On Discriminating between GCM Forcing Configurations Using Bayesian Reconstructions of Late-Holocene Temperatures*

MARTIN TINGLEY,[+] PETER F. CRAIGMILE,[#] MURALI HARAN,[@] BO LI,[&] ELIZABETH MANNSHARDT,[**] AND BALA RAJARATNAM[++]

[+] *Department of Meteorology, and Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania*
[#] *Department of Statistics, The Ohio State University, Columbus, Ohio*
[@] *Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania*
[&] *Department of Statistics, University of Illinois at Urbana–Champaign, Urbana, Illinois*
[**] *Department of Statistics, North Carolina State University, Raleigh, North Carolina*
[++] *Department of Environmental Earth System Science, Department of Statistics, and the Woods Institute for the Environment, Stanford University, Stanford, California*

(Manuscript received 17 March 2015, in final form 8 July 2015)

ABSTRACT

Several climate modeling groups have recently generated ensembles of last-millennium climate simulations under different forcing scenarios. These experiments represent an ideal opportunity to establish the baseline feasibility of using proxy-based reconstructions of late-Holocene climate as out-of-calibration tests of the fidelity of the general circulation models used to project future climate. This paper develops a formal statistical model for assessing the agreement between members of an ensemble of climate simulations and the ensemble of possible climate histories produced from a hierarchical Bayesian climate reconstruction. As the internal variabilities of the simulated and reconstructed climate are decoupled from one another, the comparison is between the two latent, or unobserved, forced responses. Comparisons of the spatial average of a 600-yr high northern latitude temperature reconstruction to suites of last-millennium climate simulations from the GISS-E2 and CSIRO models, respectively, suggest that the proxy-based reconstructions are able to discriminate only between the crudest features of the simulations within each ensemble. Although one of the three volcanic forcing scenarios used in the GISS-E2 ensemble results in superior agreement with the reconstruction, no meaningful distinctions can be made between simulations performed with different estimates of solar forcing or land cover changes. In the case of the CSIRO model, sequentially adding orbital, greenhouse gas, solar, and volcanic forcings to the simulations generally improves overall consensus with the reconstruction, though the distinctions are not individually significant.

1. Introduction

Reconstructions of past climate from natural proxies provide a means for out-of-sample assessment of simulations from the general circulation models that are used to project future climate (Jansen et al. 2007; Masson-Delmotte et al. 2013). Recently, a number of modeling groups have produced ensembles of climate simulations covering the last millennium using different combinations or estimates of pre-instrumental greenhouse gas concentration, solar irradiance, volcanic forcing, and other important climate drivers. In what follows, we focus on last-millennium simulation ensembles produced with GISS Model E2 (GISS-E2; Schmidt et al. 2011, 2012) and CSIRO Mark 3L (CSIRO Mk3L; Phipps et al. 2013). These ensembles allow for a formal investigation of the discriminating power of proxy-based climate reconstructions in terms of their ability to select between simulations conducted under different forcing scenarios. We view this endeavor as a necessary precursory step before addressing the more challenging problem of using paleoclimate reconstructions to calibrate parameters of a single climate model run under a given forcing scenario that is then used to project future climate.

Ensemble-based paleoclimate reconstructions using hierarchical Bayesian methods are becoming more prevalent in the literature (e.g., Li et al. 2007; Tingley and Huybers

* Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/JCLI-D-15-0208.1.s1.

Corresponding author address: Martin Tingley, Dept. of Meteorology, The Pennsylvania State University, 510 Walker Bldg., University Park, PA 16802.
E-mail: martin.tingley@gmail.com

DOI: 10.1175/JCLI-D-15-0208.1

2010a,b; Li et al. 2010; Tingley et al. 2012; Tingley and Huybers 2013; Werner et al. 2013; Ahmed et al. 2013; Barboza et al. 2014). These reconstructions feature the robust uncertainty quantification required for making meaningful comparisons between simulated and reconstructed climate, and for assessing the viability of discerning between climate simulations produced using different configurations or estimates of past forcings. In what follows, we use the 600-yr temperature reconstruction of Tingley and Huybers (2013); the methodology we develop here, however, is readily applicable to other reconstructions produced by hierarchical Bayesian models.

Recent discussions and reviews of techniques for climate reconstruction–simulation comparisons can be found in Moberg (2013), Masson-Delmotte et al. (2013), and Schmidt et al. (2014). Previous efforts for the late Holocene have included qualitative comparisons of time series or maps (e.g., Mann et al. 2009; Kaufman et al. 2009), fuzzy logic (e.g., Guiot et al. 1999), simple distance-based metrics to select an optimal member from an ensemble of simulations (e.g., Goosse et al. 2006), data assimilation (e.g., Goosse et al. 2010), and statistical approaches based on a detection and attribution framework (e.g., Hegerl et al. 2007, 2011).

The ranking technique developed in Sundberg et al. (2012) and applied in Hind et al. (2012), Hind and Moberg (2013), and Moberg et al. (2015) points to the numerous challenges involved in developing a statistical formalism for selecting between climate model simulations based on proxy observations. In our own development, we follow Sundberg et al. (2012) in treating the problem as a regression between latent quantities: any common structure in the simulated and reconstructed climate is due to their shared response to an external forcing, with the unforced components of their respective variabilities independent of one another.

In what follows, we develop and fit a Bayesian hierarchical model to describe the conditional dependencies that link the simulated and reconstructed climates. In contrast to Sundberg et al. (2012), where the statistical model is used to motivate two frequentist hypothesis tests, we base conclusions on the posterior distributions of model parameters. As a second contrast to Sundberg et al. (2012), who base their analysis on point estimates of temperature from a classical calibration analysis of proxy series (Brown 1993; Christiansen 2014), we take as our starting point the ensemble of posterior draws resulting from a hierarchical Bayesian reconstruction. Since each posterior draw is, conditional on the observations and modeling assumptions, equally likely, our approach to linking the simulated and reconstructed climates explicitly takes into account the time-variable uncertainty in the reconstruction. Because of the increased interest in Bayesian reconstructions (Li et al. 2007; Tingley and Huybers 2010a,b; Li et al. 2010; Tingley et al. 2012; Tingley and Huybers 2013; Werner et al. 2013; Ahmed et al. 2013; Barboza et al. 2014), we anticipate that the tools developed here will be widely applicable for future reconstruction–simulation comparisons.

We first introduce the proxy-based reconstruction and climate simulations (section 2) and describe the Bayesian hierarchical model (section 3); we then present results of comparisons between the reconstruction and the GISS and CSIRO simulations, respectively (section 4). In section 5, we discuss extensions and connections to other areas, such as the model developed in Sundberg et al. (2012), Hind et al. (2012), and Moberg et al. (2015); data assimilation; detection and attribution (e.g., Allen and Stott 2003); and computer model calibration. Concluding remarks are provided in section 6. Details of the Markov chain Monte Carlo (MCMC) algorithm used to fit the Bayesian hierarchical model are provided in the supplemental material.

2. Climate reconstruction and model runs

a. *Climate reconstructions from hierarchical Bayesian models*

We use the paleoclimate reconstruction of April–September temperature anomalies described in detail in Tingley and Huybers (2013), based on tree-ring density, ice core, and varved lake sediment observations, along with the University of East Anglia Climatic Research Unit gridded instrumental temperature anomaly data product (CRUTEMv3; http://www.cru.uea.ac.uk/cru/data/temperature/). The reconstruction is performed using the hierarchical Bayesian algorithm for reconstructing climate anomalies in space and time (BARCAST; Tingley and Huybers 2010a,b), which assumes temperature anomalies are first-order autoregressive in time, with an exponentially decaying spatial covariance function. Instrumental observations are modeled as noisy versions of the true latent temperature anomalies, and each proxy type is separately modeled as linear in the true temperature anomalies with additive noise. Further details are available in Tingley and Huybers (2010a). The end product of the analysis is a large ensemble (here of size 4000) of posterior draws of the parameters that define the statistical model as well as the reconstructed temperature anomalies in space and time.

The reconstruction of Tingley and Huybers (2013), extending back to 1400 CE, is defined between 45° and 85°N on a 5° by 5° latitude–longitude grid and includes only those grid boxes that include a minimal fraction of

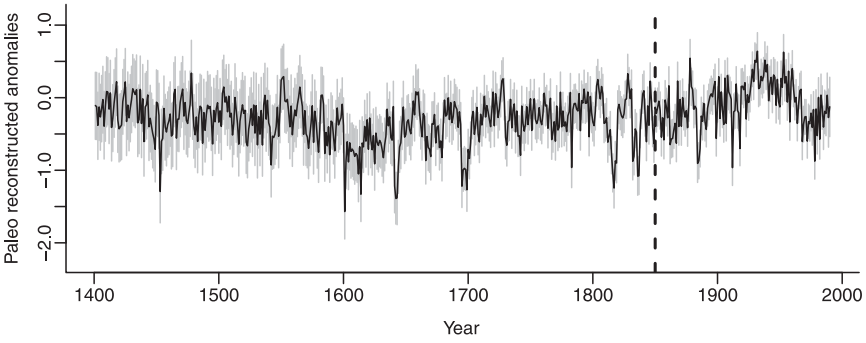

FIG. 1. The black line is a time series plot of the pointwise-in-time posterior means of April–September temperature anomalies from the proxy-based reconstruction. The light gray regions delimit the posterior 0.025 and 0.975 quantiles. The dashed vertical line at 1850 indicates when the GISS-E2 model runs end. The reconstruction uses only the proxy observations to predict past temperatures.

land. For the analysis considered here, we focus on the spatial-average time series, formed by weighting, at each year and for each posterior ensemble member, the grid boxes by the area of land they contain.

As our interest is in assessing the ability of the proxies, not the instrumental records, to discriminate between forcing scenarios, we restrict our usage of instrumental records to a minimum. Therefore, we primarily use results from running BARCAST in a reduced, or predictive, mode, as described in Tingley and Huybers (2013). In this predictive mode, only the proxies are used to predict past temperatures, and all scalar parameters are resampled from an earlier calibration analysis that employs both instrumental and proxy datasets. As discussed in Tingley et al. (2012) and Tingley and Huybers (2013), improved inferences on past climate are achieved if information from both the proxies and instruments are used for both prediction and parameter estimation. Indeed, if the goal of the analysis is to use all available information to select between forcing scenarios, then using reconstruction resulting from running BARCAST in predictive mode would not be appropriate.

