This work has been accepted to Journal of Climate. The AMS does not guarantee that the copy provided here is an accurate copy of the final published work.

A data assimilation approach to last millennium temperature field reconstruction using a limited high-sensitivity proxy network

Jonathan M. King*

Department of Geosciences and Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona

Kevin J. Anchukaitis

School of Geography, Development, and Environment and Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona

Jessica E. Tierney

Department of Geosciences, University of Arizona, Tucson, Arizona

Gregory J. Hakim

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

Julien Emile-Geay and Feng Zhu

Department of Earth Sciences, University of Southern California, Los Angeles, California

Rob Wilson

School of Earth and Environmental Sciences, University of St Andrews, St Andrews, UK

**Corresponding author*: Jonathan King, jonking93@email.arizona.edu

ABSTRACT

We use the Northern Hemisphere Tree-Ring Network Development (NTREND) tree-ring database to examine the effects of using a small, highly-sensitive proxy network for paleotemperature data assimilation over the last millennium. We first evaluate our methods using pseudo-proxy experiments. These indicate that spatial assimilations using this network are skillful in the extratropical Northern Hemisphere and improve on previous NTREND reconstructions based on Point-by-Point regression. We also find our method is sensitive to climate model biases when the number of sites becomes small. Based on these experiments, we then assimilate the real NTREND network. To quantify model prior uncertainty, we produce 10 separate reconstructions, each assimilating a different climate model. These reconstructions are most dissimilar prior to 1100 CE, when the network becomes sparse, but show greater consistency as the network grows. Temporal variability is also underestimated before 1100 CE. Our assimilation method produces spatial uncertainty estimates and these identify treeline North America and eastern Siberia as regions that would most benefit from development of new millennial-length temperature-sensitive tree-ring records. We compare our multi-model mean reconstruction to five existing paleo-temperature products to examine the range of reconstructed responses to radiative forcing. We find substantial differences in the spatial patterns and magnitudes of reconstructed responses to volcanic eruptions and in the transition between the Medieval epoch and Little Ice Age. These extant uncertainties call for the development of a paleoclimate reconstruction intercomparison framework for systematically examining the consequences of proxy network composition and reconstruction methodology and for continued expansion of tree-ring proxy networks.

1. Introduction

Past variations in surface temperatures can be used to investigate a number of key characteristics of the Earth's climate system, including the response to radiative forcing, the regional effects of such forcings, and the role of internal modes of coupled ocean-atmosphere variability (Hegerl et al. 1997; Stott and Tett 1998; Delworth and Mann 2000; Meehl et al. 2004; Lean and Rind 2008; Stott and Jones 2009; Stott et al. 2010; Solomon et al. 2011; Phipps et al. 2013; Hegerl and Stott 2014; Kaufman 2014; Guillet et al. 2017; Neukom et al. 2019; Zhu et al. 2020). Paleoclimate temperature reconstructions using natural archives like tree-rings are particularly useful because they extend the short instrumental record to centennial and longer timescales. These provide an opportunity to characterize the patterns and magnitude of forced climate response and internal variability (Hegerl et al. 2003, 2007; Schurer et al. 2013; Masson-Delmotte et al. 2013). Climate field reconstructions (CFRs) can additionally capture the spatial fingerprints of large-scale temperature anomalies caused by radiative forcing and ocean-atmosphere dynamics (Mann et al. 1998; Evans et al. 2001; Seager et al. 2007; Cook et al. 2010a,b; Phipps et al. 2013; Anchukaitis and McKay 2015; Goosse 2017). CFRs have been developed using a number of methods (Tingley et al. 2012; Smerdon and Pollack 2016) including point-by-point methods (Cook et al. 1999, 2010a,b; Anchukaitis et al. 2017), variants of regularized expectation maximization (RegEM; Schneider 2001; Rutherford et al. 2003; Mann et al. 2009; Smerdon et al. 2011; Guillot et al. 2015), and reduced space approaches (Fritts 1991; Cook et al. 1994; Mann et al. 1998; Evans et al. 2002; Gill et al. 2016).

Recently, data assimilation (DA) has emerged as a promising CFR technique (e.g. Widmann et al. 2010; Bhend et al. 2012; Goosse et al. 2012; Steiger et al. 2014; Hakim et al. 2016; Matsikaris et al. 2015; Okazaki and Yoshimura 2017; Steiger et al. 2018; Franke et al. 2020). Assimilation methods integrate the climate signals recorded in paleoclimate proxies with dynamical constraints provided by climate models to produce spatially continuous climate field reconstructions and associated uncertainty estimates. There are several existing paleoclimate DA paradigms, including pattern nudging / forcing singular vectors (Van der Schrier and Barkmeijer 2005), particle filters (Goosse et al. 2012; Dubinkina and Goosse 2013; Matsikaris et al. 2015), and ensemble Kalman filters (Bhend et al. 2012; Steiger et al. 2014; Hakim et al. 2016; Dee et al. 2016; Perkins and Hakim 2017; Steiger et al. 2018; Tardif et al. 2019; Franke et al. 2020). Here, we focus on the ensemble Kalman filter (EnKF) approach (Steiger et al. 2014; Hakim et al. 2016), which has been shown to perform well compared to other DA methods in a paleoclimate context (Liu et al. 2017). EnKF methods update an ensemble of climate states to more closely match paleoclimate proxy records. These climate states are produced using one of two approaches: the "online" method, in which the ensemble is generated by a set of transient model simulations that propagate updates forward through time (e.g. Perkins and Hakim 2017); and the "offline" (or "no-cycling") method (Oke et al. 2002; Evensen 2003), in which ensembles are constructed from pre-existing climate model output (e.g. Bhend et al. 2012; Annan and Hargreaves 2012; Steiger et al. 2014; Hakim et al. 2016; Valler et al. 2019; Tardif et al. 2019; Franke et al. 2020). We focus here on the offline approach, which has been shown to perform favorably to online methods in paleoclimate contexts with reduced computational costs (Matsikaris et al. 2015; Acevedo et al. 2017). A key requirement of EnKF methods is the ability to estimate equivalent proxy values from climate model output. This is achieved through the use of forward models that translate climate state variables, like surface temperature, into proxy values, like tree-ring width (TRW) or maximum latewood density (MXD). These forward models can range in complexity from a simple linear relationship to more detailed Proxy Systems Models (PSMs) incorporating the physical processes that transform climate signals to proxy records (Evans et al. 2013). The use of forward models helps separate data and process level models in the data assimilation framework (Goosse 2016).

An important decision in any assimilation is the selection of the proxy network. Ultimately, this choice must balance spatiotemporal coverage with sensitivity to the reconstructed field and associated proxy uncertainties (Esper et al. 2005; Frank et al. 2010; Wang et al. 2015; Wilson et al. 2016; Anchukaitis et al. 2017; Esper et al. 2018; Franke et al. 2020; Cort et al. 2021). In general, large networks maximize coverage, but their size often results from the inclusion of proxy records with comparatively weak, complex, seasonally varying, or multivariate sensitivity to reconstructed variables. By contrast, smaller curated networks consisting of well-understood and strongly-sensitive proxies provide a higher ratio of signal to noise at the cost of reduced coverage (Frank et al. 2010). An additional consideration concerns the implementation of forward models: highly sensitive networks with a known climate response and seasonal window facilitate physically realistic forward models, potentially improving assimilation skill. Given the complexity of these trade-offs, network selection is not necessarily intuitive. Noisy proxies that covary poorly with climate fields are down-weighted by the Kalman filter algorithm; if this down-weighting renders the effects of climate-insensitive proxies negligible on a reconstruction, then a large network incorporating many proxies might appear preferable. However, work by Franke et al. (2020) indicates that EnKF temperature reconstructions using large proxy networks do not correlate with target temperatures as well as reconstructions produced using smaller, more sensitive networks. This result is supported by Tardif et al. (2019), who found that additional screening of proxy records for temperature sensitivity in an assimilation framework improved their ability to reconstruct salient pre-industrial climate features, such as cooling during the Little Ice Age. The importance of proxy sensitivity is further highlighted by Steiger and Smerdon (2017) who note that skillful hydroclimate DA requires proxies sensitive to the target reconstruction field.

Curated temperature sensitive proxy networks for data assimilation include the PAGES2k (PAGES2k Consortium 2013, 2017) and NTREND networks (Wilson et al. 2016; Anchukaitis et al. 2017). The PAGES2k network has been commonly used in paleo-DA applications (Hakim et al. 2016; Dee et al. 2016; Okazaki and Yoshimura 2017; Perkins and Hakim 2017; Tardif et al. 2019; Neukom et al. 2019) and consists of proxy records identified as temperature-sensitive and meeting minimum temporal coverage and age model precision criteria during the Common Era (PAGES2k Consortium 2017). DA reconstructions using this network may implement additional proxy screening but usually incorporate several hundred proxy records. The NTREND network has stricter requirements for inclusion: it consists of 54 published tree-ring chronologies selected by dendroclimatologists for demonstrating an established and reasonable biophysical association with local seasonal temperatures (Wilson et al. 2016). Franke et al. (2020) proposed that the additional coverage of the PAGES2k network is preferable to the increased sensitivity of the smaller NTREND network for global and hemisphere-scale temperature reconstructions but found the NTREND network provided the best reconstruction in the extratropical Northern Hemisphere. To produce a maximally skillful reconstruction for this region, we focus on assimilating the NTREND network but acknowledge that this choice is accompanied by a reduced spatial extent.

Before performing an assimilation, we seek to understand the advantages and tradeoffs of offline EnKF related to both the proxy data and climate model priors. We implement these sensitivity tests using pseudo-proxy experiments (Mann and Rutherford 2002; Zorita et al. 2003; Smerdon 2012), which allow us to test the DA method's ability to reconstruct known climate fields within a controlled setting. Here, we note the importance of model selection in DA pseudo-proxy experiments and distinguish between "perfect-model" and "biased-model" experimental designs. In a perfect-model experiment, the same model is used to generate the target field and as the model prior. Such designs are common in DA analyses (Annan and Hargreaves 2012; Steiger et al. 2014;

Okazaki and Yoshimura 2017; Acevedo et al. 2017; Zhu et al. 2020), where they are powerful tools for testing sensitivity to variables like proxy noise, network distribution, and calibration intervals. Biased-model paradigms use different climate models to generate target fields and assimilated model priors and can help examine the effects of biases in a model prior's mean state and spatial covariance. Dee et al. (2016) found model biases a potentially major source of error in paleo-EnKF reconstructions, so we employ both perfect and biased-model experiments in our investigations.

In this study, we begin by first evaluating the sensitivity of our DA method to proxy noise, network attrition, and climate model biases in a suite of pseudo-proxy experiments. We also use the pseudo-proxy framework to compare the skill of our DA method to point-by-point regression (PPR), the technique used for the original NTREND temperature field reconstruction (Anchukaitis et al. 2017). We then assimilate the real NTREND tree-ring network to reconstruct mean May through August (MMJA) temperature anomalies. We produce an ensemble of real reconstructions by assimilating NTREND with output from multiple climate models in the Coupled Modeling Intercomparison Project Phase 5 (CMIP5; Taylor et al. 2012) and the Community Earth System Model (CESM) Last Millennium Ensemble (LME; Otto-Bliesner et al. 2016). We quantify the skill of the DA reconstructions using spatial temperature anomaly fields, mean Northern Hemisphere extratropical (30°N–90°N) May through August time series, and withheld proxy data. Finally, we examine the climate response of the ensemble-mean reconstruction to radiative forcings and compare these responses against existing temperature field reconstructions.

2. Methods

a. Proxy Network

The NTREND network is a curated set of 54 published annual resolution tree-ring based summer-temperature proxy records selected by dendroclimatologists to maximize sensitivity to boreal summer temperatures while minimizing the response to other climate variables (Figure 1; Wilson et al. 2016; Anchukaitis et al. 2017). Although tree growth at the NTREND sites is primarily limited by summer growing temperatures, the optimal summer season varies between sites. Wilson et al. (2016) determined the season of highest temperature sensitivity for each site and identified mean MJJA temperatures anomalies as the optimal reconstruction target for the network as a whole. The network only includes sites between 40°N and 75°N as lower latitude trees tend to exhibit sensitivity to multiple climate influences, especially moisture limitations. Each record is derived from ring-width measurements (TRW), maximum latewood density (MXD; Schweingruber et al. 1978), or a mixture of TRW, MXD, and blue intensity (BI; McCarroll et al. 2002; Björklund et al. 2014; Rydval et al. 2014; Wilson et al. 2019). The network extends from 750 - 2011 CE, with maximum coverage over the period from 1710-1988 CE. Spatial coverage is greater over Eurasia (39 sites) than North America (15 sites), with a distinct spatial imbalance prior to 1000 CE (20 vs. 3). We end all reconstructions in 1988 CE as network attrition limits the utility of assimilated NTREND reconstructions after this point (Anchukaitis et al. 2017).

b. Data Assimilation

Our data assimilation method uses an ensemble Kalman filter (EnKF) (Evensen 1994; Steiger et al. 2014)

$$\mathbf{X_a} = \mathbf{X_p} + \mathbf{K}(\mathbf{Y} - \mathbf{Y_e}) \tag{1}$$

to update an initial ensemble of climate states ($\mathbf{X_p}$) given proxy data ($\mathbf{Y}$) and model estimates of the proxy data ($\mathbf{Y_e}$). These data are combined via the Kalman Gain ($\mathbf{K}$; detailed in Appendix A1) to produce an updated ensemble ($\mathbf{X_a}$) in each reconstructed annual time step. We use an EnKF variant known as the ensemble square root Kalman filter (EnSRF; Andrews 1968), with an "offline" (or "no-cycling") approach (Oke et al. 2002; Evensen 2003). The complete details of our approach are given in Appendix A1 and described in Steiger et al. (2014) and Hakim et al. (2016). The Kalman Filter can be expressed as a recursive Bayesian filter (Chen et al. 2003; Wikle and Berliner 2007), wherein new information ($\mathbf{Y}$) updates estimates of state parameters ($\mathbf{X}$). Hence, we will often refer to $\mathbf{X_p}$ as the model prior, and the updated ensemble $\mathbf{X_a}$ as the model posterior.

