures (Rutherford et al. 2005; Mann et al. 2005). As in Rutherford et al. (2005) and Mann et al. (2005), the ridge regression procedure is used to regularize the EM algorithm. Prior to reconstruction, a weight is applied to each standardized proxy series to ensure that the error variance of the signal in the series are homogenous among all records. The weight for the ith proxy series is defined as $\sqrt{\mathrm{Var}(\mathbf{P}^{(i)})/\mathrm{Var}(\mathbf{S}^{(i)})}$, where $\mathbf{P}^{(i)}$ is the ith proxy series and $\mathbf{S}^{(i)}$ is the signal in the ith proxy series. However, $\mathbf{S}^{(i)}$ is unknown in real-world applications. An approximation of this weight can be provided by the sample correlation coefficient between the proxy series and the associated grid box temperature over the calibration period (M. Mann and S. Rutherford, personal communication, 2006). Such a weighting approach is not implemented in Rutherford et al. (2005) and Mann et al. (2005). However, through our experiments with the RegEM method (not shown) and personal communications (2006) with M. Mann and S. Rutherford, we confirmed that reconstruction of the CSM hemispheric mean temperature is sensitive to whether weighted proxy series are used. However, this only applies to reconstructions that use ridge regression to regularize the EM algorithm. Mann et al. (2007) found that results obtained using TTLS regression for regularization are insensitive to whether weights are used. They also pointed out that regularization using TTLS regression is less computationally intensive and tends to provide more robust results than regularization with ridge regression. Hence regularization using TTLS regression would be preferable. However, we were not aware of these advantages at the time of running our experiments, and hence we report on results obtained with the ridge regression procedure. Mann et al. (2005) found that RegEM reconstructions are relatively insensitive to the use of a shorter calibration period and hence, in this analysis, RegEM experiments were only carried out for the 1860–1970 calibration period.

For the state-space model approach, we have run two separate types of experiments in which the variable $\mathbf{F}_t$ in Eq. (6) is defined differently. The first type of experiment accounts for the impact of external forcing when reconstructing the unknown temperature. This is achieved by setting $\mathbf{X}_t = [1 \quad \mathbf{GS}_t \quad \mathbf{VOL}_t \quad \mathbf{SOL}_t]^{\mathrm{T}}$. The second type of experiment only uses the information from the proxy data for reconstruction and is done by using $\mathbf{X}_t = 1$ and $\Upsilon = \tau$. For experiments that investigate the impact of external forcing, the variable $\mathbf{X}_t$ is obtained from an energy balanced model (EBM) driven with reconstructed solar, volcanic and anthropogenic forcings (Hegerl et al. 2003, 2007) . The EBM simulation used here is the same as that used in Hegerl et al. (2007), from which we had available the 30N–90N average response to greenhouse gas, sulfate aerosol, volcanic and solar forcing.

In all the experiments, all parameters except one in Eq. (6) are estimated through the EM algorithm (see Appendix for details). The exception is the parameter ϕ in the exogenous variable $\mathbf{F}_t$, which needs to be estimated outside of the EM algorithm. The value of ϕ will be data dependent because ϕ represents the lag-one autocorrelation of internal variability. For our analysis, it is estimated by calculating the lag-one autocorrelation of the residuals that result from fitting Eq. (4) using the CSM or the ECHO-G model simulated northern hemisphere mean temperature as $\mathbf{T}_t$ and the EBM simulated response to forcings mentioned above as $\mathbf{X}_t$. For the CSM and ECHO-G simulations, ϕ is estimated to be 0.581 and 0.741, respectively. The exogenous variable $\mathbf{F}_t$ used in our analysis is then obtained using these ϕ values. The precise choice of ϕ turns out to have very little impact on the resulting reconstruction (not shown).

Annual mean data are used for all reconstruction methods with some exceptions. Following the typical CPS procedure, to estimate the parameters α and β in the forward regression and the variance matching method, decadally smoothed data are used. The estimated parameters are then used to scale the annual mean pseudo-proxy series to provide the reconstructed annual hemispheric mean temperature. For comparison, we also reconstruct the temperature using parameters that are estimated with annual mean data and we will denote such methods as the non-smoothed forward regression and non-smoothed variance matching methods. On the other hand, as in Mann et al. (1998), the pseudo-instrumental record used in the principle component analysis of the MBH method is

expressed as monthly means and annual mean PCs are subsequently obtained from the monthly mean PCs for analysis.

Figure 1a shows examples of reconstructed NH temperature evolution simulated by the CSM. The reconstructions are based on 15 pseudo-proxy series with SNR = 0.5. It should be noted that the same pseudo-proxy series and pseudo-instrumental record was used in each method to ensure that differences in performance are solely due to the methods themselves. Among the methods that are considered, most did provide a faithful estimate for the simulated NH temperature. The forward regression (smoothed and non-smoothed) and the non-smoothed variance matching methods are exceptions. Comparing the series reconstructed with inverse regression and total least squares regression, it is clear that the two series are visually indistinguishable, suggesting that the neglect of sampling uncertainty in the instrumental record when estimating ζ (Eq. 3) does not have a substantial impact on the results. On the other hand, neglecting the measurement error in the composite record when estimating α (Eq. 1) can cause a substantial bias in the reconstruction. This can be observed by comparing the reconstructed series obtained with forward regression to that obtained with total least squares regression. Experiments using 100 pseudo-proxy series with SNR = 0.5 are displayed in Fig. 1b. The increase in the number of pseudo-proxy series does seem to alleviate the problem in the smoothed forward regression method. The improvement gained when going from 15 to 100 pseudo-proxy series is expected because the noise variance in the composite record of 100 pseudo-proxy series is only 15% of that with 15 pseudo-proxy series. However, such improvement may be less significant in real-world applications given that errors might be spatially correlated. Results obtained with the ECHO-G simulation are very similar and thus not shown.

The test results for a reconstruction method can be affected by both the specific realizations of noise that are added when creating the pseudo-proxy record, and the locations of the pseudo-proxy record. One would expect the reconstruction to differ as a result of sampling variability from at least these two sources. Therefore, to get a better picture of the performance of each reconstruction method, it is necessary to apply them to a number of realizations of the pseudo-proxy record. Hence, we reconstructed the hemispheric mean temperature series 100 times for each method using 100 different realizations of pseudo-proxy records. For each realization, the locations of the pseudo-proxies also change randomly within the 321 grid boxes that are specified before, as well as the noise. For the network with 100 proxy series, only 40 reconstructions are produced for each method due to computational limitations. However, 100 reconstructions were produced for each method for the smaller 15 locations proxy network.

To provide a quantitative assessment for each reconstruction method, we computed the relative root mean squared error (RRMSE) of the reconstruction error, expressed relative to the variability of the model simulated NH temperature during the pre-calibration period. The RRMSE is simply defined as

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\mathbf{T}_t - \hat{\mathbf{T}}_t)^2}{\sum_{t=1}^{n}(\mathbf{T}_t - \overline{\mathbf{T}})^2}}$$

where $\mathbf{T}_t, \hat{\mathbf{T}}_t$ and $\overline{\mathbf{T}}$ are the model simulated NH temperature, the reconstructed NH temperature and the temporal mean of the model simulated NH temperature during the pre-calibration period respectively. In general, a smaller RRMSE means a better reconstruction. The RRMSE can lie between zero and infinity and a RRMSE value of less than 1 indicates that the reconstructed series is better than a reconstruction that has a constant value equal to the climatology of the pre-calibration period.