The ensemble of spatially averaged reconstructions (Fig. 1) displays three noteworthy features: (i) there are temporal trends in the reconstructed series, (ii) the series is skewed toward lower values, likely because of influential volcanic events, and (iii) the uncertainty increases into the past.

b. *GISS-E2 simulations*

The ensemble of GISS-E2 simulations is part of phase 5 of CMIP (CMIP5)/PMIP phase 3 (PMIP3) ''last millennium'' experiment (Taylor et al. 2012; Braconnot et al. 2011) conducted for the IPCC Fifth Assessment Report. The simulations feature different forcings over the 850–1850 CE interval (Table 1). As the post-1850 forcings are common between the simulations, following the CMIP5/PMIP3 ''historical'' experiment specifications, we confine comparisons with the reconstructed climate to the 1400–1850 CE interval. The simulations employ common transient greenhouse gas and orbital forcings but different combinations of estimated volcanic forcing [none, CEA (Crowley et al. 2008), or GRA (Gao et al. 2008)], estimated solar forcing [SBF (Steinhilber et al. 2009) or VSK (Vieira et al. 2011)], and estimated anthropogenic changes to land use and land cover (LULC) [PEA (Pongratz et al. 2009) or KK10 (Kaplan et al. 2011)]; see Schmidt et al. (2011, 2012) for further details. From a statistical perspective, the design is an incomplete factorial experiment (see, e.g., Dean and Voss 1999) with no replication.

As discussed in the online documentation associated with the GISS-E2 last millennium simulations,[1] the volcanic forcing for the three simulations that employ the GRA estimate (p122, p125, and p128; see Table 1) was specified as a factor of 2 larger than intended. As our goal is to assess the ability of the proxy-based reconstruction to select between different forcing scenarios, this error represents an opportunity: any feasible simulation–reconstruction comparison should be able to identify the simulations under doubled GRA volcanic forcing as less reasonable than those featuring the CEA volcanic forcing. A complicating issue, however, is that the tree-ring densities used in the climate reconstruction may overestimate volcanic cooling following large eruptions (Tingley et al. 2014; Stine and Huybers 2014).

In addition to the forced simulations, the GISS-E2 ensemble includes a preindustrial control simulation, with volcanic forcing fixed at the 850–1999 average (according to CEA) and all other forcings fixed at 850 CE

[1] See http://data.giss.nasa.gov/modelE/ar5/ for discussion and data access.

TABLE 1. Summary of forcing configurations for the GISS-E2 last millennium simulations, as described in Schmidt et al. (2011, 2012). SBF (Steinhilber et al. 2009) and VSK (Vieira et al. 2011) correspond to two independent calibrations of cosmogenic isotope records to solar irradiance. CEA (Crowley et al. 2008) and GRA (Gao et al. 2008) correspond to two independent reconstructions of volcanic aerosol optical depth, both derived from ice cores. As noted by GISS (http://data.giss.nasa.gov/modelE/ar5/), the pre-1850 volcanic forcing used in simulations p122, p125, and p128 was approximately twice as large as intended from Gao et al. (2008). PEA (Pongratz et al. 2009) and KK10 (Kaplan et al. 2011) correspond to two independent estimates of anthropogenic changes to LULC prior to 1850. All simulations use the same orbital and transient greenhouse gas forcings.

Run ID	Solar forcing	Volcanic forcing	LULC
p121	SBF	CEA	PEA
p122	SBF	GRA	PEA
p123	SBF	None	PEA
p124	VSK	CEA	PEA
p125	VSK	GRA	KK10
p126	VSK	None	PEA
p127	VSK	CEA	KK10
p128	VSK	GRA	PEA

conditions (solar according to SBF and LULC according to PEA). The control run features low frequency variability, which we capture (see section 3) using a stationary time series model.

Figure 2 compares time series plots of temperatures from the nine GISS-E2 simulations with the posterior mean of the reconstructed temperatures, after removing the mean value from each simulation. The level shift between the GISS-E2 and reconstructed anomaly series occurs because the reconstructed temperatures do not have a mean of zero over the period 1400–1850. Further differences between simulated and reconstructed temperatures emerge when examining trends, dependence, and tail structure. Simulations that use CEA volcanic forcing look more similar to the reconstructed climate, while the GRA estimate of volcanic forcing leads to tails that are too long, consistent with misspecification of the magnitude of volcanic forcing.

c. *CSIRO simulations*

The ensemble of CSIRO simulations covers the 500–2000 interval and is based on a different experimental design than the GISS-E2 ensemble. A single control run, at fixed 1 CE forcing values, was used to initialize forced simulations under four different forcing configurations, adding in turn orbital (O), greenhouse (G), solar (S), and volcanic (V) forcing; see Phipps et al. (2013) for details. The solar forcing is from Steinhilber et al. (2009), and the volcanic forcing is derived from Gao et al. (2008); these correspond to the GRA volcanic and SBF solar forcing used in the GISS-E2 simulations. Three separate simulations are performed for each forcing configuration, differing only in the year of the control run used to initialize each, yielding a total of 13 simulations, including the control. As with the GISS-E2 ensemble, the additive experimental design is incomplete, in the sense that simulations are not performed with every possible combination of forcings. In contrast to the GISS-E2 ensemble, the CSIRO ensemble does not include different estimates of the same forcing. The information that can be learned by comparing the simulations to the reconstruction is therefore different for the CSIRO and GISS-E2 ensembles.

Time series plots comparing the ensemble averages of the CSIRO simulated temperature at the four forcing conditions and the control run, along with the posterior mean of the reconstructed temperatures, are shown in Fig. 3. Visually, the simulations that include orbital, greenhouse, solar, and volcanic forcing appear to be in better agreement with the reconstructed temperatures.

3. A hierarchical statistical modeling framework linking simulated and reconstructed climate

The statistical model linking the reconstructed and simulated spatially averaged temperature series is specified via a collection of conditional distributions. We first specify how the reconstruction relates to the unobserved, or latent, true temperature series, which in turn is modeled as the sum of forced and unforced components. In analogous fashion, we decompose the simulated-temperature time series into forced and unforced components. We specify the relationship between the reconstructed and simulated temperatures as a linear regression between their respective forced components. This hierarchical Bayesian modeling approach, combined with parameter inference via MCMC, permits for numerous sources of error to be accounted for and propagated throughout the analysis.

For clarity in the development of the statistical model, we explain how the reconstruction is related to an ensemble of simulations produced using a single forcing scenario. We then fit this statistical model separately for the climate simulation(s) created using each forcing scenario included in the GISS-E2 and CSIRO ensembles. Although the unknown parameters of the statistical model are therefore specific to the climate simulation used to fit the statistical model, we suppress this dependence to simplify notation. A summary of the model components is provided in Table 2.

a. *Modeling the reconstructed temperatures*

The temperature reconstruction takes the form of a random sample (ensemble) of spatially averaged reconstructed temperature anomaly time series given the proxy

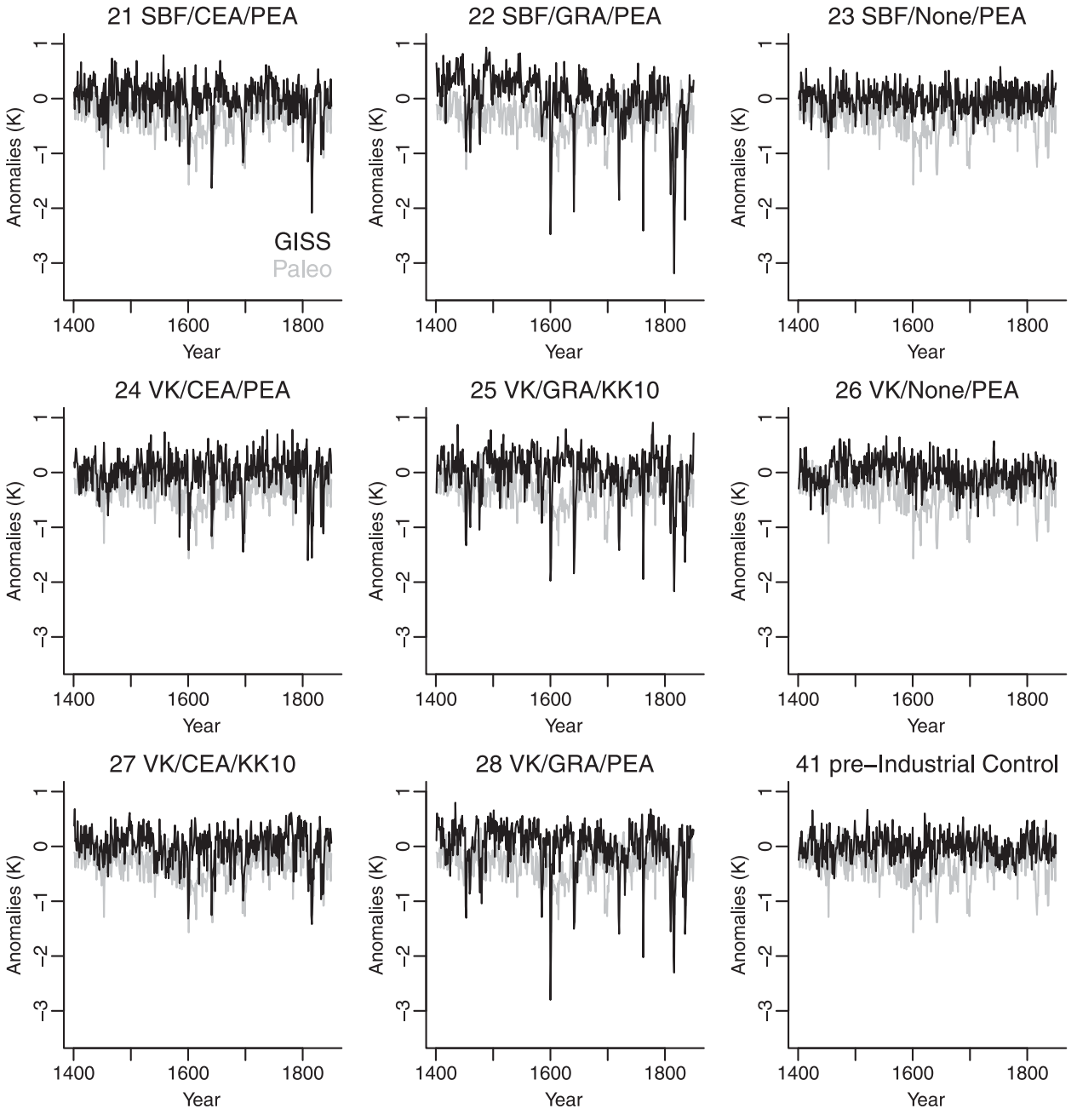

FIG. 2. Time series plots in black of the GISS-E2 simulated temperature anomalies run at eight different numbered forcing combinations (from 21 to 28) along with a control run (41). The temporal mean has been removed in each case. The posterior mean anomalies from the proxy-based reconstruction are in gray.

data and modeling assumptions. Denoting the kth such series by $\{X_{k,t}: t = 1, \ldots, T\}$, where $k = 1, \ldots, K$, a reasonable model for $X_{k,t}$ is

$$X_{k,t} \mid C_t \sim N(C_t, \lambda_t), \tag{1}$$

assuming conditional independence over k and t (i.e., given C_t, the draws of $X_{k,t}$ are independent over the different series and time indices), The series $\{C_t: t = 1, \ldots, T\}$ represents the underlying climate (here temperature anomalies) as inferred from the data and is never directly observed. The term $\lambda_t > 0$ captures the uncertainty in the reconstruction of the underlying climate and inherits time dependence from the time-varying availability of data.