We implement a covariance localization scheme, which limits the influence of proxies outside of a specified radius. Localization was originally developed to limit spurious covariance arising from sampling noise in small ensembles of $m \leq 50$ (Houtekamer and Mitchell 2001). Our offline approach enables the use of much larger ensembles ($m > 1000$), but we note that spurious covariances may still arise from biases in a climate model's covariance structure. Consequently, localization may improve the quality of assimilated paleoclimate reconstructions even for large prior ensembles. The localization radius is an important free parameter in this method and must be assessed independently for different model priors, reconstruction targets, and proxy networks (Tables 2, S1). The process used to select localization radii for these experiments is detailed in Appendix A2.

To generate model estimates of the proxy values, we follow the methodology of Tardif et al. (2019) and use linear univariate forward models trained on the mean temperature of each site's optimal growing season (Wilson et al. 2016), such that:

$$\mathbf{y}_{\mathbf{e}_j} = \alpha_j + \beta_j \mathbf{T}_j. \tag{2}$$

Here, $\mathbf{T}_j$ is a vector of mean growing-season temperature anomalies extracted from the prior. The coefficients α_j and β_j are determined by regressing assimilated observations ($\hat{\mathbf{y}}_j$) against mean growing-season temperature anomalies from the closest grid cell of the target field. We emphasize that these target fields vary by application. For pseudo-proxy experiments, the target field is a specific model realization, whereas the real assimilation uses CRU-TS 4.01 (Harris et al. 2014). Regardless of the target, we perform each regression over the years in which the real NTREND records overlap data from the closest land grid cell in CRU-TS 4.01; this ensures that both pseudo-proxy and real reconstructions use regressions with the same temporal span. The variance of each record's regression residuals is used as the observation uncertainty (R_{jj}) in the Kalman Filter (Appendix A1). This uncertainty ranges from 0.23 to 1.34 proxy units over the network.

We construct prior ensembles using output from the past1000 and historical experiments of the Coupled Modeling Intercomparison Project Phase 5 (CMIP5; Taylor et al. 2012) as well as the Last Millennium Ensemble (LME; Otto-Bliesner et al. 2016). For a given assimilation, we use values from a single climate model and designate each year of available output as a unique ensemble member. We use static model priors, whereby the same prior is used for each reconstructed time step. This scheme is justified by the limited forecast skill of climate models beyond the annual reconstruction timescale (Bhend et al. 2012) and is common in paleo-DA applications (e.g. Steiger et al. 2014; Dee et al. 2016; Tardif et al. 2019). A summary of the model ensembles is given in Table 1. The past1000 CMIP5 data for each model are from the ensemble member designated *r1i1p1*, and LME output was selected from full-forcing run 2. We assimilate temperature anomalies relative to the 1951-1980 CE mean; this helps avoid the effects of climate model mean state biases, but we note that model covariance biases are unaffected. In all reconstructions, we update the mean May through August (MJJA) temperature anomaly field, rather than individual months. We assess the skill of each assimilation by comparing the Pearson's correlation coefficients, root mean square errors (RMSEs), mean biases, and standard deviation ratios.

c. Pseudo-proxy Reconstructions

Before assimilating the real NTREND network, we first examine the skill of our DA method in a pseudo-proxy framework (Smerdon 2012). This approach allows us to test the method's ability to reconstruct known climate field targets within a controlled setting. Here, we specify the target fields as surface temperatures from the years 850-2005 CE from either the Last Millennium Ensemble full-forcing run 2 (CESM; Otto-Bliesner et al. 2016), or from the combined last millennium and historical runs of the Max Planck Institute for Meteorology Earth System Model (MPI; Marsland et al. 2003; Stevens et al. 2013). While this experimental design is intentionally tractable, we caution that the observed spatial patterns of skill will depend on the specific models used (Smerdon et al. 2011). Here, we are interested in examining the sensitivity of EnSRF to the proxy network and climate model prior, so we systematically explore the effects of noisy proxy records, network attrition, and biased climate models on DA performance. To examine the effects of model covariance biases, we test each combination of target field and model prior for LME and MPI, which allows us to alternate between perfect-model and biased-model experimental designs.

After selecting a target field, we generate pseudo-proxies using:

$$\hat{\mathbf{y}}_j = a_j + b_j \mathbf{T}_j^{\text{target}} + \epsilon_j \tag{3}$$

where $\hat{\mathbf{y}}_j$ is the j^{th} pseudo-proxy record and $\mathbf{T}_j^{\text{target}}$ is the vector of mean growing season temperature anomalies from the grid cell closest to the proxy site in the target climate field. The coefficients a_j

and b_j are the intercept and slope obtained by regressing the real NTREND network against mean growing-season temperature anomalies from the nearest land cells in CRU-TS 4.01; in this way, the pseudo-proxies mimic the temperature response of the real NTREND network for at least the instrumental period.

We examine the effects of proxy noise by selectively neglecting or adding Gaussian white noise to the pseudo-proxies, such that:

$$\epsilon_j \sim \begin{cases} 0, & \text{Perfect} \\ \mathcal{N}(0, \mathrm{R_{jj}}), & \text{Noisy} \end{cases} \tag{4}$$

Here, $\mathrm{R_{jj}}$ is the proxy-uncertainty weight for the j^{th} NTREND record and is the variance of the NTREND-CRU regression residuals. When testing noisy proxies, we perform 101 assimilations using different noise matrices and report the median skill metrics. Here, we use white noise because it allows us to directly tune the $\mathrm{R_{jj}}$ weight in the Kalman Filter. The median signal-to-noise ratio is 0.80 for the CESM pseudo-proxies and 0.85 for the MPI pseudo-proxies, which is consistent with values found in other pseudo-proxy experiments (Smerdon 2012). In each test, we examine the effects of network attrition by first assimilating the full set of pseudo-proxies over the entire period and then comparing this to an assimilation where the pseudo-proxies are subjected to the same temporal attrition as the real NTREND network.

After generating pseudo-proxies for a given experiment, we generate pseudo-proxy estimates by applying equation 2 to the prior ensemble. The coefficients α_j and β_j are determined by regressing the pseudo-proxies against the target field. Note that pseudo-proxy noise and sampling errors will affect the statistics obtained from these regressions, so α_j and β_j are estimates of the coefficients a_j and b_j used to generate the pseudo-proxies. This mimics how noise and sampling errors can introduce errors into forward models calibrated on real NTREND data. Once we obtain pseudo-proxy estimates, we then determine an optimal localization radius (Appendix A2, Table S1).

A key feature of pseudo-proxy experiments is that the target reconstruction is known. Consequently, we can assess skill directly against the correct answer. Here, we examine pseudo-proxy reconstruction skill using mean Northern Hemisphere extratropical (30°N–90°N) MJJA temperature time series, and spatial grid point time series over the full reconstruction period (850 CE to 1988 CE).

We compare the most realistic (biased-model, noisy-proxy, temporal-attrition) pseudo-proxy DA reconstructions to analogous reconstructions generated using point-by-point regression (PPR). PPR is a "region of interest" CFR technique that iteratively calculates a nested multivariate principal components regression model between predictor network and each point in the target field (Cook et al. 1999). The method was motivated by the premise that proxies near a reconstructed grid point are more likely to reflect climate at that site. Consequently, PPR uses a strict search radius to select proxy predictor series for each grid point reconstruction. The method was first used for drought reconstructions (Cook et al. 1999, 2010a,b) and later adapted for continental temperature anomalies (Cook et al. 2013). Anchukaitis et al. (2017) used the method to reconstruct hemispheric temperature anomalies, and we follow their implementation in this study.

In brief, given a target of gridded climate observation, the method first identifies proxy sites within 1000 km of each grid point centroid. If no proxy records are found within 1000 km, the search radius is expanded in 500 km increments to a maximum of 2000 km until proxy sites are found within the radius. All proxy sites found within the search radius are then used as predictor sites for that grid point. If no predictors are found within 2000 km, then no reconstruction is performed for the grid. These radii are based on decorrelation decay lengths in the observational temperature field from Cowtan and Way (2014). A multivariate regression model is then calibrated against the MJJA temperature values of the target field (Cowtan and Way 2014) for each grid point over the period 1945 to 1988 CE, and the reconstructions are validated using withheld temperature data for the period 1901 to 1944 CE. As the number of records declines back through time, the regression model is recalibrated and validated for each change in network size and scaled to match the mean and variance of the predictand during their overlapping time period (Meko 1997; Cook et al. 1999). For a given grid point, temperature anomalies are obtained for all years in which at least one predictor record remains within the initial search radius. Following Anchukaitis et al. (2017), we then screen the final reconstructed field in each time step to only include grid cells where the reduction of error (RE; Cook et al. (1994)) statistic is greater than zero. We use this screened field here as the final PPR MJJA temperature reconstruction.

d. Real NTREND Reconstruction

We next assimilate the real NTREND network. To examine the effects of prior selection, we produce 10 real DA reconstructions each using a different climate model to generate the prior (Table 1). Since each prior is itself an ensemble, these 10 reconstructions effectively create an ensemble of ensembles. To minimize ambiguity, we will henceforth refer to the set of 10 reconstructions as the "multi-model ensemble", and the DA ensemble for each individual reconstruction as a "prior/posterior ensemble".

Forward model estimates of the NTREND records in each reconstruction are determined by applying equation 2 to CRU-TS 4.01. We assess the skill of each reconstruction using time-series of mean Northern Hemisphere extratropical (30°N–90°N) MJJA temperature, instrumental spatial field grid points, and independent proxy records. The skill of the extratropical time series is determined using a Monte Carlo calibration-validation procedure (Appendix A2). Spatial skill is computed against the Berkeley Earth surface temperature field (BEST; Rohde et al. 2013) over the period 1901 - 1988 CE. The BEST instrumental record is not used in the forward model and localization calibrations, which instead leverage the CRU product. However, we caution that BEST is not a truly independent dataset, as both BEST and CRU are partly based on the same instrumental climate data. As an additional validation we assess the ability of DA to reconstruct withheld proxy time series. We perform a series of leave-one-out assimilations for each model by iteratively removing a single proxy time-series from the NTREND network and assimilating the remaining 53 records. In these experiments, we construct the prior from the average temperatures over the removed site's optimal growing season at the grid point closest to the removed site. This allows us to apply Equation 2 to the posterior to estimate the removed record from the reconstruction. We then compare this estimate to the real withheld NTREND record.

We next calculate a mean reconstruction for the multi-model ensemble. To do so, we first calculate ensemble-mean values from the posterior of each of the reconstructions. The mean of the multi-model ensemble is then calculated as the mean of these 10 posterior ensemble means. We quantify uncertainty of the multi-model mean using first the mean of the 10 posterior ensemble widths:

$$\sigma^2_{\text{multi-model mean}} = \frac{1}{10}\Sigma_{i=1}^{10}\sigma^2_{\text{posterior ensemble i}} \tag{5}$$

and then the 2σ width of the multi-model ensemble for the series. We first determine the multi-model ensemble-mean for the extratropical MJJA time series. We next compute a mean spatial reconstruction for the multi-model ensemble by linearly interpolating each reconstruction to the lowest model resolution and averaging at each grid point.

We compare the multi-model mean spatial product to several recent temperature CFRs summarized in Table 3. In brief, Guillet et al. (2017) focused on reconstructing high-frequency temperature anomalies associated with known volcanic eruptions using a network of a similar size and composition to the NTREND network in a linear regression framework and their work provides a comparison point with Anchukaitis et al. (2017). The LMR 2.1 reconstruction applied an offline EnSRF DA to the PAGES2k network and allows us to compare DA reconstructions using different proxy networks (Tardif et al. 2019). From Zhu et al. (2020), we examine the reconstruction of mean June through August (JJA) temperatures using PAGES2k trees. The Neukom et al. (2019) DA offers another comparison point, using a proxy network of intermediate size derived from a screened version of PAGES2k. Neukom et al. (2019) performed an ensemble of reconstructions using different methods and recommend using the ensemble mean reconstruction for climate analysis; however, we only focus on the DA product to emphasize the differences in reconstructions that arise when using similar methodologies.

We examine the temperature response to external forcing for both the reconstruction ensemble and temperature CFRs. We compare temperature anomalies between the Medieval Climate Anomaly (MCA; 950 - 1250 CE) and the Little Ice Age (LIA; 1450 - 1850 CE) (Masson-Delmotte et al. 2013; Anchukaitis et al. 2017), and separately use superposed epoch analysis (Haurwitz and Brier 1981) to determine composite mean responses to major tropical volcanic eruptions. For the volcanic events, we follow Sigl et al. (2015) and identify years containing a global eruption forcing magnitude equal to or larger than the 1884 Krakatoa eruption ($n = 20$), which yields the following event years: 916, 1108, 1171, 1191, 1230, 1258, 1276, 1286, 1345, 1453, 1458, 1595, 1601, 1641, 1695, 1809, 1815, 1832, 1836, and 1884 CE (Sigl et al. 2015; Anchukaitis et al. 2017). We calculate temperature anomalies relative to the mean of the five years preceding each of these event years.