Figure 2 shows the median of the RRMSEs obtained from the 100 (or 40) realizations, as a function of the degree of smoothing of the climate model simulated annual mean series $\mathbf{T}_t$ and the reconstructed annual mean series $\hat{\mathbf{T}}_t$. An estimated 5–95% uncertainty range of the RRMSE is also displayed in the figure, which is obtained by using the 5th and 95th percentiles of the sample of RRMSEs. Comparing the results obtained between the two simulations, the results are robust for most of the methods. The performance of most of the methods considered is very similar at decadal and lower resolution. The non-smoothed forward regression, non-smoothed variance matching and MBH methods are exceptions. The performance of all methods was found to be insensitive to the two choices of calibration period, and thus results obtained using the shorter calibration periods are not shown. In fact, for both the CSM and ECHO-G simulation, there is almost no change in the median value of RRMSE when the shorter calibration period is used.

At the annual resolution, the RegEM and state-space model approaches produce the smallest RRMSEs. This is a result of the more sophisticated procedures that are involved in these methods, which in effect filter out the measurement errors in the pseudo-proxy series. In contrast, the reconstructed series from a typical CPS method is merely a scaled version of the composite series, with the result that the measurement error contained in the composite series is directly transferred to the reconstructed series. This problem is more severe when only 15 proxy series are available for reconstruction (Fig. 3). Hence, the simple CPS methods should be avoided if the goal is to reconstruct high frequency climate variation. At the same time, the MBH

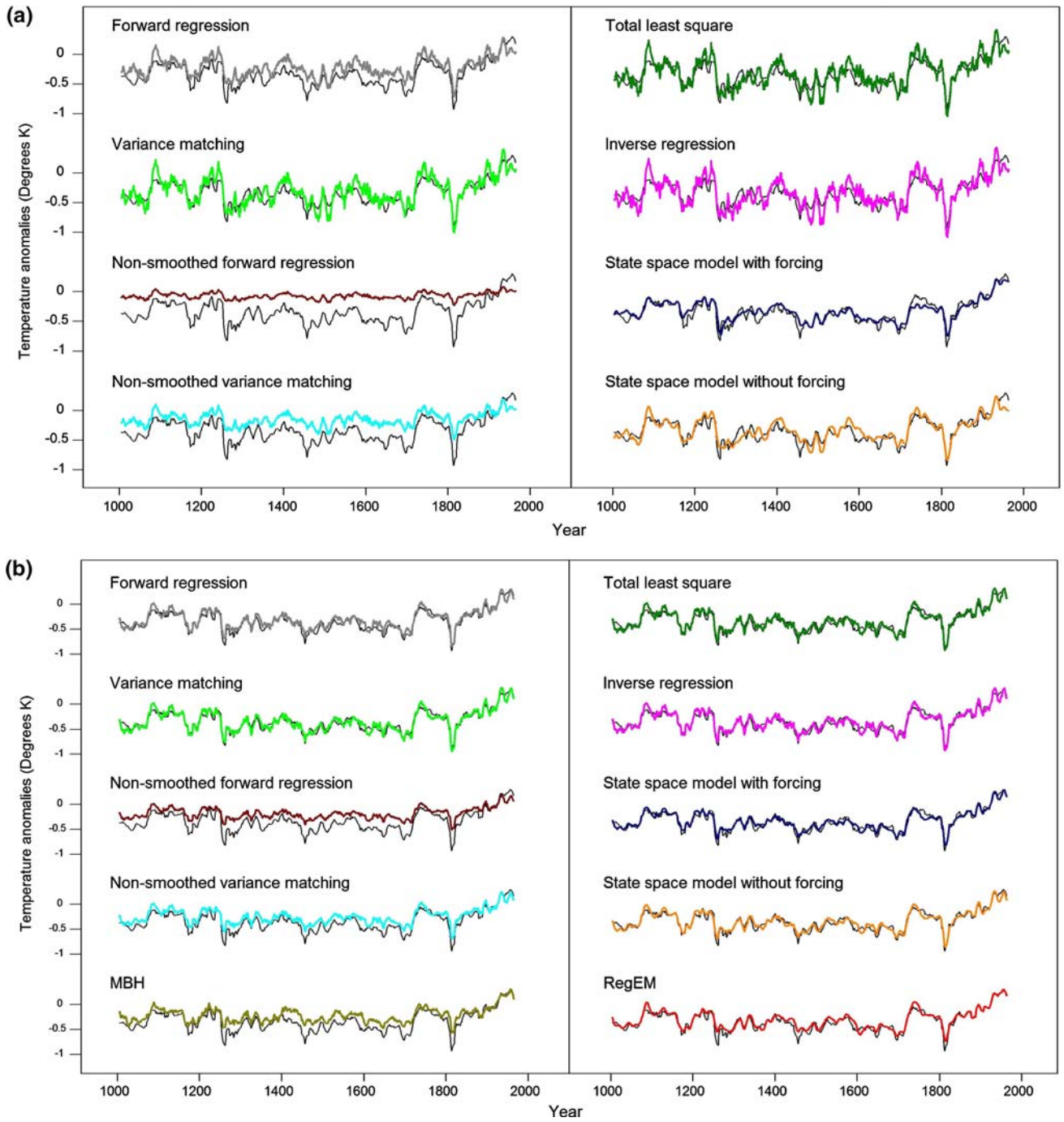

Fig. 1 Examples of reconstructed CSM northern hemisphere mean temperature series. Experiments were run using SNR = 0.5 and the 1860–1970 calibration period with **a** 15 and **b** 100 pseudo-proxy series. 11-year moving averages are shown. The CSM NH mean is shown in *black*. Reconstructions are shown as *coloured lines*. All series are express as anomalies relative to the calibration period. Units (degrees Kelvin)

method is observed to be the worst performer when SNR = 0.5. However, when SNR is increased to 1, the MBH method, at annual resolution, produces comparable RRMSEs to that of most CPS methods considered.

The estimated RRMSE uncertainty ranges provide information on the sensitivity to sampling variability for each method. From Fig. 2, it is obvious that such sensitivity is larger when only 15 pseudo-proxy series are used, reflecting the greater sampling variability in the composite record when only 15 proxy series are used. Inter-comparison between the different methods suggests that the sensitivity to sampling variability at annual resolution is