The latent climate time series C_t is then modeled as

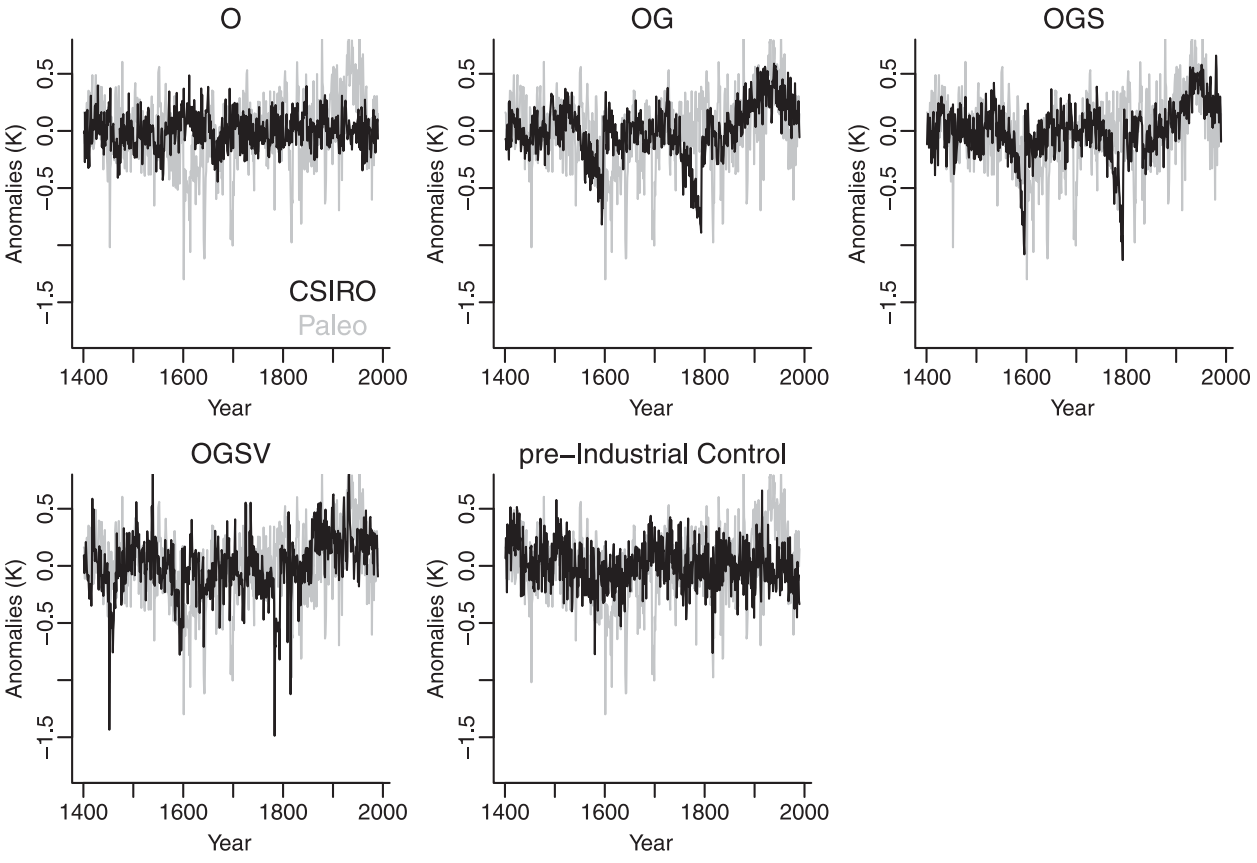

FIG. 3. Time series plots in black of the averaged-over-the-ensembles CSIRO simulated temperature anomalies run at four forcing combinations (O, OG, OGS, and OGSV) — orbital (O), greenhouse gas (G), solar (S), and volcanic (V)—including a control run. The posterior mean anomalies from the proxy-based reconstruction are in gray.

$$C_t = F_t^P + U_t^P, \quad t = 1, \ldots, T, \tag{2}$$

where $\{F_t^P : t = 1, \ldots, T\}$ and $\{U_t^P : t = 1, \ldots, T\}$ represent the forced and unforced, or internal variability, components of the latent temperature series, respectively, as inferred from the proxies (superscript P). To capture temporal dependence in the unforced time series, we assume that $\{U_t^P\}$ is a zero-mean first-order stationary autoregressive [AR(1)] process with autocorrelation parameter $-1 < \phi < 1$ and innovation variance $\sigma^2 > 0$.

For simplicity, we model the forced component as normally distributed and independent across time:

$$F_t^P \,|\, \delta, \kappa \sim N(\delta, \kappa^2).$$

Although this specification is a simplification, we note that the resulting posterior distribution, which is not constrained to be independent across time, still captures the time dependence in the forced component. We provide additional discussion of these issues in sections 4 and 5. To complete the specification of the Bayesian model, it is necessary to specify priors for the unknown parameters. Throughout, we assume that the prior parameters are mutually independent, with conjugate and weakly informative distributions (see, e.g., Gelman et al. 2003). A full list of the priors and hyperparameters is presented in Table 3.

b. *Modeling the simulated temperatures*

We model the spatially averaged simulated-temperature time series $\{Y_{j,t} : t = 1, \ldots, T\}$, where $j = 1, \ldots, J$ indices the simulations at a given forcing configuration ($J = 1$ for GISS-E2 and $J = 3$ for CSIRO), as

$$Y_{j,t} = F_t^M + U_{j,t}^M, \quad t = 1, \ldots, T. \tag{3}$$

As with the latent climate time series C_t, we decompose the simulation into two parts, with the latent time series $\{F_t^M : t = 1, \ldots, T\}$ and $\{U_{j,t}^M : j = 1, \ldots, J, t = 1, \ldots, T\}$ capturing the forced and unforced components of the simulated or modeled (superscript M) climate, respectively. We assume for each j that $\{U_{j,t}^M : t = 1, \ldots, T\}$

TABLE 2. Summary of the hierarchical Bayesian model. Throughout, t denotes the time index, j indices the simulations at a given forcing configuration, and k indices the reconstructed temperatures. Conditional independence is assumed in the distributions over indices.

Modeling the reconstructed temperatures: $X_{k,t} \mid C_t \sim N(C_t, \lambda_t)$	
$X_{k,t}$	The reconstructed temperatures. Observed.
C_t	The underlying climate, as inferred from the proxy data. Latent.
λ_t	The uncertainty in reconstructing the underlying climate. Latent.
Modeling the underlying climate inferred from proxy data: $C_t = F_t^P + U_t^P$	
F_t^P	The forced component of the underlying climate. Latent.
U_t^P	The unforced component (internal variability) of the underlying climate. Assumed to follow an AR(1) process over time. Latent.
Modeling the simulated temperatures: $Y_{j,t} = F_t^M + U_{j,t}^M$	
$Y_{j,t}$	The simulated temperature series. Observed.
F_t^M	The forced component of the simulated temperatures. Latent.
$U_{j,t}^M$	The unforced component (internal variability) of the simulated temperatures. Assumed to follow an AR(1) process over time, with parameters matching U_t^P. Latent.
Linking forced reconstructed and simulated temperatures: $F_t^M = \beta_0 + \beta_1 F_t^P + D_t$	
β_0, β_1	Regression intercept and slope relating the two forced series. Latent.
D_t	Captures the discrepancy between the two forced series modulo the regression. Assumed independent $N(0, \tau^2)$ over time. Latent.

is a zero-mean AR(1) process with autocorrelation parameter ϕ and innovation variance σ^2, with independence of the $\{U_{j,t}^M\}$ processes over the replicate j. We further assume that the parameters ϕ and σ^2 are common to both $\{U_{j,t}^M\}$ and $\{U_t^P\}$; that is, each of the $J+1$ realizations of internal variability are independent draws from a common statistical process. We take an empirical Bayesian approach to the parameters ϕ and σ^2, estimating their values using the control-run time series and then treating them as fixed. Assuming that the variability of the unforced simulation, or control run, is an adequate representation of the unforced variability of the climate is a common assumption in detection and attribution studies (e.g., Hegerl et al. 2000).

c. *Linking reconstructed and simulated temperatures*

The reconstructed (P) and simulated (M) temperature series are each decomposed into their respective forced and unforced components, and we link the two forced components via a linear-regression relationship:

$$F_t^M = \beta_0 + \beta_1 F_t^P + D_t. \tag{4}$$

Here $\{D_t : t = 1, \ldots, T\}$ captures the discrepancy between the two forced series modulo the regression relationship, and we assume that $D_t \sim N(0, \tau^2)$ with independence over time. This regression framework treats the unforced component of climate variability as an additional error, present in the observations of both independent and response variables. We specify an inverse gamma prior for τ^2 and a bivariate normal prior for $\beta \equiv (\beta_0, \beta_1)^{\mathrm{T}}$; hyperparameters are provided in Table 3. As the reconstruction is in anomaly units and the simulations in kelvins, the intercept β_0 corrects for differences in location and level shifts and is not otherwise interpretable.

The regression relationship [Eq. (4)] provides a link between the latent forced components of the reconstructed and simulated temperatures, and we use posterior distributions of the parameters β_1 and τ^2 to assess the agreement between the simulated and reconstructed climate. A value of unity for the slope β_1 indicates that the magnitude of the forced response is the same in both reconstruction and simulation. We interpret values of β_1 closer to one as indicative of better agreement between the simulation and reconstruction. The variance of the discrepancy term τ^2 is indicative of the magnitude of the mismatch between reconstructed and simulated temperatures, with lower values indicative of better agreement. We also introduce a variance fraction summary measure:

$$\upsilon_f = \frac{\mathrm{var}(\beta_1 F_t^P)}{\mathrm{var}(F_t^M)}. \tag{5}$$

We calculate the posterior distribution of υ_f using posterior draws of β_1, F_t^P, and F_t^M. The variance fraction gives the fraction of the variability in the forced component of the simulation (denominator) that can be explained, according to the regression relationship [Eq. (4)], by the forced response in the reconstruction (numerator). Larger values of υ_f are indicative of more preferable forcing configurations—the forced component

TABLE 3. Prior distributions and associated hyperparameters of the Bayesian model.