3. Results

a. Pseudo-proxy experiments

The pseudo-proxy reconstructions are most skillful in the extratropical Northern Hemisphere (Figure 2). In this region, ocean basin correlations are lower relative to land with notable exceptions over the eastern and north-western edges of the Pacific. Correlations generally decline with increasing distance from the extratropical Northern Hemisphere and the tree-ring network, although significant spatial heterogeneity exists throughout the tropics. The climate model covariance biases cause the largest reductions in correlation coefficients and sharply reduce skill outside of the extratropical Northern Hemisphere. Network attrition and proxy noise have comparatively minor effects over the full period. Results for other skill metrics show similar behavior (Figures S1, S2, and S3).

We next compare the most realistic (biased-model, noisy-proxy, temporal-attrition) DA experiments to PPR reconstructions. Given the strict reconstruction radius in PPR, and the spatial pattern of DA skill, we consider only the extratropical Northern Hemisphere in our discussion. The skill metrics for the mean extratropical time series are similar for the two methods (Table S2; Figures S4, S5). The regional spatial correlations of the DA and PPR reconstructions for the CESM and MPI targets (Figures 3 and S6, respectively) are also comparable: each exhibits correlations with the target field greater than 0.7 in Scandinavia, western Siberia, and western Canada, and these regions correspond to the best coverage by the proxy network. Similarly, both methods exhibit low correlations in southeastern Canada, eastern Siberia, and in the region of the Black and Caspian Seas. The DA does however exhibit a broader spatial region of high correlation than PPR, and DA correlations are higher than PPR values at nearly all grid points. Similarly, DA reconstructions exhibit lower RMSE values at most grid points. Standard deviation ratios indicate that the DA reconstructions underestimate temporal temperature variability, but this effect is less severe near the proxy sites. In contrast with DA, PPR time series σ ratios neither strictly overestimate nor strictly underestimate temporal variability, instead demonstrating a mixed response over the hemisphere. In general, our DA reconstructions underestimate variability more strongly than the PPR analogues. Mean biases are comparable, with both methods exhibiting similar spatial patterns and bias magnitudes, although it is interesting to note that the spatial patterns of bias change markedly depending on the target field.

b. Real NTREND Reconstruction

For the real NTREND data assimilation, validation statistics for the mean extratropical MJJA time series are similar across all priors (Table 2) with mean correlations of 0.70, RMSE of 0.19 °C, and absolute mean bias of 0.06 °C. Temporal variability is close to the target with mean standard deviation ratios of 1.11. Time series obtained using different model priors (Figure S7) have a mean range of 0.22 °C over the period of full coverage (1750-2988 CE; n = 54). However, the reconstructed time series diverge as the network becomes sparse, with a range of 0.76 °C by the first year of the reconstruction (750 CE; n = 4). The model ensemble-mean time series exhibits similar skill values as the reconstructions for the individual models (Table 2) with a correlation of 0.72, RMSE of 0.18 °C, temporal σ ratio of 1.06, and a mean bias of 0.05 °C.

We compare the extratropical MJJA time series for the multi-model mean to analogous time series extracted from the Berkeley Earth (BEST) instrumental record and the Anchukaitis et al. (2017) NTREND PPR reconstruction (Figure 4). The DA series shows similar behavior to BEST from 1880-1988 CE, although both the DA and PPR reconstructions of Anchukaitis et al. (2017) diverge from this dataset over the earliest period from 1850-1879 CE. This may reflect a warm bias (Parker 1994; Frank et al. 2007; Böhm et al. 2010) and limited spatial coverage (Rohde et al.

2013; Anchukaitis et al. 2017) in the early instrumental temperature record. The DA and PPR time series show similar behavior over most of the record, with a correlation coefficient of 0.88. Temporal variability is generally higher in the PPR series than in the DA. Prior to about 1100 CE, the series' running standard deviations show larger differences, which is caused by the decrease in DA reconstructed variability.

Most spatial validation statistics show similar patterns to those observed in the pseudo-proxy experiments (Figure 5). Correlation coefficients and standard deviation ratios indicate the highest skill over Scandinavia, central and northern Asia, and northwestern North America, the regions of densest network coverage. Correlation coefficients approach 0.8 and standard deviation ratios approach 1 near the proxy sites themselves. Over land, mean biases are typically below 0.5 °C, with the largest largest over central Canada and eastern Siberia and smallest over the Arctic Archipelago, Alaska, and west-central Asia. Away from the proxy sites, temporal variability is underestimated, particularly over the oceans. However, most land grid points exhibit σ ratios near 1 with a slight overestimate in central Asia and northern Japan. Much of the temporal variability in the extratropical mean time series is driven by land grid points, and this tendency helps reconcile Figure 5 with extratropical mean time series σ ratios near 1. RMSE values are typically less than 0.6 °C, but rise to values near 1 °C over the North Pacific, central Canada, and north of the Caspian Sea.

Independent proxy validation statistics (Table 4) show median correlation coefficients near 0.5, and RMSE values near 1°C. Temporal variability is underestimated relative to the target series with σ ratios typically between 0.3 and 0.4. Mean biases are variable and depend on the prior model used. Not surprisingly given the sparsity of the NTREND network, removing even a single proxy record from the assimilation can substantially reduce the ability to reconstruct temperature anomalies at nearby grid cells. Consequently, the leave-one-out assimilation process we use to assess independent proxy skill almost certainly underestimates overall field validation skill. Nevertheless, these values are comparable to previous efforts with median correlation coefficients somewhat higher than those in Hakim et al. (2016) and Tardif et al. (2019).

c. Epochal Temperature Changes

We next examine the temperature change between the Medieval Climate Anomaly (MCA; 950 - 1250 CE) and the Little Ice Age (LIA; 1450 - 1850 CE) (Masson-Delmotte et al. 2013; Anchukaitis et al. 2017). The reconstructions nearly all indicate warmer temperatures during the MCA throughout the high latitudes with maximum anomalies typically over northeastern Canada (Figure 6). However, anomaly magnitudes vary across reconstructions with values ranging from over 1.6 °C (for CCSM4, MIROC, MPI priors) to less than 0.8 °C (IPSL and FGOALS priors). The spatial pattern also varies by model prior. Many reconstructions show stronger anomalies in Fennoscandia, northeastern Asia, and northwestern North America, but these patterns do not occur in all models.

Comparing the MCA-LIA difference for our multi-model mean reconstruction with other CFRs (Figure 7), we find our spatial anomaly patterns most similar to Anchukaitis et al. (2017). Anomaly magnitudes are also comparable, except over northeastern Canada. In the Anchukaitis et al. (2017) reconstruction, this region exhibits anomalously high medieval temperatures (> 3 °C), which they attribute to a detrending artifact in a tree-ring record from Quebec. By contrast, our DA reconstruction produces a maximum medieval anomaly of 1 °C for this region, in better agreement with other proxy reconstructions (e.g. 0-1.5°C; Sundqvist et al. 2014). Comparing the results of this study to Neukom et al. (2019), we observe that both NTREND DA and Neukom et al. (2019) exhibit a positive anomaly over most of the high-latitude Northern Hemisphere; however, the anomalies in the Neukom et al. (2019) product have much larger magnitudes and the maxima of the North America features occur in different locations. Zhu et al. (2020) also indicate positive anomalies in the Northern Hermisphere, but these are lower magnitude than the other products and more spatially localized. By contrast, the LMR2.1 product (Tardif et al. 2019) exhibits an anomaly pattern notably different from the other reconstructions, with a strong positive anomaly in the Arctic Ocean north of Siberia. Since the Guillet et al. (2017) reconstruction reflects high-pass filtered reconstructed temperatures, we do not consider it in this comparison.

d. Volcanic Response

We next examine the composite mean response to major tropical volcanic eruptions. Our 10 reconstructions show broadly similar responses to large tropical volcanic eruptions (Figure 8), with the spatial pattern characterized by a strong cold anomaly in northern Canada and a second region of cooling extending from Fennoscandia east of the Caspian Sea toward central Asia. However, the extent and magnitude of these vary between the different reconstructions. Several regions also exhibit markedly different spatial patterns across the 10 reconstructions. In particular, the response in central North America and eastern Asia appears highly sensitive to the choice of model prior.

Comparing the volcanic pattern for our multi-model mean reconstruction with the other existing CFRs (Figure 9) shows large differences in spatial patterns, magnitudes, and even sign of the anomalies. In general, most CFRs show some combination of cooling anomalies in northern North America and northern Asia, with a slight neutral or warming anomaly in the North Pacific. However, these features are not present in all the CFRs and vary in maximum magnitude. The mean of our model ensemble, Anchukaitis et al. (2017), and Guillet et al. (2017) products all exhibit the northern Canada and western Asia cooling features and the spatial extent is similar for the two NTREND products. In contrast, the Guillet et al. (2017) Canadian feature is centered farther east, and its northern Asian feature is stronger (near 1.5 °C) with a maximum more strongly localized to northern Siberia. These two features are also present in Zhu et al. (2020), but maximum cooling is smaller in magnitude. The LMR2.1 does not show distinct north Asian terrestrial cooling, although an anomaly of 0.6 C is reconstructed in the Arctic Ocean north of Siberia. This reconstruction also demonstrates a North American response pattern similar to Zhu et al. (2020) with a reduced magnitude of cooling in northern Canada. The Neukom et al. (2019) product again shows the largest anomalies, with values greater than 1.5 °C over much of northern Siberia and Fennoscandia. This feature does not extend as far south as in the NTREND DA ensemble-mean but is zonally wider. Neukom et al. (2019) also show a single strong North American feature with cooling magnitudes near 1.2 °C. Interestingly, Neukom et al. (2019) exhibits a North Pacific warming response that strengthens one year after the volcanic event, a feature also evident in the Anchukaitis et al. (2017) reconstruction that may reflect changes in atmospheric circulation following an eruption (e.g. Robock 2000; Stenchikov et al. 2006; Christiansen 2008; Schneider et al. 2009)

4. Discussion

The pseudo-proxy experiments indicate that regions of high reconstruction skill for the assimilated NTREND network is limited to the extratropical Northern Hemisphere when using biased climate model priors. This finding supports work by Franke et al. (2020) and suggests that analyses of temperatures using the NTREND network should be limited to this region, consistent with Wilson et al. (2016) and Anchukaitis et al. (2017). In comparison with Anchukaitis et al. (2017) (NTREND PPR), our DA method exhibits similar skill at reconstructing mean Northern Hemisphere extratropical MJJA time series using the NTREND network, but also provides continuous field estimates of past temperature and improves the spatial correlation and RMSE. We suggest this improvement arises at least in part from the contrast between PPR's strict-limited search radius and the DA's longer localization radii. Many NTREND sites exhibit statistically significant covariance with the MJJA temperature field outside of PPR's 2000 km maximum search radius (see Figure 5 of Anchukaitis et al. (2017)), and these distal covariances are not used to improve the PPR reconstruction. By contrast, the DA uses no localization in these pseudo-proxy experiments (Table S1) and if the model prior provides a good estimate of a proxy site's field covariance, the proxy record can inform the reconstruction of distal grid points. Ultimately, these results suggest that our DA method improves on the spatial component of Anchukaitis et al. (2017) for reconstructing a Northern Hemisphere temperature history of the Common Era from the NTREND network. We note that, as is the case for most field reconstruction methods (Ammann and Wahl 2007; Tingley et al. 2012), our offline DA method implicitly assumes the broad-scale covariance patterns can be considered stationary through time. Transient offline (e.g. Bhend et al. 2012; Valler et al. 2019; Franke et al. 2020) or online assimilation techniques (e.g. Perkins and Hakim 2017) may offer additional improvements.

Our results also highlight the sensitivity of the DA reconstructions to the model prior. In the pseudo-proxy experiments, the introduction of model covariance bias reduces widespread global skill to the high latitude Northern Hemisphere and the regions nearest the proxy sites. Network attrition and proxy noise cause comparatively small effects over the full period, a finding in agreement with Dee et al. (2016). Given this potential for perfect-model experiments to exaggerate the magnitude and spatial extent of DA skill, we encourage future DA proof-of-concept and sensitivity studies to consider perfect-model experiments in conjunction with biased-model cases. In contrast with these results, previous assimilation efforts have found little sensitivity to the choice of prior (Hakim et al. 2016). The small size of the NTREND network may exacerbate this sensitivity, but even assimilations using larger networks may be sensitive to the choice of priors in those periods with reduced proxy coverage.

Reconstructions are most sensitive to the prior when the proxy network becomes small. For example, despite using the same proxy network and reconstruction technique, mean extratropical MJJA temperature time series diverge by more than 0.5 °C in the earliest parts of the reconstruction when the number of sites in our network is limited (Figure S7). The use of different priors also produces noticeable differences in spatial MCA-LIA temperature anomaly patterns (Figure 6), which we interpret as arising from the reduced size of the proxy network during the MCA. In contrast, the volcanic response maps present a more consistent spatial pattern (Figure 8), which we attribute to the larger size of the proxy network during most of the volcanic events. The magnitude of the forced response may also contribute to similarity across the priors; however, the volcanic response maps still exhibit different spatial patterns in regions like east Asia where the proxy network is sparse.