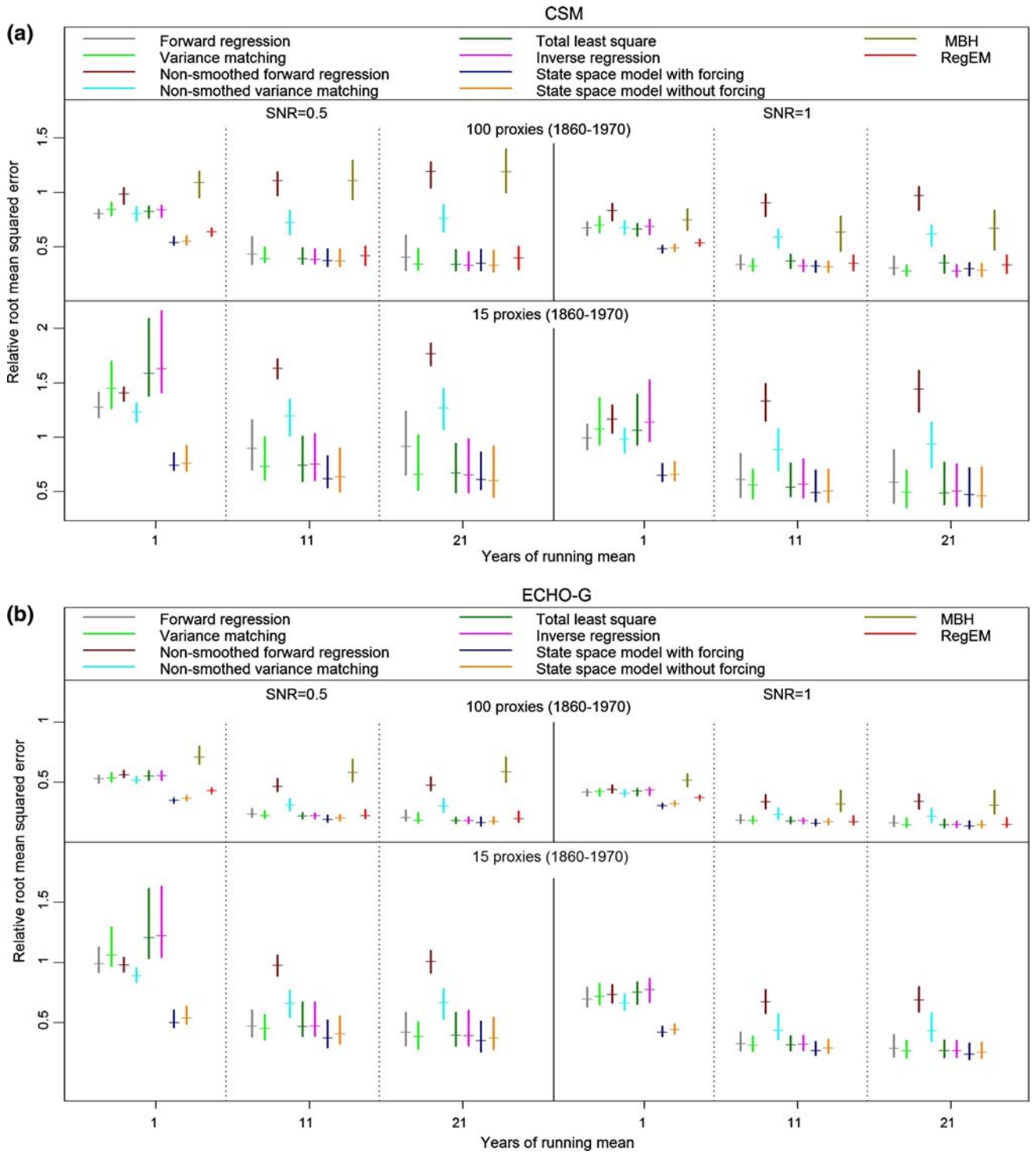

Fig. 2 Relative root mean squared error (*RRMSE*) of the reconstruction error, expressed relative to the variability of the simulated hemispheric temperature. Units (degrees Kelvin). The median RRMSE is indicated with horizontal bars and the estimated 5–95% range of the RRMSEs is shown with vertical lines. Results using the 1860–1970 calibration period with different signal-to-noise ratios and varying number of pseudo-proxies are shown for the two climate model simulations: **a** CSM and **b** ECHO-G

somewhat smaller for the RegEM and state-space model approaches. At decadal or lower resolution, the RRMSE uncertainty range is very similar across most methods.

By comparing the results of the two state-space model approaches (with external forcing response and without), it is clear that the RRMSE is insensitive to the inclusion or

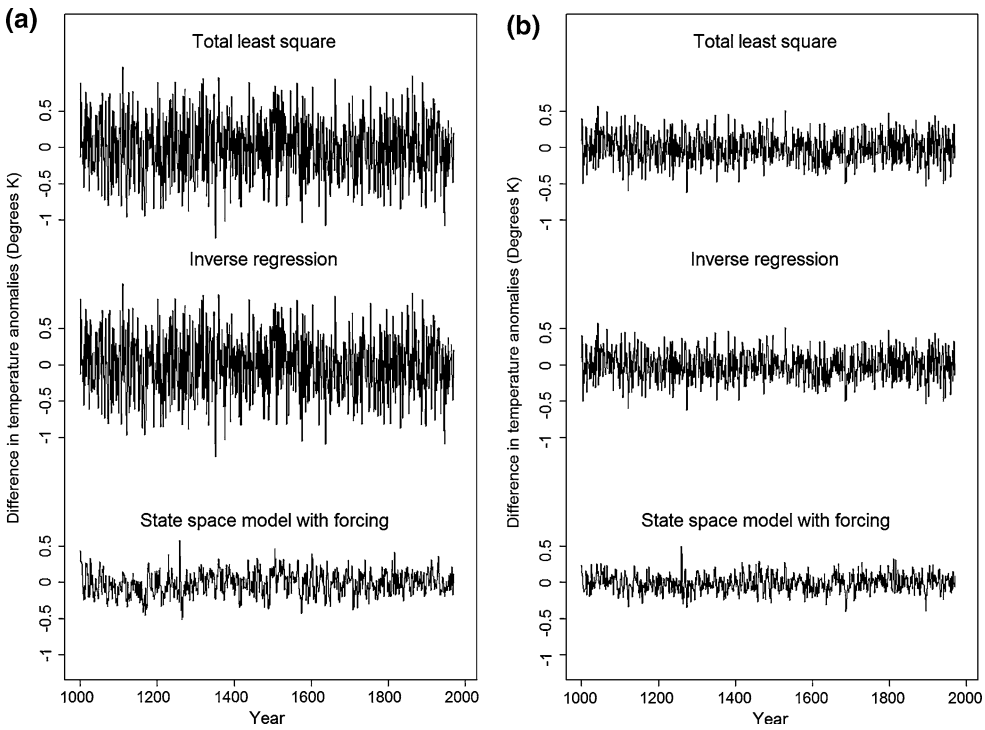

Fig. 3 Example of the difference in temperature anomalies between the ECHO-G simulation and the reconstructed ECHO-G series using SNR = 0.5 and the 1860–1970 calibration period, plotted at annual resolution with **a** 15 pseudo-proxy series and **b** 100 pseudo-proxy series

exclusion of the EBM estimated forcing response information. Nevertheless, the reconstructed series from the two approaches are slightly different when only 15 pseudo-proxy series are used (Fig. 1a). This suggests that the Kalman filter and smoother algorithm relies more heavily on the proxy series than the estimated response to forcings to estimate the unknown temperature. Even though the reconstructions from the two state-space model approaches are very similar, the approach that takes forcing changes into account may be more useful in some instances since it may be possible to use it to provide a detection assessment and perhaps also a projection of future climate.

As mentioned above, one can use the state-space model to simultaneously reconstruct the NH temperature and conduct detection analysis. Both the CSM and ECHO-G simulations are forced with a combination of external forcing factors and therefore we should be able to detect the effect of external forcing in our experiments provided that these models respond similarly to forcing as the EBM. Figure 4 shows the confidence bounds on the forcing response coefficients for the experiments using SNR = 0.5. For $\delta_{\mathbf{GS}}$ and $\delta_{\mathbf{VOL}}$, the confidence bounds do not include zero for almost all experiments that were run, indicating that the EBM simulated GS and VOL signals are detectable in the reconstruction of the CSM and ECHO-G simulations. However, the response to solar forcing as simulated by the EBM is not detectable in about half of the CSM reconstructions obtained using 15 pseudo-proxy series. This fraction is reduced to near zero when the SNR is increased to 1. In contrast, the EBM simulated SOL signal is detectable in all the CSM experiments that use 100 pseudo-proxy series and in all ECHO-G experiments. The inability to detect the SOL forcing in some experiments may be due to the fact that the climate response to solar forcing is relatively weaker than that to the other forcings and hence may be harder to detect when the noise contamination in the pseudo-proxy series increases. At the same time, unlike the ECHO-G and EBM simulations, the solar forcing estimates used in the CSM simulation excluded the 11-yr solar cycle and this may also contribute to the varying detection results for the response to solar forcing. The inability to consistently detect the response to SOL forcing in our experiments is consistent with detection work on real-world paleo-reconstructions (Hegerl et al. 2003, 2007).