Parameter	Prior	Hyperparameters	Parameter	Prior	Hyperparameters
β	$N_2(m_\beta, V_\beta)$	$m_\beta = (0,0)^{\mathrm{T}}$, $V_\beta = 100I_2$	τ^2	IGa (s_τ, r_τ)	$s_\tau = 0.01$, $r_\tau = 0.01$
δ	$N(m_\delta, v_\delta)$	$m_\delta = 0$, $v_\delta = 100$	κ^2	IGa (s_κ, r_κ)	$s_\kappa = 0.01$, $r_\kappa = 0.01$
λ_t for each t	IGa (s_λ, r_λ)	$s_\lambda = 0.01$, $r_\lambda = 0.01$			

of the simulation is in closer agreement with the forced component of the reconstruction. We find below that larger values of the slope β_1, indicative of a stronger association between the forced components of the reconstruction and simulation, are often accompanied by larger values of the discrepancy variance τ^2, and the variance fraction v_f provides a useful means of balancing these competing results into a single assessment of simulation–reconstruction agreement.

The nature of the regression relationship in Eq. (4) suggests that obtaining accurate, stable inference concerning the relationship between the simulated and reconstructed temperatures is likely to be difficult. Both the independent and dependent variables are latent quantities, in the sense that they are not observed directly. The observed climate simulations are the sum of a forced response F^M and an unforced component. In the case of the reconstructions, the forced response appears even deeper in the model: the observed $X_{k,t}$ depends on C_t, which itself is the sum of the forced component F_t^P and an unforced component. The addition of the unforced components, which contribute a substantial fraction of the variability to the observed series, results in a difficult inference problem, despite our assumption that both the stochastic structure and parameters of the unforced component are known.

d. *Inference based on the posterior distribution*

For both the CSIRO and GISS-E2 simulations, we fit the statistical model separately for each forcing configuration. For the CSIRO simulations, we fit the model in two different ways, either separately for each of the three simulations under each forcing configuration or by pooling information across each three-member ensemble. To simplify notation, let $Y_j = (Y_{j,t}: t = 1, \ldots, T)^{\mathrm{T}}$ denote simulation j under a specific forcing configuration, $Y = (Y_1, \ldots, Y_J)^{\mathrm{T}}$, and $F^M = (F_1^M, \ldots, F_T^M)^{\mathrm{T}}$. For the reconstruction, let $X_k = (X_{k,1}, \ldots, X_{k,T})^{\mathrm{T}}$, with $X = (X_1, \ldots, X_K)^{\mathrm{T}}$. Finally, let $C = (C_1, \ldots, C_T)^{\mathrm{T}}$ denote the latent temperature vector implicit to the reconstruction, and let $F^P = (F_1^P, \ldots, F_T^P)^{\mathrm{T}}$ denote the forced component of the latent temperature series. For each forcing configuration r, the parameters that must be inferred are then $\theta_r = (C, F^M, F^P, \beta, \tau^2, \delta, \kappa^2, \{\lambda_t\})$. Although each element of θ_r implicitly depends on the particular forcing configuration r used to create the simulation, we suppress this dependence to simplify the notation. The term $\pi(\theta_r \mid Y, X)$, the posterior distribution of the parameter vector θ_r given the observations (the simulations Y under forcing configuration r and reconstruction X), is proportional to

$$
\begin{aligned}
&f(Y \mid F^M) f(X \mid C, \{\lambda_t\}) f(C \mid F^P) f(F^M \mid F^P, \beta, \tau^2) f(F^P \mid \delta, \kappa^2) \pi(\beta) \pi(\tau^2) \pi(\delta) \pi(\kappa^2) \left[\prod_{t=1}^{n} \pi(\lambda_t) \right] \\
&\quad = \left[\prod_{j=1}^{J} f(Y_j \mid F^M) \right] \left[\prod_{k=1}^{K} f(X_k \mid C, \{\lambda_t\}) \right] \left[\prod_{t=1}^{T} f(C_t \mid F_t^P) \right] f(F^M \mid F^P, \beta, \tau^2) f(F^P \mid \delta, \kappa^2) \\
&\qquad \times \pi(\beta) \pi(\tau^2) \pi(\delta) \pi(\kappa^2) \left[\prod_{t=1}^{n} \pi(\lambda_t) \right].
\end{aligned}
$$

As the posterior density is not available in closed form, we sample from it using an MCMC algorithm (e.g., Gelman et al. 2003); further details of the algorithm are provided in the supplemental material. We ran multiple Markov chains with different initial values. Our MCMC estimates were very similar across these multiple chains, which gives us confidence that we have run the chain for long enough to eliminate potential biases due to initial values. We also checked autocorrelation plots to diagnose problems due to slow mixing in our Markov chain. We are satisfied that the autocorrelations, particularly after thinning, are sufficiently small, and therefore the variability (MCMC standard errors) of our estimates is also sufficiently small. Results below are based in each instance on 2000 samples from the posterior (for each of two chains, 10 000 further draws thinned by a factor of 10 after a burn-in of 1000 samples).

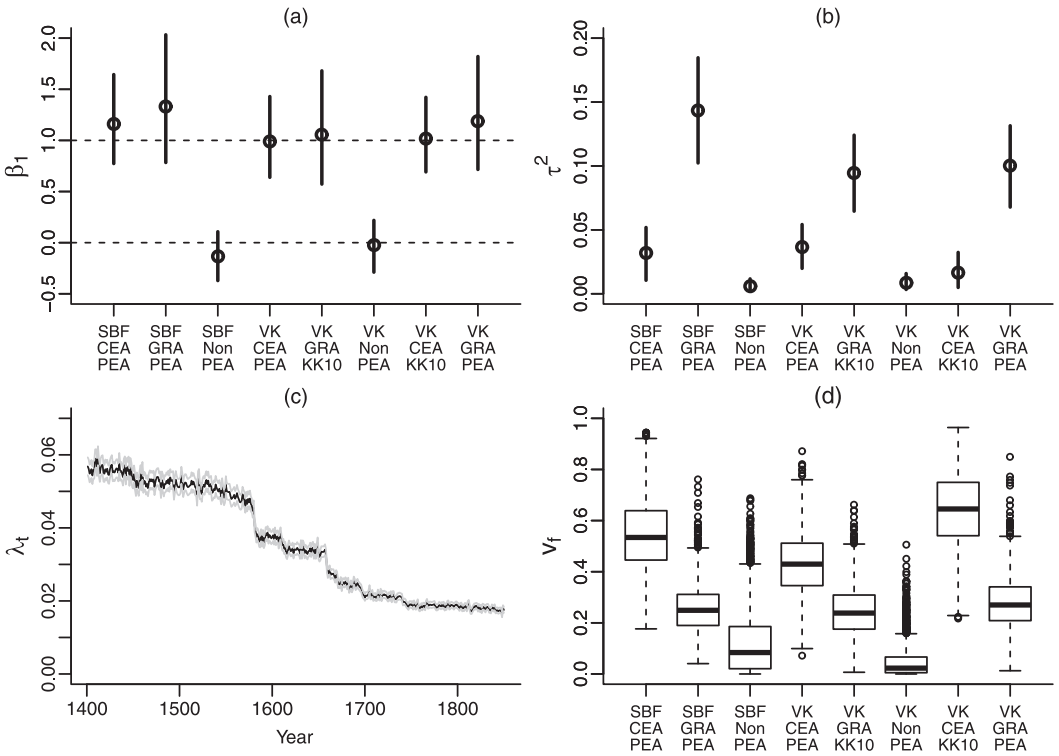

FIG. 4. Posterior summaries for the hierarchical Bayesian models fit to the GISS-E2 simulated temperatures at each of the eight forcing scenarios (black circles); the vertical lines indicate estimated 95% credible intervals for each parameter. (a) Posterior means of the slope parameter β_1; the horizontal dashed lines indicate β_1 values of 0 and 1. (b) As in (a), but for the variance of the discrepancy term τ^2. (c) The posterior mean of λ_t vs t for the first forcing scenario is shown by the black lines; the gray lines denote pointwise credible intervals. (d) Box-and-whisker plots summarizing posterior distributions of the variance fractions. The box extends from the first to the third quartile, with the median marked by a thick horizontal line. Whiskers extend to the most extreme values within 1.5 interquartile ranges of the box edges, and all more extreme values are plotted as circles.

4. Results

a. Results for GISS-E2 simulations

As the GISS-E2 last millennium experiments meld into the commonly forced historical runs after 1850, we limit comparisons with the reconstructed climate to the 1400–1850 interval. An AR(1) process appears to be a reasonable model for the GISS-E2 control run (Box–Ljung test for residuals; 30 lags; p value = 0.47), and the estimated autocorrelation parameter and innovation variance are $\hat{\phi} = 0.15$ [standard error (SE) = 0.05] and $\hat{\sigma}^2 = 0.05$. Using these empirical estimates, we fit the Bayesian model (section 3) to each of the eight forced GISS-E2 simulations (Fig. 4; Table 4).

The year-specific error variance in the proxy reconstruction $\{\lambda_t\}$ is determined largely by the proxy-based reconstructions and is therefore similar for all forcing configurations. In general, $\{\lambda_t\}$ shows a decrease toward the present, as more proxies are available in recent centuries (Fig. 4c). The large drops in variance correspond to stepwise changes in data availability [see, e.g., Fig. S1 of Tingley and Huybers (2013)].

For each of the two forcing scenarios that exclude volcanic forcing, the posterior distribution of the slope parameter β_1 includes zero, indicating little or no agreement between these forced simulations and the reconstructions (Fig. 4a). The corresponding posterior distributions of the discrepancy variances τ^2 are sharply peaked at near-zero values (Fig. 4b), indicating that the simulations without volcanic forcing have statistical distributions similar to that of the control run. This observation follows from combining Eqs. (3) and (4) and expressing each simulated temperature series $Y_{j,t}$ as

$$Y_{j,t} = \beta_0 + \beta_1 F_t^P + D_t + U_{j,t}^M \quad t = 1, \ldots, T. \qquad (6)$$

When both β_1 and τ^2, the variance of D_t, are close to zero, the simulated temperatures exhibit statistical properties matching those of the control-run process. Posterior distributions of the variance fraction v_f are likewise close to

TABLE 4. Posterior means and 95% credible intervals (in parentheses) for a selection of parameters in the hierarchical Bayesian model, for fits to the eight GISS-E2 simulations (for forcing configuration, see Table 1). Rows in boldface indicate posterior credible intervals for the slope parameter β_1 that do not include zero.