The consistency with which the DA underestimates the temporal variability of the target field, particularly over the oceans and far from the proxy sites, requires consideration. In this study, we focus on time series derived from the posterior ensemble-mean at each time step. However, this focus on the ensemble-mean neglects the width of the full posterior ensemble. Like many offline EnSRF studies (e.g. Hakim et al. 2016; Dee et al. 2016; Steiger et al. 2018), our method uses a stationary prior in each time step; thus, the prior ensemble-mean is constant through time. As the proxy network becomes sparse, update magnitudes decrease, and the posterior ensemble more closely resembles the prior. When this occurs, the reconstructed ensemble-mean time series will closely resemble the mean of the prior ensemble, and the time series' temporal variability will approach zero. Similarly, regions far from the proxy network will exhibit smaller update magnitudes, so grid point time series far from the proxy sites have lower σ ratios. However, this reduction in temporal variability is balanced by increased posterior ensemble width, which will remain near the spread of the prior ensemble. Incorporating the width of the posterior with ensemble-mean time series can produce a range that encompasses target time-series variability, but it is not always clear how to use these ranges in spatiotemporal analyses. Hence, we emphasize that users of DA products with constant priors should carefully consider how changes in the proxy network affect the temporal variability of posterior ensemble-mean time series and make use of the posterior range when possible. We also note that allowing the model prior to vary in each time step may help mitigate these effects, which again may argue for expanded future use of transient offline priors (e.g. Bhend et al. 2012; Valler et al. 2019; Franke et al. 2020) or online assimilation techniques (e.g. Perkins and Hakim 2017) where possible.

The prior sensitivity and temporal variability effects underscore the importance of understanding how the proxy network affects the quality of the reconstruction (Esper et al. 2005; Wang et al. 2014). A key feature of DA techniques is the ability to estimate reconstruction uncertainty in each time step from the width of the posterior ensemble. Figure 10 provides an example of such an analysis for the multi-model mean by examining the temperature response following the 1257 CE (Lavigne et al. 2013) and 1600 CE (De Silva and Zielinski 1998) volcanic eruptions in conjunction with the full posterior width. The uncertainty maps for both events show maxima in central North American and northeastern Asia and suggest that associated temperature anomalies should be interpreted more cautiously. Notably, these regions correspond to areas that are also sensitive to the prior in Figure 8. By contrast, central and east-central Asia, Fennoscandia, central Europe, and southwestern Canada exhibit a narrow posterior for both events, so volcanic anomalies in these regions are better constrained. Interestingly, the temperature response in 1601 CE is relatively small over much of central Europe and reconstruction uncertainty is relatively low, which suggests this feature may be a robust feature of the post-eruption climate anomaly. In addition to supporting analysis of reconstructed climate features, these uncertainty estimates can help identify regions that would benefit from increased network density (Comboul et al. 2015). In particular, we observe that northern North America and eastern Siberia would benefit from the development of new millennial-length temperature-sensitive tree-ring records.

The CFR comparison reveals the highly variable nature of spatial patterns and magnitudes of reconstructed temperature anomalies that result from different selections of proxy networks, target fields, and reconstruction methodologies. For example, despite using the same proxy network and target field, the DA multi-model mean and PPR result from Anchukaitis et al. (2017) have MCA-LIA anomalies that differ by over 2 °C in northeastern Canada (Figure 7), which relates to the outsized effect of the Quebec tree-ring width record (Gennaretti et al. 2014) on the Anchukaitis et al. (2017) reconstruction. We note that the localization radii used in our reconstructions ($\geq$ 9500 km) allow proxies to influence grid cells farther away than the maximum 2000 km search radius used by Anchukaitis et al. (2017), so distant proxies are able to counter the effects of the Quebec record in the DA. Even within the same DA framework, our results indicate that reconstructed temperature responses are highly variable, particularly for MCA-LIA anomalies. These differences result from targeting different fields and leveraging different proxy networks. Aside from spatial and temporal coverage, we note that using proxy records that are not strictly temperature sensitive can introduce structural biases relative to other temperature CFRs. For example, the LMR2.1 reconstruction includes proxies that are sensitive to more than just temperature, which could possibly reduce update magnitudes and help explain the smaller magnitudes of the volcanic responses. Similarly, the Neukom et al. (2019) DA product and LMR2.1 incorporate proxies like corals and lake-sediments that are not present in the tree-ring based CFRs, and it is possible that these records influence the large magnitudes of the Neukom et al. (2019) DA climate responses or the atypical LMR2.1 MCA-LIA spatial pattern. However, we emphasize that these hypotheses are strictly speculative at this moment and that the differences in reconstructed climate response by themselves do not indicate whether one proxy network or reconstruction is superior to another in representing past climate variability. Instead, our CFR comparison highlights that, despite the recent decades of progress in understanding both methods and paleoclimate data (Hughes and Ammann 2009; Frank et al. 2010; Smerdon et al. 2011; Tingley et al. 2012; Wang et al. 2014; Smerdon and Pollack 2016; Christiansen and Ljungqvist 2017; Esper et al. 2018), differences in reconstructions of past temperature still arise when using different proxy networks, different target seasons, and making different reconstruction choices, and these differences fundamentally influence our interpretation of the temperature response to radiative forcing (c.f. Wang et al. 2015). This observation calls for a revival of paleo-reconstruction intercomparison projects (e.g. Ammann 2008; Graham and Wahl 2011; Anchukaitis and McKay 2015) in order to examine the behavior, strengths, and weaknesses of different proxy networks and reconstruction choices in a systematic and community-driven manner. Furthermore, such an effort would help identify regions with consistently large reconstruction uncertainties and indicate where to prioritize the development of new or the extension of existing tree-ring records.

5. Conclusions

In this study, we assimilate a small but highly temperature-sensitive tree-ring network based on expert assessment to reconstruct summer (MJJA) temperature anomalies from 750-1988 CE. Our method is skillful in the extratropical Northern Hemisphere and improves on a previous spatial reconstruction using the same network, thereby providing a new dataset with which to examine temperature dynamics and climate response to radiative forcing over the last millennium. In a set of pseudo-proxy experiments, we find that our method is sensitive to climate model biases, so we perform an ensemble of reconstructions using 10 different climate model priors. Reconstructed temperature anomalies are sensitive to the selection of the model prior when the proxy network becomes sparse, but the reconstructed spatial patterns and time series converge to consistent values as the number of sites in the NTREND proxy network increases. As one consequence of using static offline priors, our method underestimates temporal variability particularly when the proxy network becomes small, which argues for the future use of transient offline priors, online assimilation techniques in DA paleoclimate reconstructions, and expanded proxy development. There is also a need for continued development of proxy system forward models, particularly for the important MXD metric. The influence of the proxy network coverage on the reconstructions emphasizes the importance of analyzing reconstructed temperature anomalies in conjunction with estimates of their uncertainty. These uncertainty estimates emerge naturally for both spatial fields and time series from the DA posterior ensembles and are an enhancement over previous reconstructions using the NTREND dataset. In addition to gauging reconstruction validity, the uncertainty estimates identify regions that would benefit from additional proxy records and support the development of more millennial-length temperature-sensitive tree-ring records in treeline North America and eastern Siberia especially. Comparison of our reconstruction with other temperature CFRs indicates that reconstructed temperature anomalies have highly variable spatial patterns and magnitudes, even within similar reconstruction frameworks and proxy network. These different climate responses call for a renewed paleo-reconstruction intercomparison framework in which to systematically examine the effects of network selection across reconstruction techniques and prioritize regions for future record development.

Data availability statement. The NTREND proxy data and the earlier reconstructions are available from the NOAA NCEI World Data Service for Paleoclimatology (`https://www.ncdc.noaa.gov/paleo-search/study/19743`). The NTREND-DA ensemble reconstructions will be available from NOAA NCEI World Data Service for Paleoclimatology ([insert url here once accepted]). Model priors from the CMIP5 and CESM LME are available on the Earth System

Grid (https://esgf-node.llnl.gov/projects/esgf-llnl/) and the NCAR Climate Data Gateway (https://www.earthsystemgrid.org/), respectively. The data and code used to run these analyses and a function reproducing the results and figures from this paper are available at https://doi.org/10.5281/zenodo.3989941.

Acknowledgments. The authors acknowledge support from the Climate Program Office of the National Oceanographic and Atmospheric Administration (NOAA grants NA18OAR4310420 to KJA, NA18OAR4310426 to JEG and FZ, and NA18OAR4310422 to GJH). GJH also acknowledges support from the NSF through grant AGS–1702423. JMK and KJA were supported by NSF grant AGS-1803946. JET and JMK acknowledge support from NSF grant #AGS-1602301 and Heising-Simons Foundation grant #2016-05. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output.

APPENDIX

Data Assimilation Methods

A1. The Ensemble Kalman Filter

Our data assimilation method uses an ensemble Kalman filter approach (Evensen 1994; Steiger et al. 2014; Hakim et al. 2016) to solve the update equation:

$$\mathbf{X_a} = \mathbf{X_p} + \mathbf{K}(\mathbf{Y} - \mathbf{Y_e}) \tag{A1}$$

in each reconstructed annual time step. Here $\mathbf{X_p}$ is an initial ensemble of plausible climate states, an n x m matrix where n is the number of state variables and m is the number of ensemble members. $\mathbf{X_a}$ is the updated ensemble (the analysis), also an n x m matrix. $\mathbf{Y}$ is a d x m matrix of observed proxy values, where d is the number of available proxy records in a given time step. $\mathbf{Y_e}$ is a d x m matrix consisting of model estimates of the proxy values. Each row $\mathrm{y}_{\mathrm{e}_j}$ is determined by applying the forward model for the j^{th} proxy site to the ensemble via Equation 2. $\mathbf{K}$ is the Kalman Gain, an n by d matrix that weights the covariance of proxy sites with the target field by the uncertainties in the proxy observations and estimates.

We use an EnKF variant known as the ensemble square root Kalman filter (EnSRF; Andrews 1968), which removes the need for perturbed observations (Whitaker and Hamill 2002). Consequently, $\mathbf{Y}$ is a matrix with constant rows. In the EnSRF formulation, ensemble deviations are updated separately from the mean, as per:

$$\bar{\mathbf{x}}_\mathbf{a} = \bar{\mathbf{x}}_\mathbf{p} + \mathbf{K}(\bar{\mathbf{y}} - \bar{\mathbf{y}}_\mathbf{e}) \tag{A2}$$

$$\mathbf{X}'_\mathbf{a} = \mathbf{X}'_\mathbf{p} - \tilde{\mathbf{K}}\mathbf{Y}'_\mathbf{e} \tag{A3}$$

where an overbar ($\bar{\mathbf{x}}$) denotes an ensemble average, and a tick ($\mathbf{X}'$) indicates deviations from an ensemble mean. Here, the ensemble mean is updated via the Kalman gain (**K**):

$$\mathbf{K} = \mathrm{cov}(\mathbf{X_p}, \mathbf{Y_e}) \times [\mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}]^{-1} \tag{A4}$$

and the deviations are updated via an adjusted gain ($\tilde{\mathbf{K}}$):

$$\tilde{\mathbf{K}} = \mathrm{cov}(\mathbf{X_p}, \mathbf{Y_e}) \times [(\sqrt{\mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}})^{-1}]^{\mathrm{T}} [\sqrt{\mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}} + \sqrt{\mathbf{R}}]^{-1} \tag{A5}$$

Here, **R** denotes the observation error-covariance matrix (*d* x *d*). We do not consider correlated measurement errors in this study, so **R** is a diagonal matrix whose elements are the observation uncertainties determined from the variances of the residuals for the forward model regressions.

A2. Covariance Localization

We implement a covariance localization scheme, modifying the Kalman Gain equations to:

$$\mathbf{K} = \mathbf{W}_{\mathrm{loc}} \circ \mathrm{cov}(\mathbf{X_p}, \mathbf{Y_e}) \times [\mathbf{Y}_{\mathrm{loc}} \circ \mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}]^{\mathbf{-1}} \tag{A6}$$

and

$$\tilde{\mathbf{K}} = \mathbf{W}_{\mathrm{loc}} \circ \mathrm{cov}(\mathbf{X_p}, \mathbf{Y_e}) \times [(\sqrt{\mathbf{Y}_{\mathrm{loc}} \circ \mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}})^{\mathbf{-1}}]^{\mathbf{T}} [\sqrt{\mathbf{Y}_{\mathrm{loc}} \circ \mathrm{cov}(\mathbf{Y_e}, \mathbf{Y_e}) + \mathbf{R}} + \sqrt{\mathbf{R}}]^{\mathbf{-1}}. \tag{A7}$$

Here, $\mathbf{W}_{\mathrm{loc}}$ (*n* x *d*) and $\mathbf{Y}_{\mathrm{loc}}$ (*d* x *d*) are matrices of covariance localization weights applied to the covariance of proxy sites with model grid cells ($\mathbf{W}_{\mathrm{loc}}$) and proxy sites with one another ($\mathbf{Y}_{\mathrm{loc}}$). We implement localization weights as a fifth order Gaspari-Cohn polynomial (Gaspari and Cohn

1999) applied to the distance between proxy sites and model grid cells ($\mathbf{W}_{\mathrm{loc}}$) or proxy sites with one another ($\mathbf{Y}_{\mathrm{loc}}$). Weights are applied to covariance matrices via element-wise multiplication.