Figure 5 displays hindcasts of annual NH mean temperature for 1971 to 1990 with 100 pseudo-proxy series, SNR = 0.5 and the 1860–1970 calibration period. The hindcasts are produced using Eq. (9). For comparison, the sum of the responses to external forcings for the average of 30N–90N as simulated by the EBM is also displayed in the figure. We have produced two hindcasts for each climate model using the same set of pseudo-proxy series, one using the full 1000–1970 period to estimate the parameters for the state-space model and another using only data from 1800 to 1970. It can be observed that the skill of the hindcast varies between the two analysis periods. In fact, the estimates of the forcing response coefficients, which are influential to the hindcast values, are substantially different for the two analysis periods (Table 1). However, these

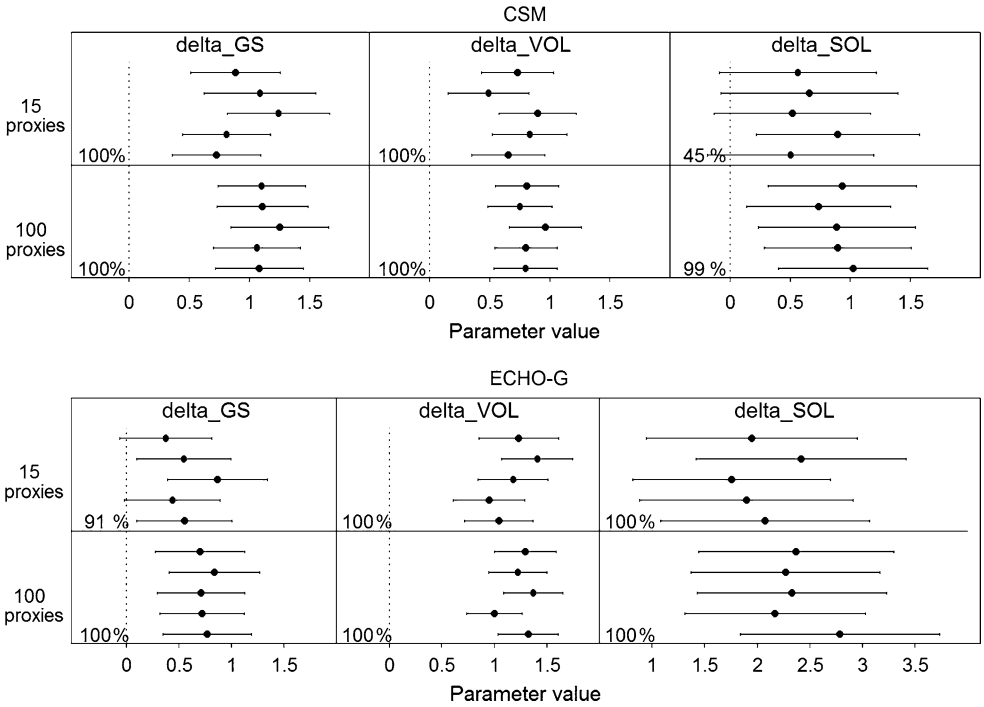

Fig. 4 95% confidence bounds for the coefficients used to scale the EBM simulated responses to external forcing in the state-space model when the pseudo-proxy SNR is 0.5 and using the 1860–1970 calibration period. Only results from 5 out of the 100 (or 40) experiments are displayed. The number in the *bottom left corner of each box* indicates the percentage of confidence bounds (out of 100 or 40) that excludes zero

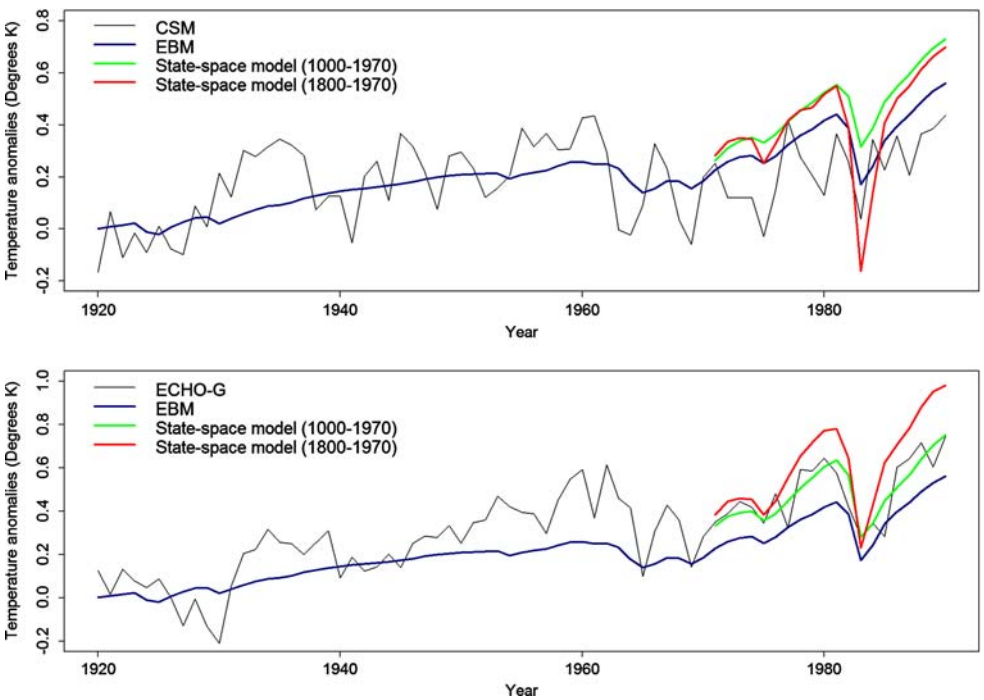

Fig. 5 Hindcasts of annual mean NH temperature based on the estimated state equation from 100 proxy series for two analysis periods: 1000–1970 and 1800–1970. The sum of response to external forcings for the 30N–90N average as simulated by the energy balance model is also displayed for comparison purposes. All series are expressed as anomalies relative to the 1860–1970 period

differences have only a minor influence on the reconstructed series (not shown) because the Kalman smoother algorithm relies more heavily on the proxy data than the EBM to reconstruct the unknown temperature.

Figure 6 displays examples of reconstructed NH temperature from the MBH method using the two different simulations and different SNR with the 1860–1970 calibration period. Von Storch et al. (2004) included a detrending step in their test of the MBH method. We therefore repeated our experiments with detrended calibration data and those results are also shown in Fig. 6. When the variance of the noise added in the pseudo-proxy series is the same as the grid-box temperature variance, that is, when SNR=1, the non-detrended MBH method was able to provide a reasonable reconstruction of the ECHO-G simulated hemispheric mean temperature. However, results

Table 1 Estimated parameter values of the state-space model obtained with 100 pseudo-proxy series using two analysis periods

Parameter	CSM		ECHO-G	
	1000–1970	1800–1970	1000–1970	1800–1970
δ_{GS}	**1.315**	**2.105**	**0.775**	**1.978**
δ_{VOL}	**0.994**	**2.334**	**1.270**	**1.875**
δ_{SOL}	**0.891**	−0.298	**2.779**	**6.456**

Boldfaced values are significant

become unsatisfactory when the SNR is decreased to 0.5. For the reconstruction of CSM hemispheric mean temperature, the result is poor with both SNRs. Although our calibration period is longer than that used in Mann et al. (1998), our result does not change substantially when the shorter 1880–1960 calibration period is used (not shown). While our result does suggest the MBH method underestimates long term variability, the under-estimation is smaller when non-detrended data is used (see Wahl et al. 2006; von Storch et al. 2006 for further discussions).

We also tested the CPS methods that are mentioned in the previous section using detrended calibration data (not shown). The performance of the CPS methods varies across different realizations of pseudo-proxy records. In some cases, detrending does not affect the reconstructed series irrespective of which CPS method is considered. On the other hand, there were also realizations of pseudo proxies for which there was under-estimation of variability when detrended data are used. Therefore, in the context of pseudo-proxies constructed with white noise, detrending results in less robust reconstructions of hemispheric mean temperature variability.