Model	β_1	τ^2	v_f
21 SBF/CEA/PEA	**1.160 (0.773, 1.644)**	**0.032 (0.011, 0.052)**	**0.545 (0.291, 0.842)**
22 SBF/GRA/PEA	**1.331 (0.784, 2.033)**	**0.143 (0.102, 0.185)**	**0.257 (0.095, 0.466)**
23 SBF/None/PEA	−0.134 (−0.371, 0.107)	0.006 (0.002, 0.012)	0.126 (0.000, 0.477)
24 VSK/CEA/PEA	**0.989 (0.639, 1.429)**	**0.037 (0.020, 0.054)**	**0.431 (0.206, 0.674)**
25 VSK/GRA/KK10	**1.055 (0.573, 1.680)**	**0.094 (0.065, 0.124)**	**0.248 (0.080, 0.477)**
26 VSK/None/PEA	−0.023 (−0.289, 0.218)	0.009 (0.003, 0.016)	0.048 (0.000, 0.240)
27 VSK/CEA/KK10	**1.019 (0.692, 1.421)**	**0.017 (0.005, 0.032)**	**0.642 (0.342, 0.901)**
28 VSK/GRA/PEA	**1.188 (0.716, 1.821)**	**0.100 (0.068, 0.131)**	**0.281 (0.116, 0.508)**

zero, further confirming little agreement between these simulations and the reconstructed temperatures.

For each of the six GISS-E2 simulations that includes an estimate of volcanic forcing, the posterior distribution of the slope parameter covers unity, indicating broad agreement between the simulations and the reconstructions (Fig. 4a). The large posterior variance of β_1 for simulations that include the GRA forcing indicates an uncertain link between F_t^M and F_t^P for these forcing scenarios.

The three GISS-E2 simulations with the GRA volcanic forcing feature large discrepancy variances and low variance fractions, indicating disagreement between the simulations and the reconstruction, despite the reasonable slope parameters. The three GISS-E2 simulations that use the CEA volcanic forcing are in better agreement with the reconstructed climate, as evidenced by lower discrepancy variances and higher variance fractions. As the GRA forcing used in the GISS-E2 simulations was double the intended value (section 2b), the better agreement between the reconstruction and the simulations that include the CEA (as opposed to GRA) volcanic forcing is to be expected. Results therefore confirm the ability of our hierarchical Bayesian modeling approach to correctly identify a set of forcing scenarios as unreasonable.

The impact of different choices of solar forcing and LULC are harder to detect (Fig. 4). The simulation with VSK solar, CEA volcanic, and KK10 LULC features the smallest discrepancy variance and the highest posterior mean variance fraction, suggesting that it features the closest agreement with the reconstruction. Of the three simulations that feature CEA volcanic forcing, that with VSK solar and PEA LULC features the lowest posterior mean variance fraction, suggesting that it is in poorest agreement with the reconstruction.

We briefly discuss (but do not plot) the results of two additional versions of the analysis. In comparing results under these different analysis choices, it is important to note that there are at least four potential sources of uncertainty: the climate models, the estimated forcing series, the proxy records, and the statistical framework that links these elements. As these sources of uncertainty cannot be readily disentangled from one another, it is possible that comparative results may not lend themselves to clear interpretations.

1) *Excluding volcanic years.* In this set of experiments, years featuring volcanic eruptions with volcanic explosivity indices greater than six (taken from Simkin and Siebert 1994) were excluded from the analysis, along with the two subsequent years. These volcanic years have a disproportionately large impact on the distribution of temperatures, for both the simulations and reconstructions, so it is worthwhile to explore the agreement between simulations and reconstruction after they are excluded. Furthermore, there is evidence that the tree-ring records included in the reconstruction do not correctly capture volcanic cooling (e.g., Stine and Huybers 2014; Tingley et al. 2014). In general, the resulting slope parameters are reduced as compared with the analysis that includes all years, with a larger magnitude of decrease for the GRA forcing. For those simulations that include volcanic forcing, excluding the major volcanic events reduces the posterior means of the variance fractions by between 0.08 and 0.21 units. The pattern of change is similar for the discrepancy variances. Results suggest that the response to major volcanic events is a significant source of variation in the forced component of the GISS-E2 simulations and that a similar behavior is evident in the reconstructed climate. As before, the variance fractions for the simulations that employ the CEA estimate of volcanism, as opposed to the GRA estimate, are higher, indicating better agreement with the reconstruction.
2) *Decadal averages.* We reran the analysis using decadal averages of the simulations, control run, and reconstructions, estimating the parameters ϕ and σ^2 in the AR(1) process using the decadal averaged control run. This analysis yields marginally smaller mean values of the slope parameter β_1, discrepancy

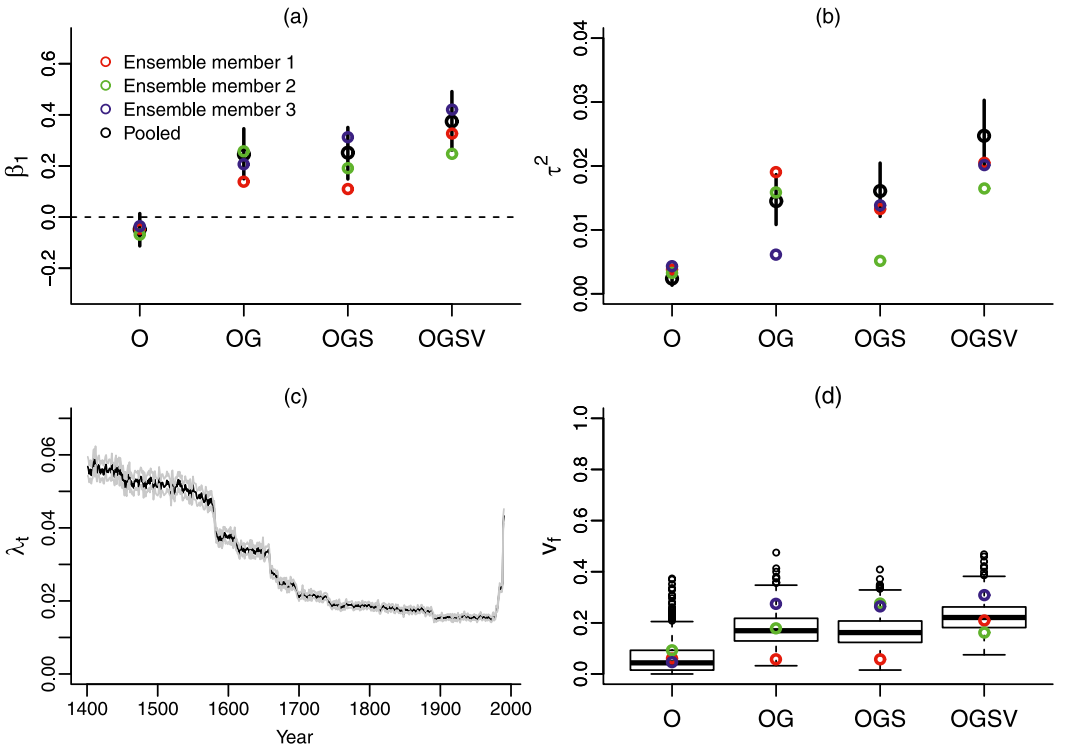

FIG. 5. As in Fig. 4, but summarizing the hierarchical Bayesian models fit to the CSIRO simulated temperatures at each of the four forcing scenarios. The comparison is to the reconstruction that predicts past temperatures using only the proxy observations over the 1400–1990 interval. (a),(b),(d) The colored circles correspond to posterior means from separate analyses of each of the three ensemble members at each forcing level, while the black circles show results from pooling information across the three ensemble members.

variance τ^2, and time-varying variance λ_t, though the posterior distributions feature substantial overlaps with the annual-scale results. The latter two changes are obvious effects of smoothing out the yearly time series. Variance fractions are generally in the same range as those for the annual results, but the posterior distributions for the different simulations that feature the CEA volcanic forcing feature greater overlap.

b. *Results for CSIRO*

The ensemble of CSIRO simulations has two advantages relative to the GISS-E2 simulations. First, the simulations extend through the twentieth century under different forcings. Second, there are three simulations for each forcing scenario. We therefore fit the model in two different ways, either separately for each of the three simulations under each forcing configuration or by pooling information across these three-member ensembles. Pooling information results in a tighter inference of model parameters, while separate fits allow for an exploration of the variability of the results between simulations under the same forcing configuration.

Results are based once more on assuming an AR(1) model for the control runs. The estimated autocorrelation parameter is $\hat{\phi} = 0.21$ (SE = 0.024), and the estimated innovation variance is $\hat{\sigma}^2 = 0.04$. We note, however, that the CSIRO control run exhibits a longer range of dependence in the sense that higher order AR coefficients may be significant. The dependence structure in the CSIRO simulations is more involved than that found in the GISS-E2 simulations, and as a stationary AR(1) model cannot capture the observed dependence, more caution is required when interpreting results.

We first consider results based on the climate reconstruction that employs only the proxy observations over the 1400–1990 interval (Fig. 5; Table 5). For all forcing configurations, the time series of year-specific variances λ_t are similar to one another (Fig. 5c) and show the same progression as for the GISS-E2 simulations during their period of overlap (Fig. 4c). This agreement is to be expected as the variances λ_t are primarily a feature of the reconstruction, which is the same across the various analyses. The proxy-based reconstruction extends to 1990, but owing to a paucity of data, the uncertainty increases over the last decade (Fig. 5c).

When pooling information across the three simulations at each forcing configuration, the slope β_1 increases as additional forcings are included in the simulations

TABLE 5. Posterior means and 95% credible intervals (in parentheses) for a selection of parameters in the hierarchical Bayesian model, for fits to the four ensembles of CSIRO simulations, each with a different forcing configuration. Rows in boldface indicate posterior credible intervals for the slope parameter β_1 that do not include zero.