The localization radius is an important free parameter that must be assessed independently for different model priors, reconstruction targets, and proxy networks. Here, we select localization radii using a two step process. For a given model prior and target field, we first assimilate the proxy network from 1901-1988 CE using each localization radius from 250 km to 50,000 km in steps of 250 km and a run with no localization. We then determine the σ ratio of each reconstructed extratropical MJJA time series in a calibration interval. We find the σ ratio closest to 1 and record the associated localization radius as "optimal". We then calculate skill metrics for the extratropical MJJA time series over a validation interval using the reconstruction with the optimal radius.

To limit the sensitivity of this method to the calibration period (Christiansen et al. 2009), we perform this optimization using each set of 44 contiguous years from 1901-1988 CE once as a calibration interval and once as a validation interval. The final localization radius is the median of the 88 "optimal" radii, and the median validation skill metrics are reported.

a. Selection Criterion

In the development of this method, we tested an RMSE selection criterion in addition to σ ratios. We find that correlation coefficients, RMSE values, and mean biases of the reconstructed mean extratropical MJJA time series are all insensitive to the choice of selection criteria (Table 2, Table A1), but that σ ratios are more sensitive. Specifically, mean σ ratios are near 0.8 for the RMSE selection criterion, but rise to 1.11 for the σ ratio scheme. Since the σ ratio localization selection criteria brings the σ ratio skill metric closer to 1 without appreciably altering the other skill metrics, and because of the tendency for our DA method to underestimate temporal variability, we use a σ ratio selection criterion.

References

Acevedo, W., B. Fallah, S. Reich, and U. Cubasch, 2017: Assimilation of pseudo-tree-ring-width observations into an atmospheric general circulation model. *Climate of the Past*, **13 (5)**, 545–557.

Ammann, C., 2008: The paleoclimate reconstruction challenge. *PAGES News*, **16 (1)**, 4.

Ammann, C. M., and E. R. Wahl, 2007: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures. *Climatic Change*, **85 (1)**, 71–88.

Anchukaitis, K. J., and N. McKay, 2015: PAGES2k: Advances in climate field reconstructions. *PAGES Magazine*, **22 (2)**, 98.

Anchukaitis, K. J., and Coauthors, 2017: Last millennium Northern Hemisphere summer temperatures from tree rings: Part II, spatially resolved reconstructions. *Quaternary Science Reviews*, **163**, 1–22.

Andrews, A., 1968: A square root formulation of the Kalman covariance equations. *AIAA Journal*, **6 (6)**, 1165–1166.

Annan, J., and J. Hargreaves, 2012: Identification of climatic state with limited proxy data. *Climate of the Past*, **8 (4)**, 1141–1151.

Bhend, J., J. Franke, D. Folini, M. Wild, and S. Brönnimann, 2012: An ensemble-based approach to climate reconstructions. *Climate of the Past*, **8 (3)**, 963–976.

Björklund, J., B. E. Gunnarson, K. Seftigen, J. Esper, and Coauthors, 2014: Blue intensity and density from northern fennoscandian tree rings, exploring the potential to improve summer temperature reconstructions with earlywood information. *Climate of the Past*, **10 (2)**, 877–885.

Böhm, R., P. D. Jones, J. Hiebl, D. Frank, M. Brunetti, and M. Maugeri, 2010: The early instrumental warm-bias: a solution for long central european temperature series 1760–2007. *Climatic Change*, **101 (1-2)**, 41–67.

Chen, Z., and Coauthors, 2003: Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, **182 (1)**, 1–69.

Christiansen, B., 2008: Volcanic eruptions, large-scale modes in the northern hemisphere, and the el niño–southern oscillation. *Journal of Climate*, **21 (5)**, 910–922.

Christiansen, B., and F. C. Ljungqvist, 2017: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia. *Reviews of Geophysics*, **55 (1)**, 40–96.

Christiansen, B., T. Schmith, and P. Thejll, 2009: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *Journal of Climate*, **22 (4)**, 951–976.

Comboul, M., J. Emile-Geay, G. J. Hakim, and M. N. Evans, 2015: Paleoclimate sampling as a sensor placement problem. *Journal of Climate*, **28**, 7717–7740, doi:10.1175/JCLI-D-14-00802.1.

Cook, E. R., K. J. Anchukaitis, B. M. Buckley, R. D. D'Arrigo, G. C. Jacoby, and W. E. Wright, 2010a: Asian monsoon failure and megadrought during the last millennium. *Science*, **328 (5977)**, 486–489.

Cook, E. R., K. R. Briffa, and P. D. Jones, 1994: Spatial regression methods in dendroclimatology: a review and comparison of two techniques. *International Journal of Climatology*, **14 (4)**, 379–402.

Cook, E. R., P. J. Krusic, K. J. Anchukaitis, B. M. Buckley, T. Nakatsuka, M. Sano, and Coauthors, 2013: Tree-ring reconstructed summer temperature anomalies for temperate East Asia since 800 CE. *Climate Dynamics*, **41 (11-12)**, 2957–2972.

Cook, E. R., D. M. Meko, D. W. Stahle, and M. K. Cleaveland, 1999: Drought reconstructions for the continental United States. *Journal of Climate*, **12 (4)**, 1145–1162.

Cook, E. R., R. Seager, R. R. Heim Jr, R. S. Vose, C. Herweijer, and C. Woodhouse, 2010b: Megadroughts in North America: Placing IPCC projections of hydroclimatic change in a long-term palaeoclimate context. *Journal of Quaternary Science*, **25 (1)**, 48–61.

Cort, G. D., M. Chevalier, S. L. Burrough, C. Y. Chen, and S. P. Harrison, 2021: An uncertainty-focused database approach to extract spatiotemporal trends from qualitative and discontinuous lake-status histories. *Quaternary Science Reviews*, **258**, 106 870, doi:10.1016/j.quascirev.2021.106870, URL https://doi.org/10.1016/j.quascirev.2021.106870.

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, **140 (683)**, 1935–1944.

De Silva, S. L., and G. A. Zielinski, 1998: Global influence of the AD 1600 eruption of Huaynaputina, Peru. *Nature*, **393 (6684)**, 455–458.

Dee, S. G., N. J. Steiger, J. Emile-Geay, and G. J. Hakim, 2016: On the utility of proxy system models for estimating climate states over the Common Era. *Journal of Advances in Modeling Earth Systems*, **8 (3)**, 1164–1179.

Delworth, T. L., and M. E. Mann, 2000: Observed and simulated multidecadal variability in the Northern Hemisphere. *Climate Dynamics*, **16 (9)**, 661–676.

Dubinkina, S., and H. Goosse, 2013: An assessment of particle filtering methods and nudging for climate state reconstructions. *Climate of the Past*, **9 (3)**, 1141–1152.

Esper, J., D. C. Frank, R. J. Wilson, and K. R. Briffa, 2005: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. *Geophysical Research Letters*, **32 (7)**.

Esper, J., and Coauthors, 2018: Large-scale, millennial-length temperature reconstructions from tree-rings. *Dendrochronologia*, **50**, 81–90, doi:10.1016/j.dendro.2018.06.001.

Evans, M., A. Kaplan, M. Cane, and R. Villalba, 2001: Globality and optimality in climate field reconstructions from proxy data. *Interhemispheric Climate Linkages*, Elsevier, 53–XV.

Evans, M. N., A. Kaplan, and M. A. Cane, 2002: Pacific sea surface temperature field reconstruction from coral δ^{18}O data using reduced space objective analysis. *Paleoceanography*, **17 (1)**, doi: 10.1029/2000PA000590.

Evans, M. N., S. E. Tolwinski-Ward, D. M. Thompson, and K. J. Anchukaitis, 2013: Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews*, **76**, 16–28.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, **99 (C5)**, 10 143–10 162.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, **53 (4)**, 343–367.

Frank, D., U. Büntgen, R. Böhm, M. Maugeri, and J. Esper, 2007: Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at a moving target. *Quaternary Science Reviews*, **26 (25-28)**, 3298–3310.

Frank, D., J. Esper, E. Zorita, and R. Wilson, 2010: A noodle, hockey stick, and spaghetti plate: a perspective on high-resolution paleoclimatology. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (4)**, 507–516.

Franke, J., V. Valler, S. Brönnimann, R. Neukom, and F. Jaume-Santero, 2020: The importance of input data quality and quantity in climate field reconstructions – results from the assimilation of various tree-ring collections. *Climate of the Past*, **16 (3)**, 1061–1074, doi: 10.5194/cp-16-1061-2020.

Fritts, H. C., 1991: *Reconstructing large-scale climatic patterns from tree-ring data: a diagnostic analysis*. University of Arizona Press.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125 (554)**, 723–757.

Gennaretti, F., D. Arseneault, A. Nicault, L. Perreault, and Y. Bégin, 2014: Volcano-induced regime shifts in millennial tree-ring chronologies from northeastern North America. *Proceedings of the National Academy of Sciences*, **111 (28)**, 10 077–10 082.

Gill, E. C., B. Rajagopalan, P. Molnar, and T. M. Marchitto, 2016: Reduced-dimension reconstruction of the equatorial pacific sst and zonal wind fields over the past 10,000 years using mg/ca and alkenone records. *Paleoceanography*, **31 (7)**, 928–952.

Goosse, H., 2016: An additional step toward comprehensive paleoclimate reanalyses. *Journal of Advances in Modeling Earth Systems*, n/a–n/a, doi:10.1002/2016MS000739, URL http://dx.doi.org/10.1002/2016MS000739.

Goosse, H., 2017: Reconstructed and simulated temperature asymmetry between continents in both hemispheres over the last centuries. *Climate Dynamics*, **48 (5-6)**, 1483–1501.

Goosse, H., J. Guiot, M. E. Mann, S. Dubinkina, and Y. Sallaz-Damaz, 2012: The Medieval Climate Anomaly in Europe: Comparison of the summer and annual mean signals in two reconstructions and in simulations with data assimilation. *Global and Planetary Change*, **84**, 35–47.

Graham, N., and E. Wahl, 2011: Paleoclimate reconstruction challenge. *PAGES/CLIVAR Newsletter*, **19 (2)**, 71–72.

Guillet, S., and Coauthors, 2017: Climate response to the Samalas volcanic eruption in 1257 revealed by proxy records. *Nature Geoscience*, **10 (2)**, 123–128.

Guillot, D., B. Rajaratnam, and J. Emile-Geay, 2015: Statistical paleoclimate reconstructions via Markov random fields. *The Annals of Applied Statistics*, **9 (1)**, 324–352.

Hakim, G. J., J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger, and W. A. Perkins, 2016: The Last Millennium Climate Reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres*, **121 (12)**, 6745–6764.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations–the CRU TS3.10 Dataset. *International Journal of Climatology*, **34 (3)**, 623–642.

Haurwitz, M. W., and G. W. Brier, 1981: A critique of the superposed epoch analysis method: its application to solar–weather relations. *Monthly Weather Review*, **109 (10)**, 2074–2079.

Hegerl, G., and P. Stott, 2014: From past to future warming. *Science*, **343 (6173)**, 844–845.

Hegerl, G. C., 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. *Journal of Climate*, **20 (4)**, 650–666.

Hegerl, G. C., T. J. Crowley, S. K. Baum, K.-Y. Kim, and W. T. Hyde, 2003: Detection of volcanic, solar and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature. *Geophysical Research Letters*, **30 (5)**, doi:10.1029/2002GL016635.

Hegerl, G. C., K. Hasselmann, U. Cubasch, J. F. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, 1997: Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dynamics*, **13 (9)**, 613–634.

Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129 (1)**, 123–137.

Hughes, M., and C. Ammann, 2009: The future of the past—an earth system framework for high resolution paleoclimatology: editorial essay. *Climatic Change*, **94 (3-4)**, 247–259.

Kaufman, D., 2014: A community-driven framework for climate reconstructions. *Eos, Transactions American Geophysical Union*, **95 (40)**, 361–362, doi:10.1002/2014eo400001.

Lavigne, F., and Coauthors, 2013: Source of the great AD 1257 mystery eruption unveiled, Samalas volcano, Rinjani Volcanic Complex, Indonesia. *Proceedings of the National Academy of Sciences*, **110 (42)**, 16 742–16 747.

Lean, J. L., and D. H. Rind, 2008: How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophysical Research Letters*, **35 (18)**, doi: 10.1029/2008GL034864.

Liu, H., Z. Liu, and F. Lu, 2017: A systematic comparison of particle filter and EnKF in assimilating time-averaged observations. *Journal of Geophysical Research: Atmospheres*, **122 (24)**, 13–155.

Mann, M. E., R. S. Bradley, and M. K. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, **392 (6678)**, 779–787.

Mann, M. E., and S. Rutherford, 2002: Climate reconstruction using 'Pseudoproxies'. *Geophysical Research Letters*, **29 (10)**, 139–1.

Mann, M. E., and Coauthors, 2009: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science*, **326 (5957)**, 1256–1260.

Marsland, S. J., H. Haak, J. H. Jungclaus, M. Latif, and F. Röske, 2003: The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modelling*, **5 (2)**, 91–127.

Masson-Delmotte, V., and Coauthors, 2013: Information from Paleoclimate Archives. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 383–464.

Matsikaris, A., M. Widmann, and J. H. Jungclaus, 2015: On-line and off-line data assimilation in palaeoclimatology: a case study. *Climate of the Past*, **11**, 81–93.