3.1.1 Analysis with red pseudo-proxy noise

Up to this point, our experiments have not taken into account the possibility that the proxies consist of a temperature signal plus correlated errors. We therefore now examine the impact of red pseudo-proxy noise on the reconstruction methods. Following Mann et al. (2007), red pseudo-proxy noise is generated from an AR(1) process with lag-one autocorrelation equal to 0.32. As before, analyses are conducted using two calibration periods and with SNR = 0.5 and 1.

Examples of reconstructed NH temperature evolution simulated by the CSM based on red pseudo-proxy series with SNR=0.5 are displayed in Fig. 7. It is clear that the redness of the pseudo-proxy noise has slightly increased the variability of the reconstructed series when 15 pseudo-proxy series are used. On the other hand, series reconstructed with 100 pseudo-proxy series do not seem to be affected. Results obtained with the ECHO-G simulation are very similar and thus not shown.

The estimated RRMSEs obtained with red pseudo-proxy series (not shown) are very similar to those shown in Fig. 2. In particular, the estimated median RRMSEs are almost unchanged and the relative ranking of the RRMSEs between the different reconstruction methods remains the same as in the case of white pseudo-proxy series. Hence, similar conclusions regarding the performance of the different methods can be drawn as before. However, the RRMSE uncertainty ranges are slightly larger than before when 15 pseudo-proxy series are used.

3.2 Analysis with real-world paleoclimate proxy data

We apply the CPS methods and state-space model approach to the paleoclimate proxy data used in Hegerl et al. (2007). This data set consists of 14 proxy series. All records are available as decadally smoothed series. As in Hegerl et al. (2007), we calculated correlation based weighted averages to form the composite record. Using 1880–1960 as the calibration period, the 30N–90N mean temperature is reconstructed for the period 1510–1960. As noted previously, the use of the state-space model approach requires fixing the parameter ϕ in the exogenous variable $\mathbf{F}_t$. Here, the value of this parameter is estimated by the lag-one autocorrelation of the decadally smoothed 30N–90N mean temperature from a control simulation of the CCCma CGCM2 (Flato and Boer 2001). The resulting ϕ value, 0.982, was used to obtain the exogenous variable $\mathbf{F}_t$.

Figure 8 compares the reconstructed series obtained from the variance matching method, state-space model approach and Hegerl et al. (2007), who used the total least squares method. The reconstruction estimates obtained from these approaches are nearly identical. Reconstructed series from the other CPS approaches also agree closely with the series in Fig. 8 (not shown). This finding is consistent with results obtained using climate model simulations in the previous section where it was seen that these methods have similar RRMSEs at the decadal resolution.

Estimates of the parameters of the state-space model (with forcing) are given in Table 2. The parameter ϕ is estimated to be close to 1, which is larger than that obtained using climate model simulations with annual mean data, which ranges from 0.3 to 0.7. This is not surprising given that decadally smoothed data is strongly dependent between successive time points. The estimate parameter value for $\delta_{\mathbf{GS}}$ and $\delta_{\mathbf{VOL}}$ is significantly different from zero, suggesting that the response to GS and VOL forcing are detectable. On the other hand, the response to SOL forcing is not detected. These detection results agree with the findings reported in Hegerl et al. (2007).

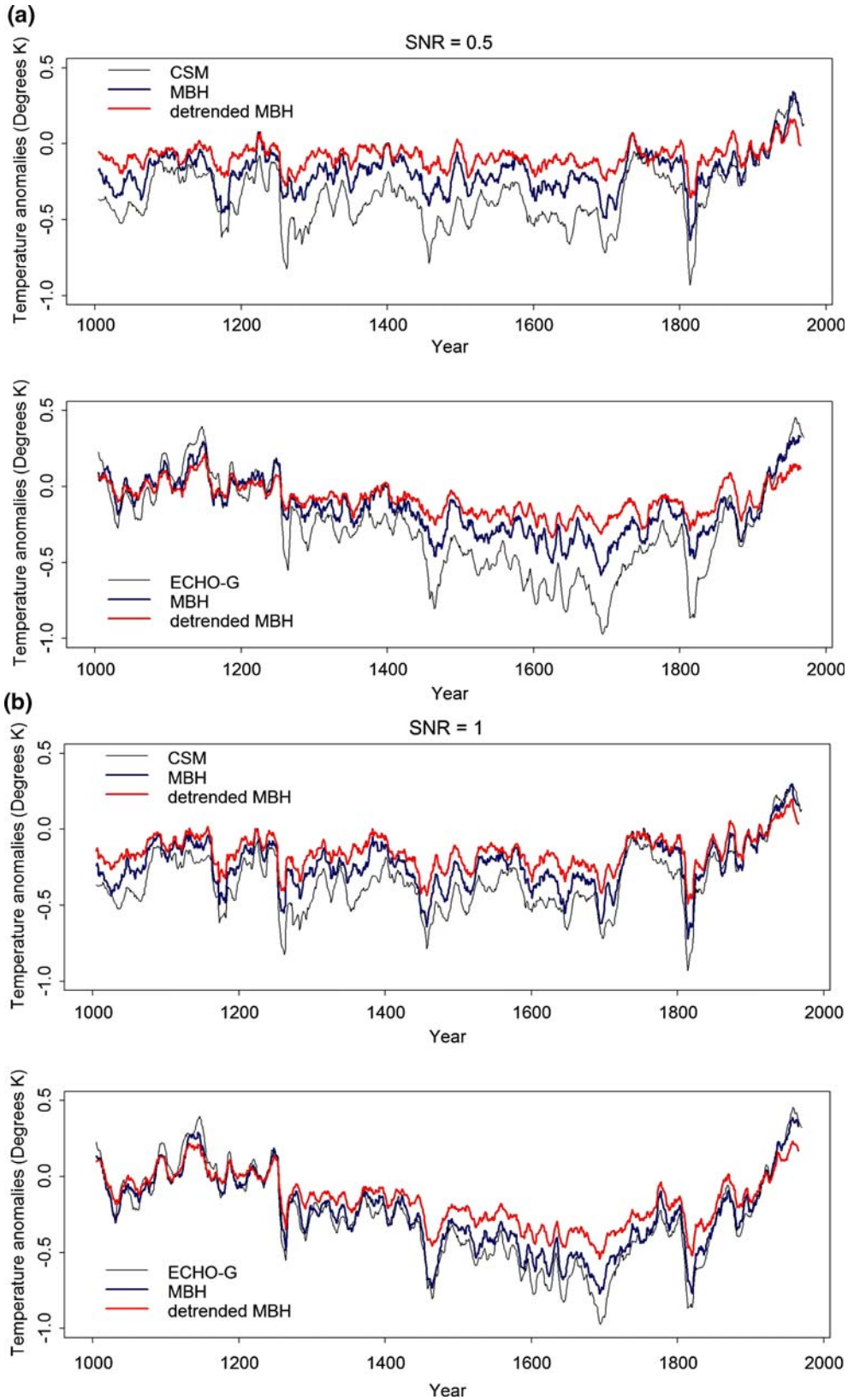

Fig. 6 Comparison of the reconstructed hemispheric mean temperature series obtained using the MBH method with non-detrended or detrended data at two signal-to-noise ratios **a** SNR = 0.5 and **b** SNR = 1 with 100 pseudo-proxy series and the 1860–1970 calibration period. All series are expressed as 11-year running means and as anomalies relative to the calibration period