Model	β_1	τ^2	v_f
O	−0.047 (−0.113, 0.014)	0.002 (0.001, 0.004)	0.062 (0.000, 0.223)
OG	**0.246 (0.150, 0.346)**	**0.015 (0.011, 0.019)**	**0.176 (0.065, 0.317)**
OGS	**0.252 (0.149, 0.351)**	**0.016 (0.012, 0.020)**	**0.166 (0.061, 0.289)**
OGSV	**0.374 (0.262, 0.491)**	**0.025 (0.020, 0.030)**	**0.224 (0.111, 0.346)**

(Fig. 5a; Table 5). The uncertainty in the slopes is smaller for the CSIRO simulations than for the GISS-E2 simulations, and in contrast to results for the GISS-E2 simulation, the posterior credible intervals for the slopes do not cover unity. For the simulations that include only orbital forcings, the 95% posterior credible interval for β_1 covers zero. For all three simulations including at least the greenhouse gas forcing, the 95% posterior credible intervals for each slope parameter are greater than zero. The addition of greenhouse gas and volcanic forcings results in substantial increases in β_1, whereas the inclusion of the solar forcing has a much smaller effect. The presence of a common signal between the reconstructions and simulations becomes more apparent when greenhouse gas and volcanic forcings are included.

The discrepancy variances τ^2 follow the same pattern as the slopes, with a near-zero value for the orbital-only forcing configuration and significant increases with the inclusion of greenhouse gas and volcanic forcings (Fig. 5b). The slopes and discrepancy variances generally increase in tandem with one another, making it difficult to use them to compare the agreement between the reconstructed climate and the various simulations; that is, as the linear relationship between the simulated and reconstructed forced response becomes steeper, there is generally more spread about the line of best fit. The variance fraction v_f is a useful metric for comparing results in this case, as it combines the competing effects of greater variability about a stronger slope into a single measure.

The variance fractions v_f generally increase when additional forcings are included (Fig. 5d), indicating that a progressively larger fraction of the variability in the simulated forced response can be explained by linear dependence on the forced response from the reconstruction. The posterior distributions of variance fractions feature substantial overlaps and are generally smaller than for the GISS-E2 simulations.

Overall, numerical summaries suggest that the agreement with the reconstructions is weaker for the CSIRO simulations as compared with the GISS-E2 simulations. A possible explanation is the different time spans covered by the two comparisons in combination with the post-1850 interval excluded from the GISS-E2 analysis featuring a prominent trend in temperatures. Another possible explanation is the stronger time series dependence observed in the CSIRO control run leading to more residual dependence in the discrepancies.

There is considerable variability in results when fitting the Bayesian model separately for each of the three simulations at each forcing level (Fig. 5). For all simulations, the posterior mean of both the slope β_1 and discrepancy variance τ^2 are smallest for the orbital-only simulations and are largest when all forcings are included. Results for the two intermediate forcings scenarios are unclear, which is to be expected given that, when pooling across the three simulations, the corresponding posterior distributions are so similar. The variance fractions v_f likewise feature variability between the three simulations, with generally smaller values for the orbital-only simulation and generally larger values when all forcings are included. The size, and even the direction, of changes in v_f as additional forcings are included, however, varies between the three ensemble members. The variability of results across the three ensemble members points to the value of running numerous simulations at each forcing configuration in this type of experiment, as conclusions based on a single simulation may not be robust.

Several variations of the analysis choices are worth discussing (Fig. 6). In the same manner as for the GISS-E2 simulations, we first exclude years that may be strongly affected by volcanism and, separately, decadal-average both the simulations and reconstructions. To avoid any potential issues associated with the late twentieth-century divergence of tree-ring density records from temperatures [for further discussion, see Tingley and Huybers (2013)], we perform separate fits after curtailing the proxy-based reconstruction at 1960 (labeled P6 on the x axis in Fig. 6). Finally, we fit the Bayesian model using reconstructions that use both the proxy and instrumental datasets to predict past temperatures, ending the reconstruction in either 1960 or 1990 (labeled I6 and I9, respectively, in Fig. 6). As there are four forcing combinations, these variations result in a total of 48 model fits (Fig. 6); we briefly summarize a few patterns that emerge from these exercises, noting that the same caveats apply as with the GISS-E2 analysis.

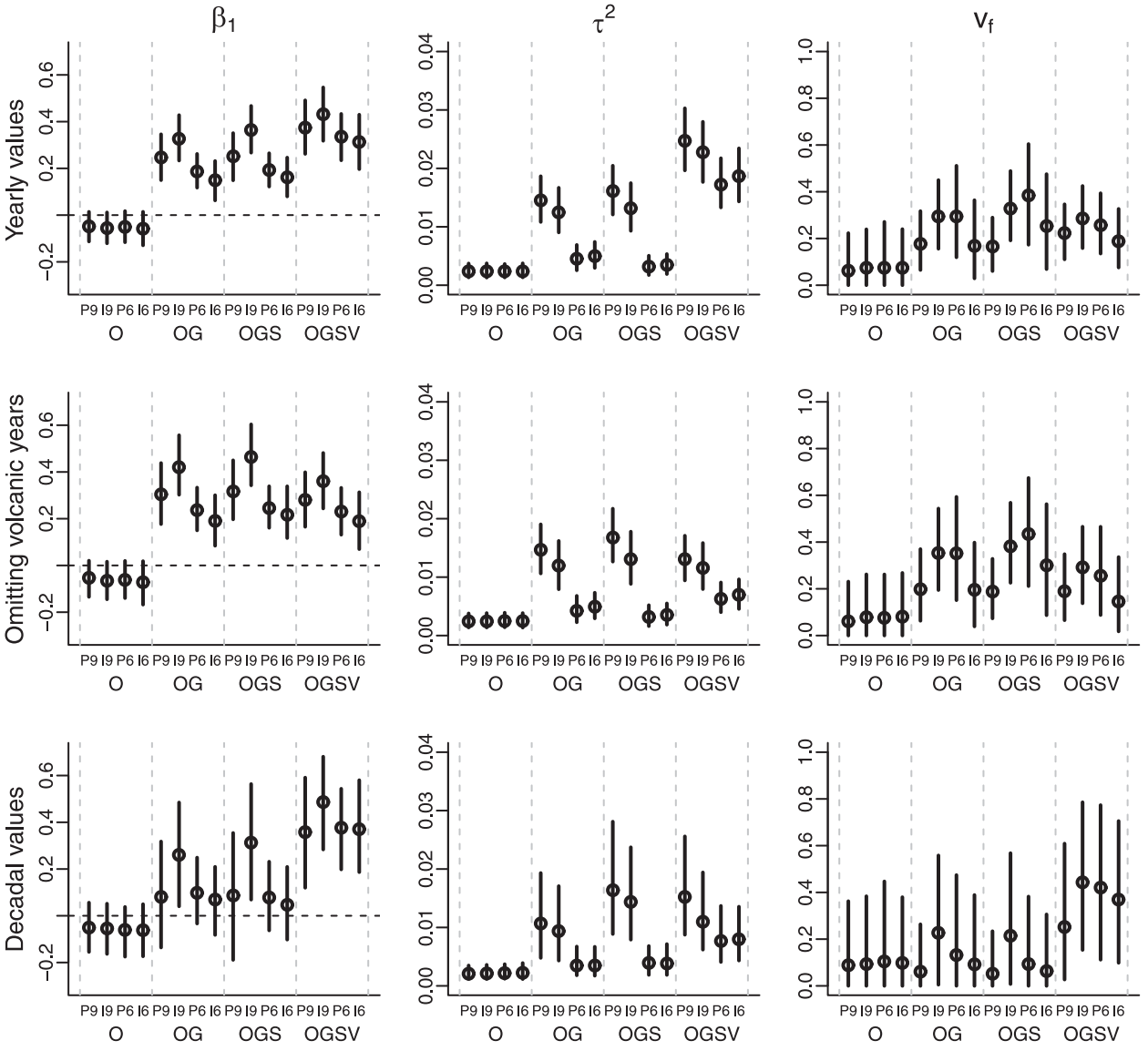

FIG. 6. Summary of results from applying the Bayesian model to the CSIRO simulations and the reconstructions in a number of ways: (left)–(right) β_1, τ^2, and v_f and (top)–(bottom) analysis on all annual values, after omitting the volcanic years, and after decadal averaging. Within each panel, O, OG, OGS, and OGSV indicate the forcings that are included in the simulations. For each forcing configuration, P9, I9, P6, and I6 indicate the data that are included in the reconstruction (P = proxy only; I = proxy and instruments) and the end point of the comparison (6 for 1960 and 9 for 1990).

1) *Excluding volcanic years.* The most notable change is a decrease in the slope parameter β_1 for the simulations that include volcanic forcing. Excluding years that are strongly affected by volcanism decreases the strength of linear association between the reconstructions and the simulations, and results for simulations that include volcanic forcing move toward the results for simulations that exclude volcanic forcing (Fig. 6). This feature conforms to intuition: if there is good agreement between the simulated and reconstructed responses to major eruptions, then removing those years should decrease the agreement between them. Results therefore suggest that the hierarchical Bayesian model is capable of identifying important aspects of agreement, or disagreement, between the reconstructed and simulated climates.
2) *Decadal averages.* Rerunning the analysis using decadal averages has a similar effect on the slopes β_1, discrepancy variances τ^2, and time-varying variances λ_t as seen for the analysis of the GISS-E2 simulations. Most notably, the posterior means of the variance fractions are generally smaller as compared

with the annual-scale analysis, and the posterior distributions feature greater variability. We interpret these features as indicating that the agreement between the CSIRO simulations and the reconstructions is better at shorter time scales.

3) *Proxy-based reconstruction, 1400–1960.* As compared with the analysis over 1400–1990, the slope parameters are moderately smaller and there is a noticeable decrease in the discrepancy variances, particularly for the simulations that exclude volcanic forcing. According to the variance fraction, the simulations that include orbital, greenhouse gas, and solar forcing now feature, on the annual scale, the best agreement with the reconstruction. On the annual scale, the subsequent addition of volcanic forcing reduces the posterior mean of the variance fraction, though the posterior distributions feature considerable overlap. On the decadal scale, the addition of the volcanic forcing increase the variance fraction.
4) *Instrumental and proxy-based reconstruction, 1400–1990 and 1400–1960.* For analyses that extend to 1990 (I90 in Fig. 6), there are small increases in the slopes, as compared with the proxy-only reconstruction (P9), and generally small decreases in the discrepancy variances. There are also substantial increases in the variance fraction for all simulations that include greenhouse gas forcings. As the post-1960 interval features rapid rises in both greenhouse gas concentrations and temperatures as well as possible divergence between tree-ring densities and temperatures, the increased agreement that results from including the instruments in the reconstruction is expected.

   Including the instruments does not have the same effect for analyses that extend only to 1960. Simulations that include greenhouse gas forcing then feature, in terms of posterior means, decreases in the slope, increases in the discrepancy variances, and decreases in the variance fractions as compared with the proxy-only reconstructions. However, there are substantial overlaps in the posterior distributions of the three diagnostics (β_1, τ^2, and υ_f), so we caution against over-interpreting this result.