McCarroll, D., E. Pettigrew, A. Luckman, F. Guibal, and J.-L. Edouard, 2002: Blue reflectance provides a surrogate for latewood density of high-latitude pine tree rings. *Arctic, Antarctic, and Alpine Research*, **34 (4)**, 450–453.

Meehl, G. A., W. M. Washington, C. M. Ammann, J. M. Arblaster, T. Wigley, and C. Tebaldi, 2004: Combinations of natural and anthropogenic forcings in twentieth-century climate. *Journal of Climate*, **17 (19)**, 3721–3727.

Meko, D., 1997: Dendroclimatic reconstruction with time varying predictor subsets of tree indices. *Journal of Climate*, **10**, 687–696.

Neukom, R., N. Steiger, J. J. Gómez-Navarro, J. Wang, and J. P. Werner, 2019: No evidence for globally coherent warm and cold periods over the preindustrial Common Era. *Nature*, **571 (7766)**, 550–554.

Okazaki, A., and K. Yoshimura, 2017: Development and evaluation of a system of proxy data assimilation for paleoclimate reconstruction. *Climate of the Past*, **13 (4)**, 379–393.

Oke, P. R., J. S. Allen, R. N. Miller, G. D. Egbert, and P. M. Kosro, 2002: Assimilation of surface velocity data into a primitive equation coastal ocean model. *Journal of Geophysical Research: Oceans*, **107 (C9)**, doi:10.1029/2000JC000511.

Otto-Bliesner, B. L., and Coauthors, 2016: Climate variability and change since 850 CE: An ensemble approach with the Community Earth System Model. *Bulletin of the American Meteorological Society*, **97 (5)**, 735–754.

PAGES2k Consortium, 2013: Continental-scale temperature variability during the past two millennia. *Nature Geoscience*, **6 (5)**, 339–346.

PAGES2k Consortium, 2017: A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data*, **4**, 170 088.

Parker, D., 1994: Effects of changing exposure of thermometers at land stations. *International Journal of Climatology*, **14 (1)**, 1–31.

Perkins, W. A., and G. J. Hakim, 2017: Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. *Climate of the Past*, **13 (5)**, 421–436.

Phipps, S. J., and Coauthors, 2013: Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years. *Journal of Climate*, **26 (18)**, 6915–6936.

Robock, A., 2000: Volcanic eruptions and climate. *Reviews of geophysics*, **38 (2)**, 191–219.

Rohde, R., and Coauthors, 2013: Berkeley Earth temperature averaging process. *Geoinformatics and Geostatistics: An Overview*, **1 (2)**, 20–100.

Rutherford, S., M. Mann, T. Delworth, and R. Stouffer, 2003: Climate field reconstruction under stationary and nonstationary forcing. *Journal of Climate*, **16 (3)**, 462–479.

Rydval, M., L.-Å. Larsson, L. McGlynn, B. E. Gunnarson, N. J. Loader, G. H. Young, and R. Wilson, 2014: Blue intensity for dendroclimatology: should we have the blues? Experiments from Scotland. *Dendrochronologia*, **32 (3)**, 191–204.

Schneider, D. P., C. M. Ammann, B. L. Otto-Bliesner, and D. S. Kaufman, 2009: Climate response to large, high-latitude and low-latitude volcanic eruptions in the Community Climate System Model. *Journal of Geophysical Research: Atmospheres*, **114 (D15)**.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, **14 (5)**, 853–871.

Schurer, A. P., G. C. Hegerl, M. E. Mann, S. F. B. Tett, and S. J. Phipps, 2013: Separating forced from chaotic climate variability over the past millennium. *Journal of Climate*, **26 (18)**, 6954–6973, doi:10.1175/jcli-d-12-00826.1.

Schweingruber, F., H. Fritts, O. Bräker, L. Drew, and E. Schär, 1978: The x-ray technique as applied to dendroclimatology. *Tree-Ring Bulletin*, **38**, 61–91.

Seager, R., N. Graham, C. Herweijer, A. L. Gordon, Y. Kushnir, and E. Cook, 2007: Blueprints for medieval hydroclimate. *Quaternary Science Reviews*, **26 (19-21)**, 2322–2336.

Sigl, M., and Coauthors, 2015: Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature*, **523 (7562)**, 543–549.

Smerdon, J. E., 2012: Climate models as a test bed for climate reconstruction methods: pseudo-proxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, **3 (1)**, 63–77.

Smerdon, J. E., A. Kaplan, E. Zorita, J. F. González-Rouco, and M. Evans, 2011: Spatial performance of four climate field reconstruction methods targeting the Common Era. *Geophysical Research Letters*, **38 (11)**, doi:10.1029/2011GL047372.

Smerdon, J. E., and H. N. Pollack, 2016: Reconstructing Earth's surface temperature over the past 2000 years: the science behind the headlines. *Wiley Interdisciplinary Reviews: Climate Change*, **7 (5)**, 746–771.

Solomon, A., and Coauthors, 2011: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: implications for prediction. *Bulletin of the American Meteorological Society*, **92 (2)**, 141–156.

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. *Journal of Climate*, **27 (1)**, 426–441.

Steiger, N. J., and J. E. Smerdon, 2017: A pseudoproxy assessment of data assimilation for reconstructing the atmosphere–ocean dynamics of hydroclimate extremes. *Climate of the Past*, **13 (10)**, 1435–1449.

Steiger, N. J., J. E. Smerdon, E. R. Cook, and B. I. Cook, 2018: A reconstruction of global hydroclimate and dynamical variables over the common era. *Scientific Data*, **5**, doi:10.1086/sdata.2018.86.

Stenchikov, G., K. Hamilton, R. J. Stouffer, A. Robock, V. Ramaswamy, B. Santer, and H.-F. Graf, 2006: Arctic Oscillation response to volcanic eruptions in the IPCC AR4 climate models. *Journal of Geophysical Research: Atmospheres*, **111 (D7)**.

Stevens, B., and Coauthors, 2013: Atmospheric component of the MPI-M earth system model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, **5 (2)**, 146–172.

Stott, P. A., N. P. Gillett, G. C. Hegerl, D. J. Karoly, D. A. Stone, X. Zhang, and F. Zwiers, 2010: Detection and attribution of climate change: a regional perspective. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (2)**, 192–211.

Stott, P. A., and G. S. Jones, 2009: Variability of high latitude amplification of anthropogenic warming. *Geophysical Research Letters*, **36 (10)**, doi:10.1029/2009GL037698.

Stott, P. A., and S. F. Tett, 1998: Scale-dependent detection of climate change. *Journal of Climate*, **11 (12)**, 3282–3294.

Sundqvist, H. S., and Coauthors, 2014: Arctic Holocene proxy climate database–new approaches to assessing geochronological accuracy and encoding climate variables. *Climate of the Past*, **10**, 1605–1631, doi:10.5194/cp-10-1605-2014.

Tardif, R., and Coauthors, 2019: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, **15 (4)**, 1251–1273.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, **93 (4)**, 485–498.

Tingley, M. P., P. F. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam, 2012: Piecing together the past: Statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews*, **35**, 1–22.

Valler, V., J. Franke, and S. Brönnimann, 2019: Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation. *Climate of the Past*, **15 (4)**, 1427–1441, doi:10.5194/cp-15-1427-2019.

Van der Schrier, G., and J. Barkmeijer, 2005: Bjerknes' hypothesis on the coldness during AD 1790–1820 revisited. *Climate Dynamics*, **25 (5)**, 537–553.

Wang, J., J. Emile-Geay, D. Guillot, N. P. McKay, and B. Rajaratnam, 2015: Fragility of reconstructed temperature patterns over the common era: Implications for model evaluation. *Geophysical Research Letters*, **42 (17)**, 7162–7170.

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

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, **130 (7)**, 1913–1924.

Widmann, M., H. Goosse, G. van der Schrier, R. Schnur, and J. Barkmeijer, 2010: Using data assimilation to study extratropical northern hemisphere climate over the last millennium. *Climate of the Past*, **6 (5)**, 627–644, doi:10.5194/cp-6-627-2010.

Wikle, C. K., and L. M. Berliner, 2007: A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, **230 (1-2)**, 1–16.

Wilson, R., and Coauthors, 2016: Last millennium northern hemisphere summer temperatures from tree rings: Part I: The long term context. *Quaternary Science Reviews*, **134**, 1–18.

Wilson, R., and Coauthors, 2019: Improved dendroclimatic calibration using blue intensity in the southern Yukon. *The Holocene*, **29 (11)**, 1817–1830.

Zhu, F., J. Emile-Geay, G. J. Hakim, J. King, and K. J. Anchukaitis, 2020: Resolving the differences in the simulated and reconstructed temperature response to volcanism. *Geophysical Research Letters*, **47 (8)**, e2019GL086 908.

Zorita, E., F. González-Rouco, and S. Legutke, 2003: Testing the approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. *Journal of Climate*, **16 (9)**, 1378–1390.

LIST OF TABLES

Table 1. Summary of climate models used to construct data assimilation prior ensembles. Climate models are listed along with the identifying acronym used in this study. The years of available output are provided with the experiment used to generate them. The size of the model prior generated from these years is also provided. Taylor et al. (2012) provide more details on the PMIP3 and CMIP5 experiments, and Otto-Bliesner et al. (2016) describe the LME.

Model	Acronym	Years: Experiment	Sample size (m)
BCC-CSM1-1	BCC	850-2000: past1000	1151
CCSM4	CCSM4	850-1850: past1000 1851-2005: historical	1156
CESM1.1-CAM5	CESM	850-2005: LME full-forcing	1156
CSIRO-Mk3L-1-2	CSIRO	851-1850: past1000 1851-2000: historical	1150
FGOALS-gl	FGOALS	1000-1999: past1000	1000
HadCM3	HadCM3	850-1850: past1000 1859-2000: historical	1147
IPSL-CM5A-LR	IPSL	850-1850: past1000 1851-2005: historical	1156
MIROC-ESM	MIROC	850-1849: past1000 1850-2005: historical	1156
MPI-ESM-P	MPI	850-1849: past1000 1850-2005: historical	1156
MRI-CGCM3	MRI	850-1850: past1000 1850-2005: historical	1156

TABLE 2. Calibrated localization radii. Localization radii for individual model priors are selected using the radius search and calibration-validation procedure detailed in Appendix A1. Skill metrics are the median values obtained for the mean extratropical MJJA time series relative to BEST for the set of validation periods.

Model	Localization Radius (km)	Correlation	RMSE (°C)	σ Ratio	Mean Bias (°C)
BCC	∞	0.69	0.18	1.03	0.05
CCSM4	16500	0.72	0.19	1.18	0.07
CESM	∞	0.72	0.18	1.08	0.06
CSIRO	∞	0.70	0.19	1.18	0.05
F-GOALS	∞	0.70	0.18	1.02	0.07
HadCM3	∞	0.69	0.19	1.18	0.05
IPSL	12750	0.70	0.19	1.19	0.06
MIROC	26375	0.71	0.19	1.18	0.06
MPI	27625	0.69	0.20	1.18	0.06
MRI	∞	0.71	0.17	1.01	0.05

TABLE 3. Temperature field reconstructions used to compare spatial patterns of climate response to radiative forcings in this study. We provide a reference for each CFR along with the name used in this study. We also note the maximum size of the proxy network used in each study along with the target temperature fields.

Name	Reference	Network Size	Reconstruction Target
NTREND - DA	This study	54	MJJA
NTREND - PPR	Anchukaitis et al. (2017)	54	MJJA
Guillet 2017	Guillet et al. (2017)	28	Highpass JJA
Zhu 2020	Zhu et al. (2020)	395	JJA
LMR 2.1	Tardif et al. (2019)	544	Annual (Jan. - Dec.)
Neukom (DA)	Neukom et al. (2019)	210	Annual (April - March)

1007 TABLE 4. Withheld proxy verification statistics for individual models. Reported skill metrics are the median
1008 for all individual proxy comparisons over the 54 leave-one-out assimilations.

Model	Correlation	RMSE	σ Ratio	Mean Bias °C
BCC	0.53	0.98	0.42	0.12
CCSM4	0.52	0.98	0.42	0.06
CESM	0.50	1.03	0.35	0.27
CSIRO	0.54	1.01	0.31	0.13
F-GOALS	0.47	1.04	0.34	0.06
HadCM3	0.49	1.03	0.39	0.25
IPSL	0.53	1.00	0.38	0.08
MIROC	0.53	1.01	0.37	0.25
MPI	0.53	0.99	0.39	0.11
MRI	0.55	0.98	0.32	0.16

Table A1. As in Table 2, but using the RMSE optimization scheme.