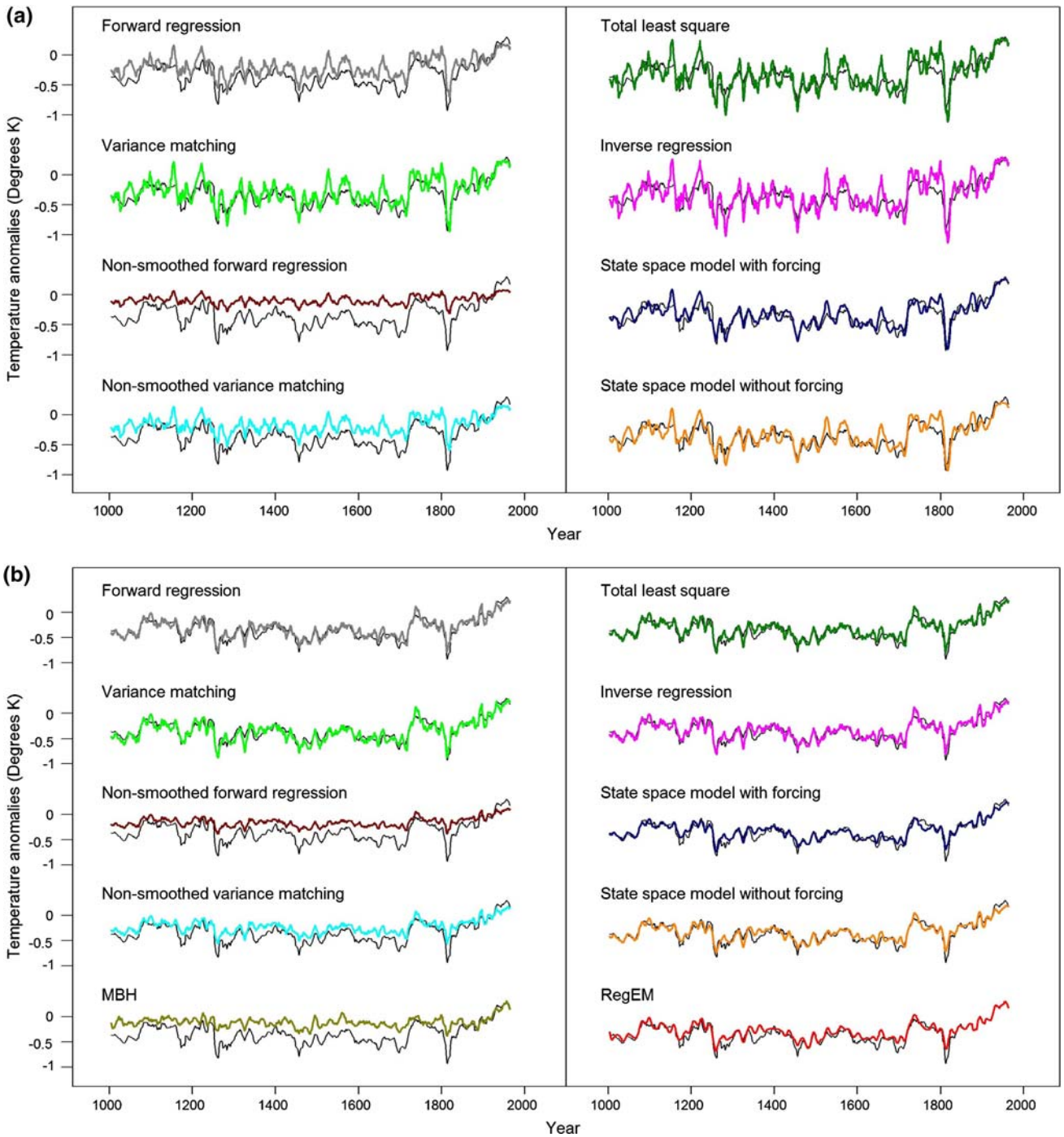

Fig. 7 Examples of reconstructed CSM northern hemisphere mean temperature series. Experiments were run using SNR = 0.5 and the 1860–1970 calibration period with **a** 15 and **b** 100 red pseudo-proxy series. 11-year moving averages are shown. The CSM NH mean is shown in black. Reconstructions are shown as *coloured lines*. All series are express as anomalies relative to the calibration period. Units (degrees Kelvin)

The 30N–90N decadally smoothed mean temperature hindcast for 1961–1990 is displayed in Fig. 9. Hindcasts are generated for three analysis periods (1270–1960, 1510–1960 and 1800–1960). Confidence bounds on the hindcast for the 1510–1960 analysis period are also displayed in the figure. The hindcasts obtained from the 1270–1960 and 1510–1960 analysis periods are very similar and is very close to the sum of the response to external forcings as simulated by the EBM. The confidence bounds on the hindcasts generated for the 1510–1960 analysis were able to include almost all the observed temperature anomalies. Similar results can be obtained for the 1270–1960 analysis

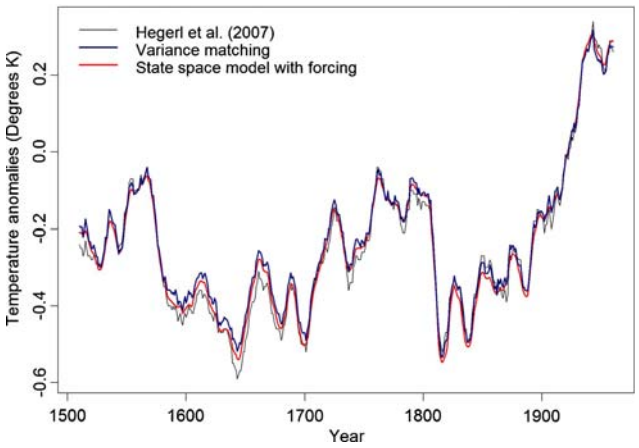

Fig. 8 Reconstructed series for the 30N–90N mean using the variance matching method and state-space model approach for real-world paleoclimate proxy data. Temperature series are expressed as 11-year moving averages. All series are expressed as anomalies relative to the 1880–1960 period and in units of degrees Kelvin

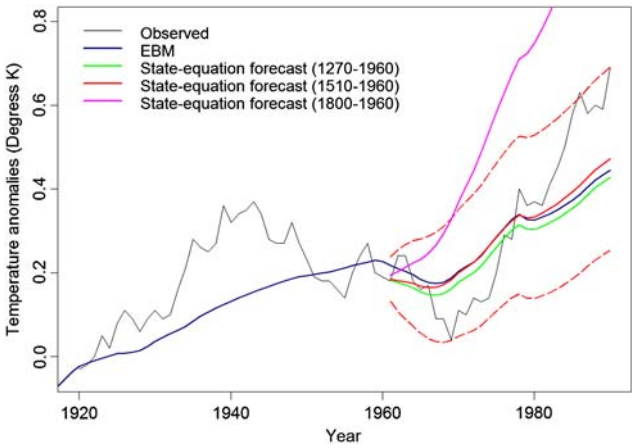

Fig. 9 State equation hindcast of decadally smoothed 30N–90N mean temperature with real-world paleoclimate proxy data. All hindcasts are based on the estimated state equation for three analysis periods: 1270–1960, 1510–1960 and 1800–1960. Nine proxy series are available for the 1270–1960 analysis period and 14 proxy series are available for the other two analysis period. 95% confidence bounds of the hindcasts for the 1510–1960 analysis period are shown as dashed lines. All series are expressed as anomalies relative to the 1880–1960 period and in units of degrees Kelvin

period (not shown). On the other hand, the hindcasts obtained from the 1800–1960 analysis period warm too quickly. This is because the estimated value of $\delta_{\mathbf{GS}}$ is four times larger than that in the other two analysis periods (Table 2).

4 Conclusion

In this paper, we have compared the skill of several different reconstruction methods using climate model simulations. At the annual resolution, the state-space model and RegEM approaches provide the best reconstructions. On the other hand, when compared at decadal or lower resolution, we find that most methods can provide satisfactory and similar results. Exceptions are the MBH, non-smoothed forward regression and non-smoothed variance matching methods. When analysed with decadally smoothed real-world paleoclimate proxy data, we find that all of the CPS methods considered provide almost identical results. The similarity in performance provides evidence that the difference between many real-world reconstructions is more likely to be due to the choice of the proxy series, or the use of difference target seasons or latitudes than to the choice of statistical reconstruction method (see also Juckes et al. 2006).