5. Extensions and connections with other methods

a. Possible extensions

The statistical framework we develop here can be extended or generalized in a number of ways. The prior specification of temporal independence for the forced response is likely not appropriate, as the major forcing types—greenhouse gas, volcanic, and solar—each have characteristic time series properties. Volcanic forcing is intermittent and features a strong negative skew, solar forcing features smooth centennial-scale variability, and greenhouse gas forcing is nearly monotonic over the 1400–1990 interval. Including prior information about the differing time series properties of the forcings may improve inference. One possible approach would be to further decompose F_t^M and F_t^P into individual forced responses, each with a different time series model. It would then be possible to study the agreement between each component of F_t^M and F_t^P. Although more accurate statistical models can yield better inferences, a more involved statistical model could also complicate the interpretation of the discrepancy between the reconstructions and the simulations. Another possible extension could involve time-scale decompositions (e.g., using a wavelet transformation; Percival and Walden 2000) of both the simulations and reconstructions to discriminate between the series over different temporal scales.

Here we have focused on large-scale spatial-average temperatures. As both the simulations and reconstructions provide spatially complete temperature fields, spatial averaging can be at any scale, and the static model could then be applied regionally or even locally. Note that as averaging areas become smaller, the ratio of forced to internal variability (akin to a signal-to-noise ratio) becomes larger, and the inference on the regression parameters linking the forced and simulated climates will become more uncertain. Alternatively, the analysis could be generalized to include a spatial component so that F^P and F^M are modeled as space–time processes. Some prior information about the space–time covariance of the forced response will be necessary to achieve inference with a reasonable assessment of uncertainty. Any spatial extension may require developing more nuanced measures of simulation–reconstruction discrepancy, as some simulations may feature spatial patterns that are broadly correct but perhaps shifted or rotated. In this regard, see Heaton et al. (2015) for an example of a model calibration analysis that allows for spatial realignment.

b. Comparison with other studies

Our development shares similarities with the model presented in Sundberg et al. (2012), applied in Hind et al. (2012) and Hind and Moberg (2013), and generalized in Moberg et al. (2015), which also casts the problem as a regression between latent quantities [cf. statistical model 1 from Sundberg et al. (2012) to Eq. (4)]. However, Sundberg et al. (2012) do not estimate the parameters of their statistical model [see discussion

in section 9 of Sundberg et al. (2012)] but rather define tests of the significance of a correlation and distance metric to assess model–reconstruction agreement. Moreover, the distance metric is interpretable only if the correlation measure is significant. In contrast, our approach provides interpretable results for all diagnostics (β_1, τ^2, and υ_f) even if, as is the case for the GISS-E2 simulations that exclude volcanic forcing, there is no apparent shared variability between the simulation and reconstruction.

The statistic T in Sundberg et al. (2012) is used to test if the squared difference between reconstructed climate and a forced simulation is significant with respect to a null that assumes the forced simulation is equivalent to an unforced simulation. In contrast, the τ^2 variance parameter we use above measures the average squared distance between the forced components of the simulation and reconstruction, and we use τ^2 to compare the agreement between a single reconstruction and a number of forced simulations. Moberg et al. (2015) generalize the model of Sundberg et al. (2012) to permit autoregressive time series dependence in the unforced simulations. In contrast to our approach of fitting the parameters of a latent regression model assuming correlated errors, Moberg et al. (2015) adjust the variances of test statistics introduced in Sundberg et al. (2012) to account for the assumed serial correlation.

Finally, Moberg et al. (2015) apply their model to 15 tree-ring records and an ensemble of climate simulations comprising single-forcing configurations (either solar, volcanic, or land cover change, with constant preindustrial greenhouse gas forcing) and a multiple-forcing configuration (solar, volcanic, land cover change, orbital, greenhouse gas, and nonvolcanic aerosol forcing) conducted using two separate amplitudes of solar forcing. In agreement with our findings from the CSIRO analysis, Moberg et al. (2015) find evidence that simulations that include multiple forcing feature greater agreement with the information from the proxies than those that include only a single forcing. In agreement with our findings from the GISS-E2 analysis, Moberg et al. (2015) find it difficult to discern between the two amplitudes of solar forcing they consider. We refrain from more detailed comparisons as both the climate simulations and proxy information differ between the two studies.

c. Climate reconstruction via data assimilation

A related research problem concerns the optimal estimation of past climate from all available sources: climate models, proxies, and the instrumental record. Each of these data sources is inadequate in some way, be it due to a short time span (instrumental record), a difficulty in quantifying relationships with latent climate processes (proxies), or an imperfect representation of physical climate processes (climate models). In this article we have limited ourselves to a one-way flow of information, whereby the proxy-based reconstruction informs a basis for selection between climate simulations. Allowing information from the simulations to adjust the proxy-based reconstruction would require a statistical framework that links each source of information to a latent climate process, and the estimation of that process would then mix across these sources of information. The utility of such an estimate is limited, however, as the mixing of information from the simulations and reconstructions precludes using the resulting climate estimate to assess model performance. A related line of research involves applying data assimilation techniques to reconstruct past climate (e.g., Steiger et al. 2014).

d. Detection and attribution

There are connections between the statistical model we use here and the optimal fingerprinting framework used in detection and attribution (D&A) studies (e.g., Hegerl et al. 2000). D&A studies generally assume that the forced response in the observations is linear in a number of patterns, or fingerprints, each corresponding to a particular forcing. The detection step then proceeds by determining whether estimates of coefficients in the regression of the observed climate onto the fingerprints are significantly greater than zero. Note that the direction of the regression is opposite to that used here, as we regress the simulated forced response on the reconstructed forced response. In contrast to D&A, our current interest is not in detecting the effects of particular forcings on the observed or reconstructed climate but rather in determining if the proxy-based reconstructions are able to select in some fashion between simulations conducted under different forcing scenarios. Further discussion of these issues is available in Sundberg et al. (2012).

e. Computer model calibration

Computer model calibration is an active area of research (e.g., Forest et al. 2002; Sansó et al. 2008; Bhat et al. 2012; Chang et al. 2014, in the context of climate models) that shares characteristics with the simulation–reconstruction comparisons developed here. Computer model calibration focuses on learning about unknown parameters of a complex computer model. These parameters are inputs to the computer model that affect the output produced by the model. In the context of climate models the parameters may represent, for instance, constants that describe the dynamics of the climate system or some approximate representations of

aggregate phenomenon (''parameterization''). Computer model calibration involves comparing, in a statistically rigorous fashion, computer model output run at various parameter settings to observational data, taking into account systematic discrepancies between the computer model and the observations along with measurement error in the observations (cf. Bayarri et al. 2007). Learning about the parameters may be of interest in its own right—for instance when the parameter is climate sensitivity (cf. Forest et al. 2002). Alternatively, inferred distributions of the parameters may be used to characterize climate model projections and to propagate parametric and other uncertainties into the projection uncertainty.

In contrast, our aim has been to explore the capability of a proxy-based reconstruction to select among disparate modeling configurations and forcings, and we have treated the climate model experiments as ensembles of opportunity. We are therefore limited to exploring whether the climate reconstructions support a particular model configuration and forcing scenario over another and have not studied what the most probable values for a parameter might be, given the reconstructions and simulations. Uncertainties about model configurations and forcing scenarios, neither of which may be easily translatable into parameters, can result in larger projection uncertainties than uncertainties due to model parameters. It is possible that, in future work, our methods may be useful in conjunction with climate model calibration approaches. This would be one more step in the direction of capturing the various complex uncertainties that go into studies of climate change (e.g., Katz et al. 2013).

6. Conclusions

We have developed a hierarchical Bayesian model for comparing simulated and reconstructed spatially averaged climate time series. A key challenge in this exercise is that any shared variability between the simulated and reconstructed climates is due to a common response to external forcing, whereas we observe in each case the sum of the forced responses and independent realizations of internal variability.

In diagnosing the agreement between the reconstructed and simulated components, we make use of the slope β_1 and error variance τ^2 in the regression linking their respective forced components [Eq. (4)]. The slope is an indication of the strength of the relationship between the two forced components (the larger the slope, the stronger the relationship), whereas the error variance is a measure of the uncertainty about the best-fit linear relationship. In practice, we find that the error variance and slope increase in tandem so that an estimate of the fraction of variability in the simulated forced response that is attributable to the reconstructed forced response υ_f [Eq. (5)] is a useful metric for assessing simulation–reconstruction agreement.

Results for both the GISS-E2 and CSIRO simulations point to the possibility, but also the challenge, of using paleoclimate reconstructions to discriminate between climate simulations run under different estimates of prehistorical forcings. As the link is between quantities that are not directly observed, inferences are uncertain, and we are able to distinguish between only the broadest differences in the forcing configurations used to generate the two ensembles of simulations. For the GISS-E2 ensemble, we correctly identify simulations using the CEA volcanic forcing as being in better agreement with the reconstruction than simulations that exclude volcanic forcing or that use a volcanic forcing that is unreasonably strong. Differences between simulations that use the CEA volcanic forcing are more difficult to detect, as the posterior distributions of the diagnostic parameters (β_1, τ^2, and υ_f) feature substantial overlap. For the CSIRO ensemble, there is a general increase in agreement as additional forcings are included, but overall, as measured by the variance ratio, the agreement between the reconstructions and simulations is lower than for the GISS-E2 simulations. Results vary among the three CSIRO simulations conducted at each forcing configuration, with even the ordering of simulations, as measured by the variance fraction, changing within the three-member ensemble. These results point to the importance of including multiple simulations at a given forcing configuration in such experiments.

One of the main uses of climate models is to project future climate under estimates of future forcings (Collins et al. 2013). The paleoclimate record can serve as an important test bed for assessing model sensitivity and parameterizations, as the models are primarily tuned to agree with instrumental, rather than proxy, observations (Flato et al. 2013). Our results suggest that reconstructions are sufficiently informative to select between the broad features of simulations from a given climate model conducted under different climate forcing configurations. A more ambitious goal, and a subject of future research, is to use proxy–model comparisons to select between climate model configurations with different sensitivities or to weight an ensemble of future projections, each based on a climate model with a different sensitivity, based on the level of agreement with proxy-based reconstructions.

Acknowledgments. This work benefited from discussions with Chris Forest, Steven MacEachern, and Steven

Phipps. We are grateful to Gavin Schmidt and Steven Phipps for facilitating access to the GISS-E2 and CSIRO climate model simulations, respectively. Peter F. Craigmile is supported in part by the National Science Foundation through NSF-DMS-1407604 and NSF-SES-1424481. Murali Haran is supported by NSF-DMS-1418090 and the Network for Sustainable Climate Risk Management under NSF Cooperative Agreement NSF-GEO-1240507. Bo Li is supported by NSF-PLR-14-1839. Elizabeth Mannshardt was funded as a Postdoctoral Research Scholar through NSF Collaborative Research Grant NSF-DMS-1107046, which also supplied Murali Haran with travel support. Bala Rajaratnam is supported by NSF-DMS-1106642, NSF-DMS-CAREER-1352656, NSF-DMS-CMG-1025465, NSF-AGS-1003823, AFOSR FA 9550-13-1-0043, and the UPS fund.