Model	Localization Radius (km)	Correlation	RMSE (°C)	σ Ratio	Mean Bias (°C)
BCC	18875	0.71	0.17	0.78	0.06
CCSM4	7375	0.71	0.18	0.81	0.07
CESM	15750	0.71	0.18	0.84	0.07
CSIRO	15750	0.70	0.18	0.80	0.06
F-GOALS	19000	0.72	0.18	0.77	0.08
HadCM3	13375	0.70	0.18	0.82	0.06
IPSL	6750	0.70	0.18	0.80	0.07
MIROC	11125	0.71	0.18	0.84	0.07
MPI	10250	0.70	0.18	0.80	0.07
MRI	20250	0.71	0.17	0.78	0.06

LIST OF FIGURES

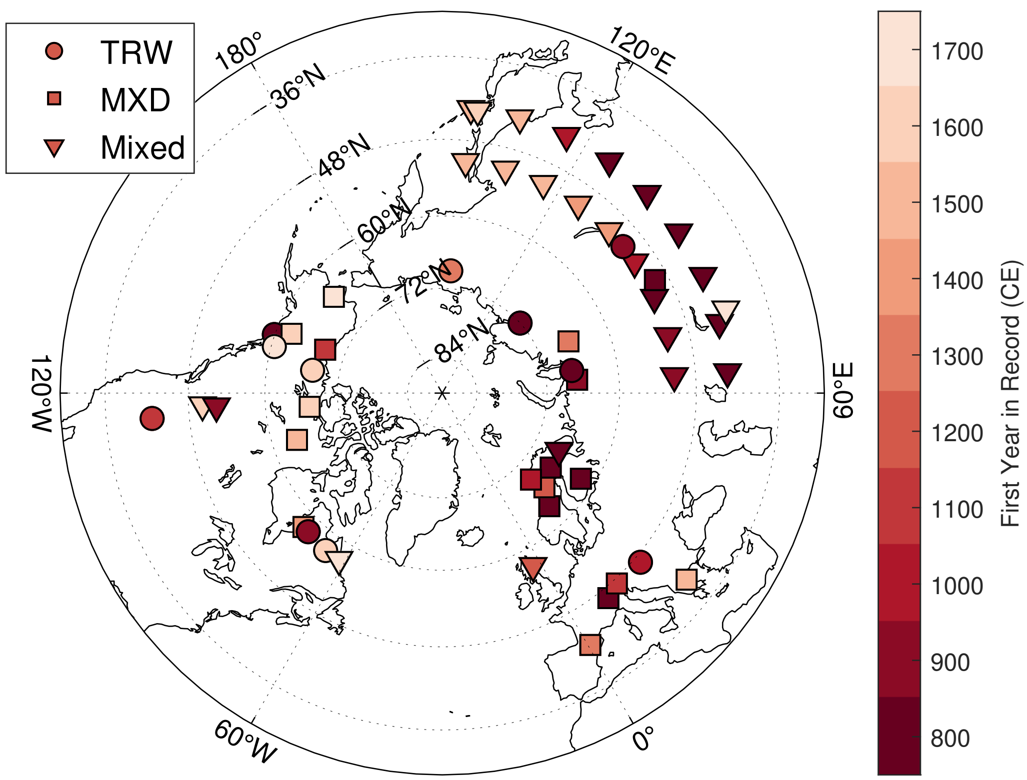


FIG. 1. Locations of the 54 NTREND sites (Wilson et al. 2016). NTREND records were developed using ring-width data (TRW; circles), maximum latewood density (MXD; squares), or a mix of TRW, MXD, and blue intensity (Mixed; triangles). Marker color denotes the century in which each record begins.

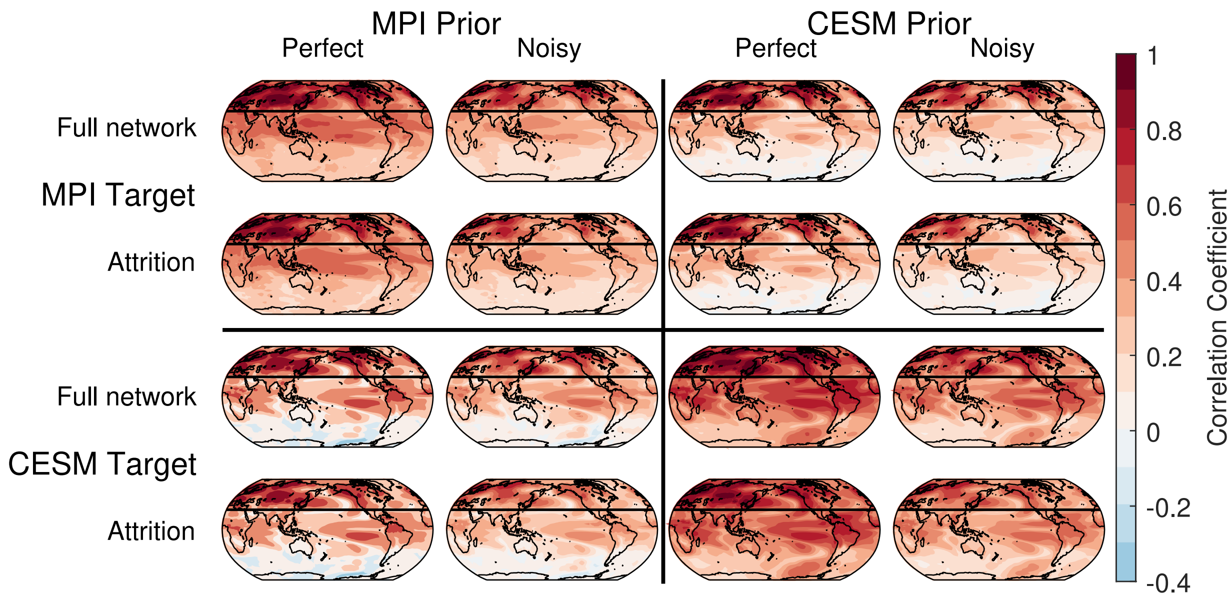


FIG. 2. Local Pearson's correlation coefficients of pseudo-proxy reconstruction temperature anomalies with the target fields. Correlation coefficients are calculated over the period 850-1988 CE. Major rows indicate the model used to generate the target field, and major columns show the model used to build the initial ensemble for each assimilation. Minor rows designate whether the proxy network exhibits no time attrition or realistic time attrition. Minor columns indicate whether reconstructions use perfect or noisy proxies. The top-left and bottom-right quadrants display the perfect-model experiments, while the top-right and bottom-left quadrants show the biased-model cases. The black line in each map indicates 30°N.

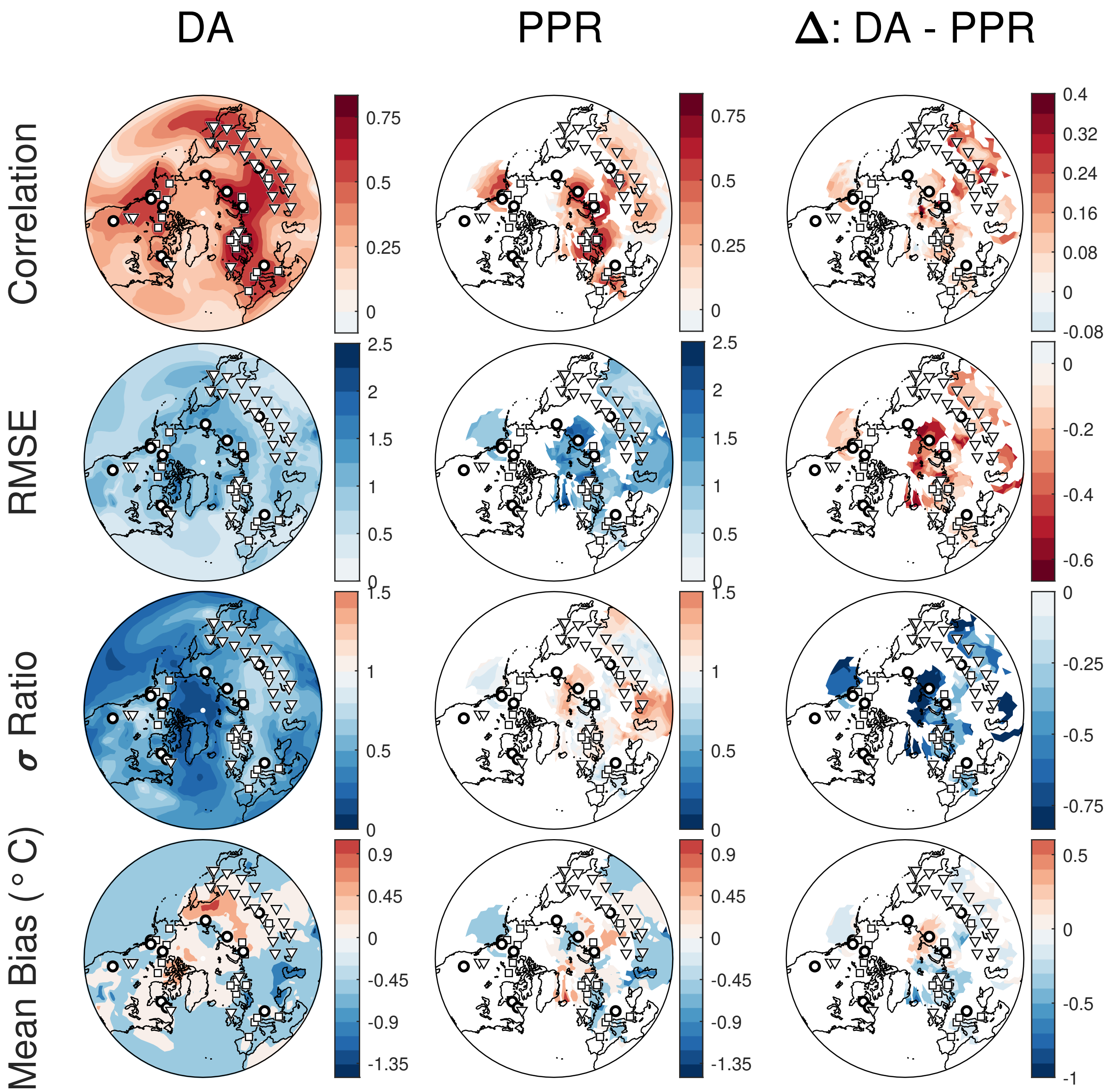


Fig. 3. Pseudo-proxy reconstruction skill for DA (left column), PPR (middle), and a comparison of the two (right). Skill metrics are relative to a CESM target field using noisy proxies and realistic temporal attrition. DA results are for a biased-model MPI prior. All skill metrics are computed over the period 850-1988 CE. In order the rows detail local Pearson's correlation coefficients, RMSE values, temporal standard deviation (σ) ratios, and mean biases. Comparison plots show DA skill minus PPR skill. The comparison plot of σ ratios only considers grid points where σ is underestimated in both the DA and PPR reconstruction.

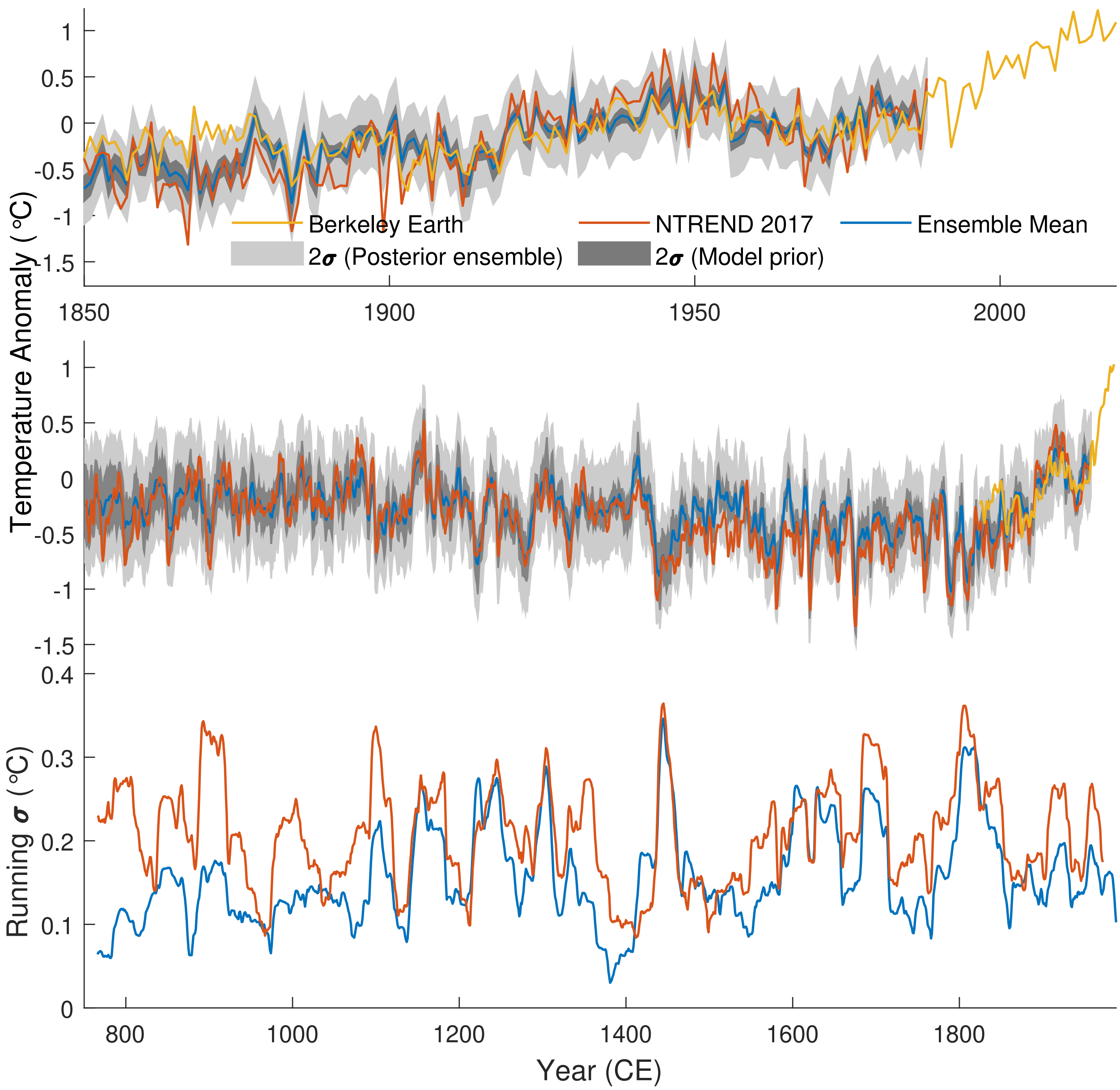


FIG. 4. Extratropical MJJA time series for the multi-model mean reconstruction (blue), Berkeley Earth instrumental records (yellow), and Anchukaitis et al. (2017) (red). We provide two different measures of uncertainty for the DA time series: the average of the 2σ posterior ensemble width taken over the 10 reconstruction (light grey), and the 2σ width of variability arising from prior model selection (dark grey). Reconstructed temperature anomalies are shown in Celsius for the instrumental era (top), and full reconstruction (middle). A three year moving average has been applied to the time series in the middle panel. The bottom panel displays the 31-year, running standard deviation of the DA ensemble-mean and Anchukaitis et al. (2017) time series.