We have also put forward another approach to historical temperature reconstruction that is based on a state-space time series model and the Kalman filter and smoother algorithm. This approach allows the possibility of incorporating additional non-proxy information into the reconstruction analysis, such as the estimated response to external forcing. However, our experiments show that the state-space model approach does not produce substantially different reconstructions when such information is included. Nevertheless, both state-space model approaches provided better reconstructions than existing CPS methods at annual resolutions. At the same time, including forcing response terms in the state-space model allows one to carry out a simultaneous reconstruction and detection analysis. It can also be used to provide forecasts of future climates. Consistent with the results of Hegerl et al. (2007), we have detected the effects of anthropogenic forcing (greenhouse gas and aerosol) and volcanic forcing in real-world paleoclimate proxy data.

Table 2 Estimated parameter values of the state-space model obtained with paleoclimate proxy data for three reconstruction periods

Parameter	1270–1960	1510–1960	1800–1960
ϕ	**0.981** (0.967, 0.994)	**0.979** (0.961, 0.997)	**0.944** (0.900, 0.988)
δ_{GS}	**1.022** (0.043, 2.001)	**1.128** (0.112, 2.143)	**4.271** (1.054, 7.489)
δ_{VOL}	**0.953** (0.685, 1.220)	**0.814** (0.522, 1.106)	**0.998** (0.461, 1.535)
δ_{SOL}	0.784 (−0.592, 2.161)	0.368 (−1.218, 1.955)	-0.203 (−3.211, 2.805)

The 95% confidence interval is listed in brackets. Boldfaced values are significant

Acknowledgments We thank Caspar Ammann, Gabriele Hegerl and Eduardo Zorita for providing their data for use in this study. We also thank Gabriele Hegerl for helpful and constructive discussion. We gratefully acknowledge that Terry Lee was supported by the Canadian Foundation for Climate and Atmospheric Science through the Canadian CLIVAR Research Network. Work by Min Tsao was supported by the Natural Sciences and Engineering Research Council through a Discovery Grant. This paper was improved by insightful and helpful comments provided by Scott Rutherford, Walter Skinner, Xuebin Zhang and an anonymous referree.

5 Appendix

5.1 Expectation maximization algorithm

Here we present the steps required to estimate the unknown parameters that specify the state-space model (Eq. 6). We use $\Theta = \{\zeta, R, \phi, \Upsilon, Q, \mu, \Sigma\}$ to represent the vector of unknown parameters. The derivations presented here are expanded from Shumway and Stoffer (1982, 2000, pp. 321–325). Detail derivations of the following procedure are given in Lee (2008). We first write down the likelihood function for $\{\mathbf{P}_t;\ t = 1, 2, ..., n\}$ as in Shumway and Stoffer (2000). Ignoring a constant, the likelihood function $L_{\mathrm{P}}(\Theta)$ can be expressed as:

$$-2\ln L_{\mathrm{P}}(\Theta) = \sum_{t=1}^{n} \ln(\Omega_t) + \sum_{t=1}^{n} (\varepsilon_t^2/\Omega_t)$$

where $\Omega_t = \zeta^2 S_t^{t-1} + R > 0$ and $\varepsilon_t = \mathbf{P}_t - \zeta \mathbf{T}_t^{t-1}$ for $t = 1, 2, ..., n$. The calibration period hemispheric mean temperatures $\{\mathbf{T}_t;\ t = n + 1, ..., n + m\}$ are not involved in the above likelihood function. To better utilize the available information, we have modified the likelihood function to include $\{\mathbf{T}_t;\ t = n + 1, ..., n + m\}$. Ignoring a constant, one can express the likelihood function for the observation data set $\{\mathbf{P}_t, \mathbf{T}_s;\ t = 1, 2, ..., n;\ s = n + 1, ..., n + m\}$, namely $L_{\mathrm{P,T}}(\Theta)$, as

$$\begin{aligned} -2\ln L_{\mathrm{P,T}}(\Theta) = &\sum_{t=1}^{n+m} \ln(\Omega_t) + \sum_{t=1}^{n+m} (\varepsilon_t^2/\Omega_t) + (m-1)\ln(Q) \\ &+ \ln(\phi^2 S_n^n + Q) + (\mathbf{T}_{n+1} - \phi \mathbf{T}_n^n - \Upsilon \mathbf{F}_{n+1})^2/ \\ &(\phi^2 S_n^n + Q) + \sum_{t=n+2}^{n+m} (\mathbf{T}_t - \phi \mathbf{T}_{t-1} - \Upsilon \mathbf{F}_t)^2/Q \end{aligned}$$

where Ω_t and ε_t is defined as before for $t = 1, 2, ..., n$. For $t = n + 1, ..., n + m$, $\Omega_t = R$ and $\varepsilon_t = \mathbf{P}_t - \zeta \mathbf{T}_t$. The goal here is to find the Θ values that maximize the likelihood function $L_{\mathrm{P,T}}(\Theta)$. Using the EM algorithm, such Θ values can be found through an iterative procedure. The estimation procedure is summarized as follows.

1. Select the starting values for the parameters $\Theta^{(0)} = \{\zeta^{(0)}, R^{(0)}, \phi^{(0)}, \Upsilon^{(0)}, Q^{(0)}, \mu^{(0)}\}$ while fixing Σ, one of the initial condition for the Kalman recursion. (In our analysis, we set $\Sigma = 0.05$. Results were almost identical when different Σ values were used.) On iteration j, ($j = 1, 2, ...$), do steps 2–4.
2. Let $N = n + m$. Using $\Theta^{(j-1)}$, compute the Kalman filter and smoother estimates using Eqs. (7) and (8) for $t = 1, 2, ..., N$ and the likelihood function $\ln L_{\mathrm{P,T}}(\Theta^{(j-1)})$. Also calculate, for $t = N - 1, N - 2, ..., 0$

$$S_t^N = S_t^t + J_t^2 (S_{t+1}^N - S_{t+1}^t)$$

and for $t = n, n - 1, ..., 0$,

$$\begin{aligned} C_t &= J_{t-1} S_t^N \\ \widetilde{\mathbf{T}}_t^N &= \mathbf{T}_t^N + (J_t J_{t+1} \ldots J_n)(\mathbf{T}_{n+1} - \mathbf{T}_{n+1}^N) \\ \widetilde{S}_t^N &= S_t^N - (J_t J_{t+1} \ldots J_n)^2 S_{n+1}^N \end{aligned}$$

and the following quantities:

$$\begin{aligned} Z_{11} &= \sum_{t=1}^{n} [(\widetilde{\mathbf{T}}_t^N)^2 + \widetilde{S}_t^N] + \sum_{t=n+1}^{N} (\mathbf{T}_t)^2 \\ Z_{00} &= \sum_{t=0}^{n} [(\widetilde{\mathbf{T}}_t^N)^2 + \widetilde{S}_t^N] + \sum_{t=n+1}^{N-1} (\mathbf{T}_t)^2 \\ Z_{10} &= \sum_{t=1}^{n} (\widetilde{\mathbf{T}}_t^N \widetilde{\mathbf{T}}_{t-1}^N + C_t) + \mathbf{T}_{n+1} \widetilde{\mathbf{T}}_n^N + \sum_{t=n+2}^{N} \mathbf{T}_t \mathbf{T}_{t-1} \\ F_{11} &= \sum_{t=1}^{n} \mathbf{F}_t \widetilde{\mathbf{T}}_t^N + \sum_{t=n+1}^{N} \mathbf{F}_t \mathbf{T}_t \\ F_{10} &= \sum_{t=1}^{n+1} \mathbf{F}_t \widetilde{\mathbf{T}}_{t-1}^N + \sum_{t=n+2}^{N} \mathbf{F}_t \mathbf{T}_{t-1} \\ F_{00} &= \sum_{t=1}^{N} \mathbf{F}_t \mathbf{F}_t^{\mathrm{T}}. \end{aligned}$$