REFERENCES

Ahmed, M., and Coauthors, 2013: Continental-scale temperature variability during the past two millennia. *Nat. Geosci.*, **6**, 339–346, doi:10.1038/ngeo1797.

Allen, M., and P. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, part I: Theory. *Climate Dyn.*, **21**, 477–491, doi:10.1007/s00382-003-0313-9.

Barboza, L., B. Li, M. P. Tingley, and F. G. Viens, 2014: Reconstructing past temperatures from natural proxies and estimated climate forcings using short-and long-memory models. *Ann. Appl. Stat.*, **8**, 1966–2001, doi:10.1214/14-AOAS785.

Bayarri, M., J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu, 2007: A framework for validation of computer models. *Technometrics*, **49**, 138–154, doi:10.1198/004017007000000092.

Bhat, K. S., M. Haran, R. Olson, and K. Keller, 2012: Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics*, **23**, 345–362, doi:10.1002/env.2149.

Braconnot, P., S. P. Harrison, B. Otto-Bliesner, A. Abe-Ouchi, J. Jungclaus, and J.-Y. Peterschmitt, 2011: The Paleoclimate Modeling Intercomparison Project contribution to CMIP5. *CLIVAR Exchanges*, No. 56, International CLIVAR Project Office, Southampton, United Kingdom, 15–19.

Brown, P. J., 1993: *Measurement, Regression, and Calibration*. Clarendon Press, 201 pp.

Chang, W., M. Haran, R. Olson, and K. Keller, 2014: Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.*, **8**, 649–673, doi:10.1214/14-AOAS733.

Christiansen, B., 2014: Straight line fitting and predictions: On a marginal likelihood approach to linear regression and errors-in-variables models. *J. Climate*, **27**, 2014–2031, doi:10.1175/JCLI-D-13-00299.1.

Collins, M., and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 1029–1136. [Available online at https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter12_FINAL.pdf.]

Crowley, T. J., G. Zielinski, B. Vinther, R. Udisti, K. Kreutz, J. Cole-Dai, and E. Castellano, 2008: Volcanism and the little ice age. *PAGES News*, No. 16, 22–23.

Dean, A. M., and D. Voss, 1999: *Design and Analysis of Experiments*. Springer, 742 pp.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866. [Available online at https://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter09_FINAL.pdf.]

Forest, C., P. Stone, A. Sokolov, M. Allen, and M. Webster, 2002: Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, **295**, 113–117, doi:10.1126/science.1064419.

Gao, C., A. Robock, and C. Ammann, 2008: Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models. *J. Geophys. Res.*, **113**, D23111, doi:10.1029/2008JD010239.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, 2003: *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC, 696 pp.

Goosse, H., H. Renssen, A. Timmermann, R. S. Bradley, and M. E. Mann, 2006: Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium. *Climate Dyn.*, **27**, 165–184, doi:10.1007/s00382-006-0128-6.

——, E. Crespin, A. de Montety, M. Mann, H. Renssen, and A. Timmermann, 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.

Guiot, J., J.-J. Boreux, P. Braconnot, and F. Torre, 1999: Data-model comparison using fuzzy logic in paleoclimatology. *Climate Dyn.*, **15**, 569–581, doi:10.1007/s003820050301.

Heaton, M. J., W. Kleiber, S. R. Sain, and M. Wiltberger, 2015: Emulating and calibrating the multiple-fidelity Lyon–Fedder–Mobarry magnetosphere–ionosphere coupled computer model. *J. Roy. Stat. Soc.*, **64C**, 93–113, doi:10.1111/rssc.12064.

Hegerl, G. C., P. Stott, M. Allen, J. Mitchell, S. Tett, and U. Cubasch, 2000: Optimal detection and attribution of climate change: Sensitivity of results to climate model differences. *Climate Dyn.*, **16**, 737–754, doi:10.1007/s003820000071.

——, T. J. Crowley, M. Allen, W. T. Hyde, H. N. Pollack, J. Smerdon, and E. Zorita, 2007: Detection of human influence on a new, validated 1500-year temperature reconstruction. *J. Climate*, **20**, 650–666, doi:10.1175/JCLI4011.1.

——, J. Luterbacher, F. González-Rouco, S. F. Tett, T. Crowley, and E. Xoplaki, 2011: Influence of human and natural forcing on European seasonal temperatures. *Nat. Geosci.*, **4**, 99–103, doi:10.1038/ngeo1057.

Hind, A., and A. Moberg, 2013: Past millennial solar forcing magnitude. *Climate Dyn.*, **41**, 2527–2537, doi:10.1007/s00382-012-1526-6.

——, ——, and R. Sundberg, 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. *Climate Past*, **8**, 1355–1365, doi:10.5194/cp-8-1355-2012.

Jansen, E., and Coauthors, 2007: Palaeoclimate. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 433–497.

Kaplan, J. O., K. M. Krumhardt, E. C. Ellis, W. F. Ruddiman, C. Lemmen, and K. K. Goldewijk, 2011: Holocene carbon emissions as a result of anthropogenic land cover change. *Holocene*, **21**, 775–791, doi:10.1177/0959683610386983.

Katz, R. W., P. F. Craigmile, P. Guttorp, M. Haran, B. Sansó, and M. L. Stein, 2013: Uncertainty analysis in climate change

assessments. *Nat. Climate Change*, **3**, 769–771, doi:10.1038/nclimate1980.

Kaufman, D., D. Schneider, N. McKay, and C. Ammann, 2009: Recent warming reverses long-term Arctic cooling. *Science*, **325**, 1236–1239, doi:10.1126/science.1173983.

Li, B., D. Nychka, and C. Ammann, 2007: The 'hockey stick' and the 1990s: A statistical perspective on reconstructing hemispheric temperatures. *Tellus*, **59A**, 591–598, doi:10.1111/j.1600-0870.2007.00270.x.

——, ——, and ——, 2010: The value of multiproxy reconstruction of past climate. *J. Amer. Stat. Assoc.*, **105**, 883–911, doi:10.1198/jasa.2010.ap09379.

Mann, M., and Coauthors, 2009: Global signatures and dynamical origins of the little ice age and medieval climate anomaly. *Science*, **326**, 1256–1260, doi:10.1126/science.1177303.

Masson-Delmotte, V., and Coauthors, 2013: Information from paleoclimate archives. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 383–464. [Available online at http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter05_FINAL.pdf.]

Moberg, A., 2013: Comparisons of simulated and observed Northern Hemisphere temperature variations during the past millennium—Selected lessons learned and problems encountered. *Tellus*, **65B**, 19921, doi:10.3402/tellusb.v65i0.19921.

——, R. Sundberg, H. Grudd, and A. Hind, 2015: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium—Part 3: Practical considerations, relaxed assumptions, and using tree-ring data to address the amplitude of solar forcing. *Climate Past*, **11**, 425–448, doi:10.5194/cp-11-425-2015.

Percival, D., and A. Walden, 2000: *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 594 pp.

Phipps, S., and Coauthors, 2013: Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years. *J. Climate*, **26**, 6915–6936, doi:10.1175/JCLI-D-12-00108.1.

Pongratz, J., T. Raddatz, C. Reick, M. Esch, and M. Claussen, 2009: Radiative forcing from anthropogenic land cover change since A.D. 800. *Geophys. Res. Lett.*, **36**, L02709, doi:10.1029/2008GL036394.

Sansó, B., C. E. Forest, and D. Zantedeschi, 2008: Inferring climate system properties using a computer model. *Bayesian Anal.*, **3**, 1–37, doi:10.1214/08-BA301.

Schmidt, G., and Coauthors, 2011: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). *Geosci. Model Dev.*, **4**, doi:10.5194/gmd-4-33-2011.

——, and Coauthors, 2012: Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.1). *Geosci. Model Dev.*, **5**, 185–191, doi:10.5194/gmd-5-185-2012.

——, and Coauthors, 2014: Using palaeo-climate comparisons to constrain future projections in CMIP5. *Climate Past*, **10**, 221–250, doi:10.5194/cp-10-221-2014.

Simkin, T., and L. Siebert, 1994: *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism during the Last 10,000 Years*. 2nd ed. Geoscience Press, 349 pp.

Steiger, N. J., G. J. Hakim, E. J. Steig, D. S. Battisti, and G. H. Roe, 2014: Assimilation of time-averaged pseudoproxies for climate reconstruction. *J. Climate*, **27**, 426–441, doi:10.1175/JCLI-D-12-00693.1.

Steinhilber, F., J. Beer, and C. Fröhlich, 2009: Total solar irradiance during the Holocene. *Geophys. Res. Lett.*, **36**, L19704, doi:10.1029/2009GL040142.

Stine, A. R., and P. Huybers, 2014: Arctic tree rings as recorders of variations in light availability. *Nat. Commun.*, **5**, 3836, doi:10.1038/ncomms4836.

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

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tingley, M. P., and P. Huybers, 2010a: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems. *J. Climate*, **23**, 2759–2781, doi:10.1175/2009JCLI3015.1.

——, and ——, 2010b: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation–maximization algorithm. *J. Climate*, **23**, 2782–2800, doi:10.1175/2009JCLI3016.1.

——, and ——, 2013: Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature*, **496**, 201–205, doi:10.1038/nature11969.

——, P. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam, 2012: Piecing together the past: Statistical insights into paleoclimatic reconstructions. *Quat. Sci. Rev.*, **35**, 1–22, doi:10.1016/j.quascirev.2012.01.012.

——, A. R. Stine, and P. Huybers, 2014: Temperature reconstructions from tree-ring densities overestimate volcanic cooling. *Geophys. Res. Lett.*, **41**, 7838–7845, doi:10.1002/2014GL061268.

Vieira, L., S. Solanki, N. Krivova, and I. Usoskin, 2011: Evolution of the solar irradiance during the Holocene. *Astron. Astrophys.*, **531**, A6, doi:10.1051/0004-6361/201015843.

Werner, J. P., J. Luterbacher, and J. E. Smerdon, 2013: A pseudoproxy evaluation of Bayesian hierarchical modeling and canonical correlation analysis for climate field reconstructions over Europe. *J. Climate*, **26**, 851–867, doi:10.1175/JCLI-D-12-00016.1.