3. Obtain $\Theta^{(j)}$ using the following:

$$\begin{aligned} \zeta^{(j)} &= \left[\sum_{t=1}^{n} [(\widetilde{\mathbf{T}}_t^N)^2 + \widetilde{S}_t^N] + \sum_{t=n+1}^{N} \mathbf{T}_t^2\right]^{-1} \\ &\quad \left[\sum_{t=1}^{n} \widetilde{\mathbf{T}}_{t|N} \mathbf{P}_t + \sum_{t=n+1}^{N} \mathbf{T}_t \mathbf{P}_t\right] \\ R^{(j)} &= N^{-1} \sum_{t=1}^{n} \left[\widetilde{S}_t^N (\zeta^{(j)})^2 + (\mathbf{P}_t - \zeta^{(j)} \widetilde{\mathbf{T}}_t^N)^2\right] \\ &\quad + N^{-1} \sum_{t=n+1}^{N} \left(\mathbf{P}_t - \zeta^{(j)} \mathbf{T}_t\right)^2 \\ \Upsilon^{(j)} &= \left(F_{11}^{\mathrm{T}} - F_{10}^{\mathrm{T}} Z_{10}/Z_{00}\right)\left(F_{00} - F_{10} F_{10}^{\mathrm{T}}/Z_{00}\right)^{-1} \\ \phi^{(j)} &= \left(Z_{10} - \Upsilon^{(j)} F_{10}\right) Z_{00}^{-1} \\ Q^{(j)} &= N^{-1}\left(Z_{11} - \phi^{(j)} Z_{10} - \Upsilon^{(j)} F_{11}\right) \\ \mu_0^{(j)} &= \widetilde{\mathbf{T}}_0^N. \end{aligned}$$

4. Repeat steps 2 and 3 until convergence. For the analysis presented in this paper, the algorithm is stopped when $\ln L_{\mathrm{P,T}}(\Theta^{(j)}) - \ln L_{\mathrm{P,T}}(\Theta^{(j-1)}) < 0.0005$.

To provide a final estimate of $\mathbf{T}_t$, the Kalman filter and smoother estimates are recalculated using Eqs. (7) and (8) with the estimate parameters $\hat{\Theta}$, for $t = 1, 2, ..., n + m$. At the time of convergence, one can also calculate the standard errors for $\hat{\Theta}$. For parameters estimated using ln $L_P(\Theta)$, the asymptotic variance of $\hat{\Theta}$ is defined as (Caines 1988, Chap. 7; Jensen and Petersen 1999):

$$\mathrm{Var}(\hat{\Theta}) = -\left[\frac{\partial^2}{\partial \Theta} \ln L_P(\Theta)\bigg|_{\hat{\Theta}}\right]^{-1}$$

where $\partial^2/\partial\Theta$ denotes the second derivatives with respect to Θ. In our application, we have used $L_{P,T}(\Theta)$ in the estimation procedure and hence a reasonable estimate of the asymptotic variance can be obtained by replacing $L_P(\Theta)$ with $L_{P,T}(\Theta)$ in the above formula. More discussions of this can be found in Lee (2008). An analytical form of the asymptotic variance is generally hard to find and hence we calculated it numerically. The asymptotic variance can be used to provide confidence bounds for the estimated parameters. In particular, the 95% confidence bound for Θ is defined as $\hat{\Theta} \pm 1.96\sqrt{\mathrm{Var}(\hat{\Theta})}$. In our application, we are interested in the confidence bounds for the parameters $\delta_{\mathbf{GS}}$, $\delta_{\mathbf{VOL}}$ and $\delta_{\mathbf{SOL}}$. These bounds can provide a detection assessment of the importance of the response to the GS, VOL and SOL forcing on the hemispheric mean temperature.

References

Allen MR, Stott PA (2003) Estimating signal amplitudes in optimal fingerprinting. Part I: theory. Clim Dyn 21:477–491

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

Briffa KR, Osborn TJ, Schweingruber FH, Harris IC, Jones PD, Shiyatov SG, Vaganov EA, (2001) Low-frequency temperature variations from a northern tree-ring density network. J Geophys Res 106:2929–2941

Brohan P, Kennedy JJ, Haris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106

Brown PJ (1993) Measurement, regression, and calibration. Oxford University Press, Oxford, 201 pp

Bürger G, Cubasch U (2005) Are multiproxy climate reconstructions robust? Geophys Res Lett 32:L23711

Bürger G, Fast I, Cubasch U (2006) Climate reconstruction by regression—32 variations on a theme. Tellus 58A:227–235

Caines PE (1988) Linear stochastic systems. Wiley, New York, 847 pp

Coelho CAS, Pezzulli S, Balmaseda M, Doblas-Reyes JJ, Stephenson D (2004) Forecast calibration and combination: a simple Bayesian approach for ENSO. J Clim 17:1504–1516

Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford, 253 pp

Esper J, Cook ER, Schweinngruber FH (2002) Low-frequency in long treering chronologies for reconstruction past temperature variability. Science 295:2250–2253

Esper J, Frank DC, Wilson RJ, Briffa KR (2005) Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. Geophys Res Lett 32:L07711

Fierro RD, Golub GH, Hansen PC, O'Leary DP (1997) Regularization by truncated total least squares. SIAM J Sci Comput 18:1223–1241

Flato G, Boer GJ (2001) Warming asymmetry in climate change experiments. Geophys Res Lett 28:195–198

Fuller WA (1987) Measurement error models. Wiley, New York, 440 pp

Harvey A (1989) Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge, 554 pp

Hegerl GC, Crowley TJ, Baum SK, Kim K-Y, Hyde WT (2003) Detection of volcanic, solar, and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature. Geophys Res Lett 30:1242 doi:10.1029/2002GL016635

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

Jensen JL, Petersen NV (1999) Asymptotic normality of the maximum likelihood estimator in state space models. Ann Stat 27:514–535

Jones PD, Osborn TS, Briffa KR (1997) Estimating sampling errors in large-scale temperature averages. J Clim 10:2548–2568

Jones, PD, Briffa KR, Barnett TP, Tett SFB (1998) High-resolution palaeoclimatic records for the last millennium: Integration, interpretation and comparison with general circulation model control run temperatures. Holocene 8:455–471

Juckes MN, Allen MR, Briffa KR, Esper J, Hegerl GC, Moberg A, Osborn TJ, Weber SL, Zorita E (2006) Millennial temperature reconstruction intercomparison and evaluation. Clim Past Discuss 2:1001–1049

Kalman R (1960) A new approach to linear filtering and prediction problems. J Basic Eng 82:34–45

Lee TCK (2008) On statistical approaches to climate change analysis. University of Victoria, PhD dissertation, to be available on https://dspace.library.uvic.ca:8443/dspace/

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 CM (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 CM (2007) Robustness of proxy-based climate field reconstruction methods. J Geophys Res (in press)

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

Osborn TJ, Raper SCB, Briffa KR (2006) Simulated climate change during the last 1,000 years: comparing the ECHO-G general circulation model with the MAGICC simple climate model. Clim Dyn 27:185–197

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

Schneider T (2001) Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J Clim 14:853–871

Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. J Time Ser Anal 3:253–264

Shumway RH, Stoffer DS (2000) Time series analysis and its applications. Springer, Heidelberg, 549 pp

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

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

Wahl ER, Ritson DM, Ammann CM (2006) Comment on "Reconstructing past climate from noisy data". Science 312:529b

Zorita E, von Storch H (2005) Methodical aspects of reconstructing non-local historical temperatures. Memorie Soc Astron Ital 76:794–801

Zorita E, Gonzalez-Rouco JF, Legutke S (2003) Testing the Mann et al. (1998) approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. J Clim 16:1378–1390