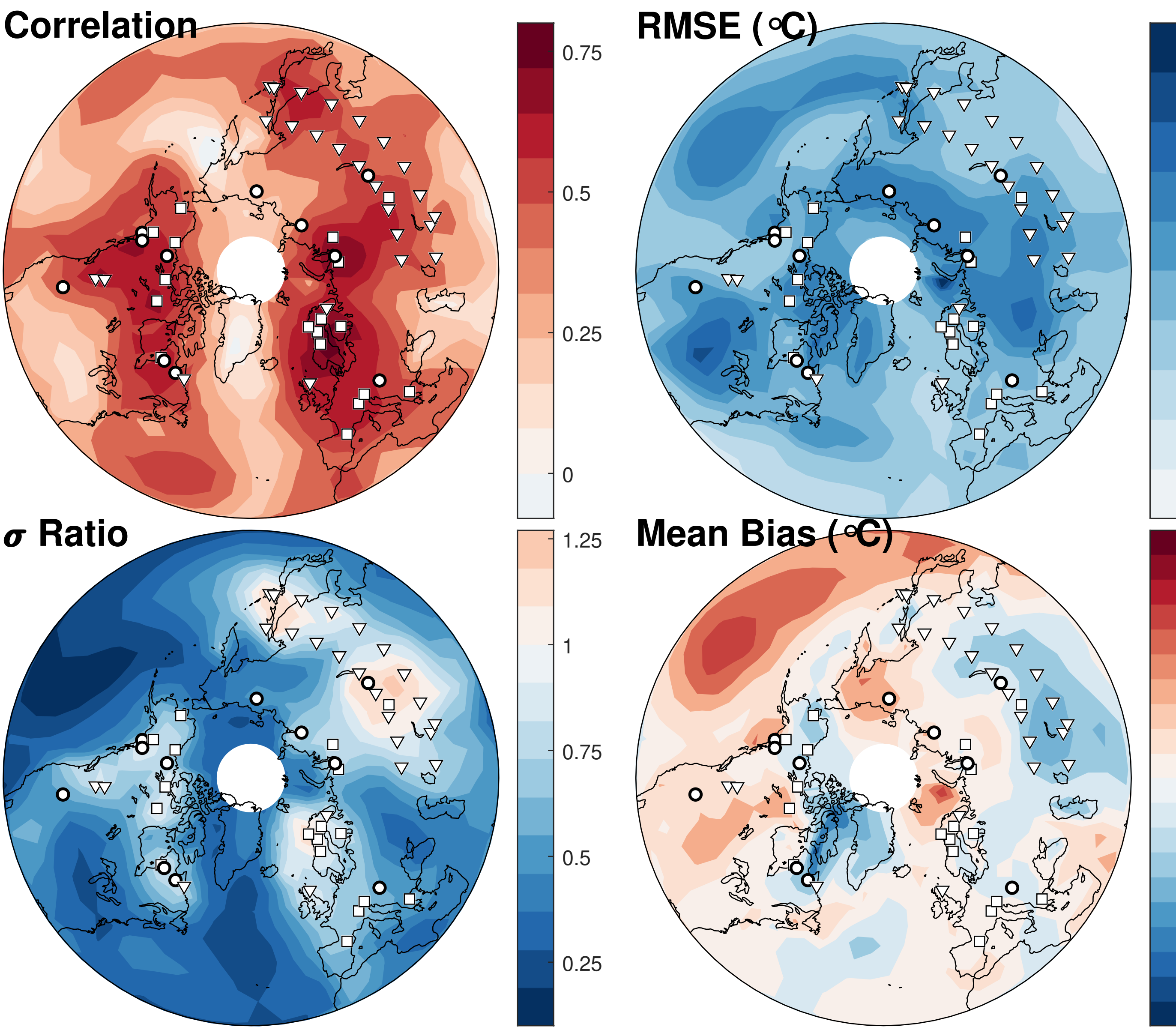
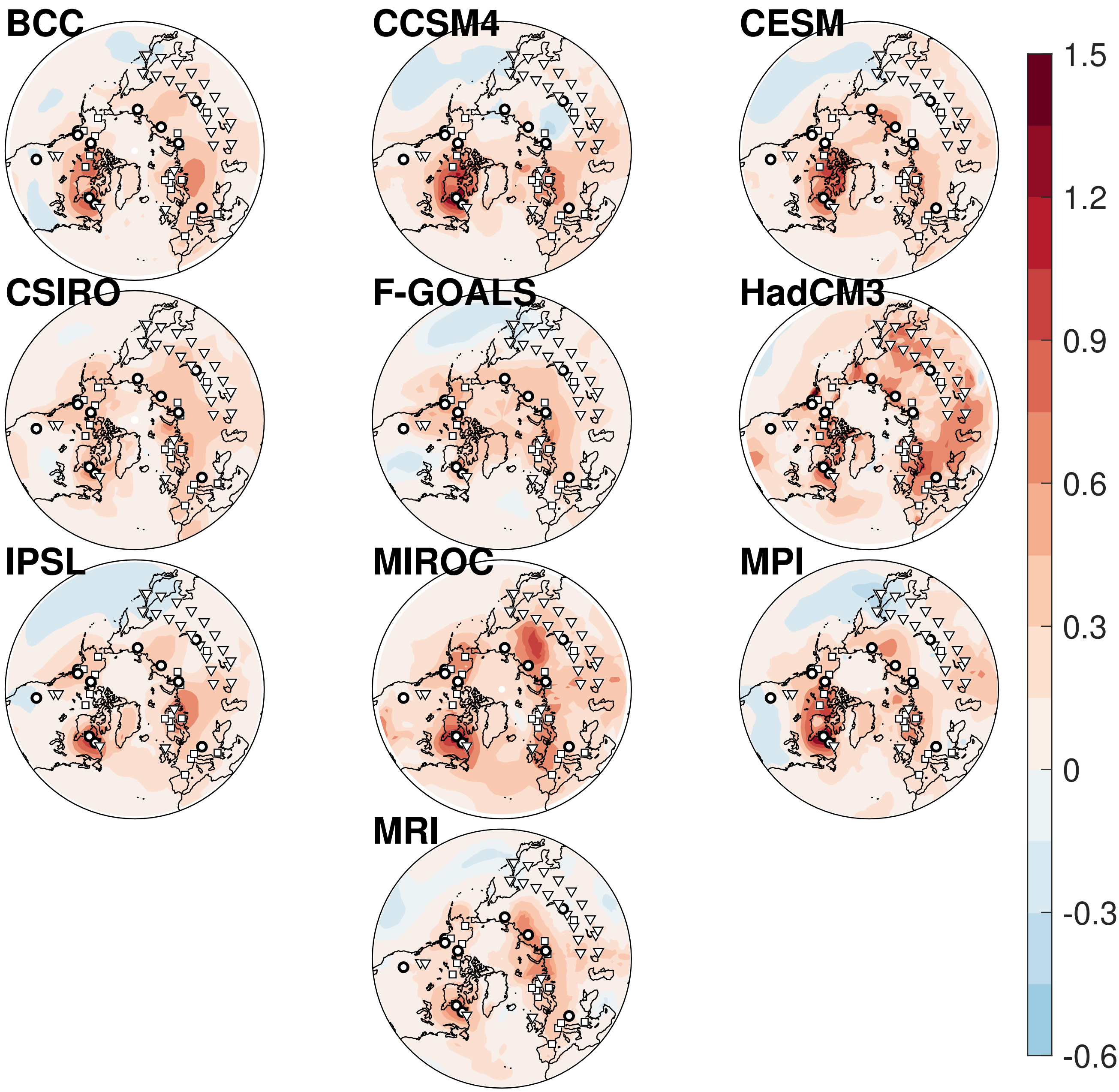


FIG. 5. Spatial skill metrics for the multi-model mean reconstruction. Maps detail Pearson correlation coefficients (top left), RMSE values (top right), σ ratios (bottom left), and mean biases (bottom right) of reconstructed grid point time series relative to the Berkeley Earth instrumental dataset over the period 1901-1988 CE. White markers show the proxy network and marker symbols follow the convention in Figure 1.



1091 FIG. 6. Reconstructed temperature anomalies (in Celsius) between the MCA (950-1250 CE) and LIA (1450-1850 CE) for the DA reconstructions. Each map shows the results for a particular model prior.

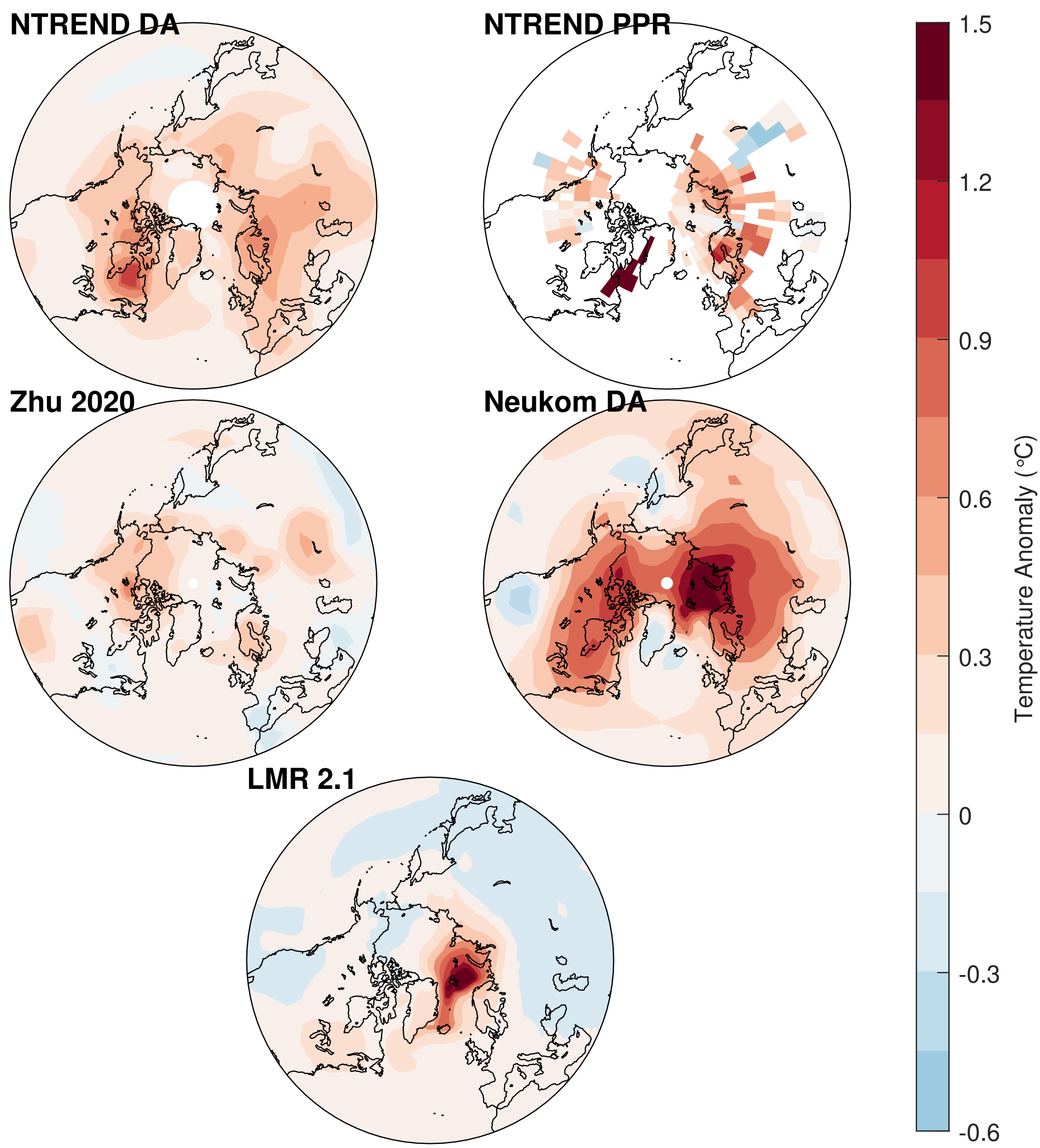


FIG. 7. As in 6, but for the temperature CFRs summarized in Table 3.

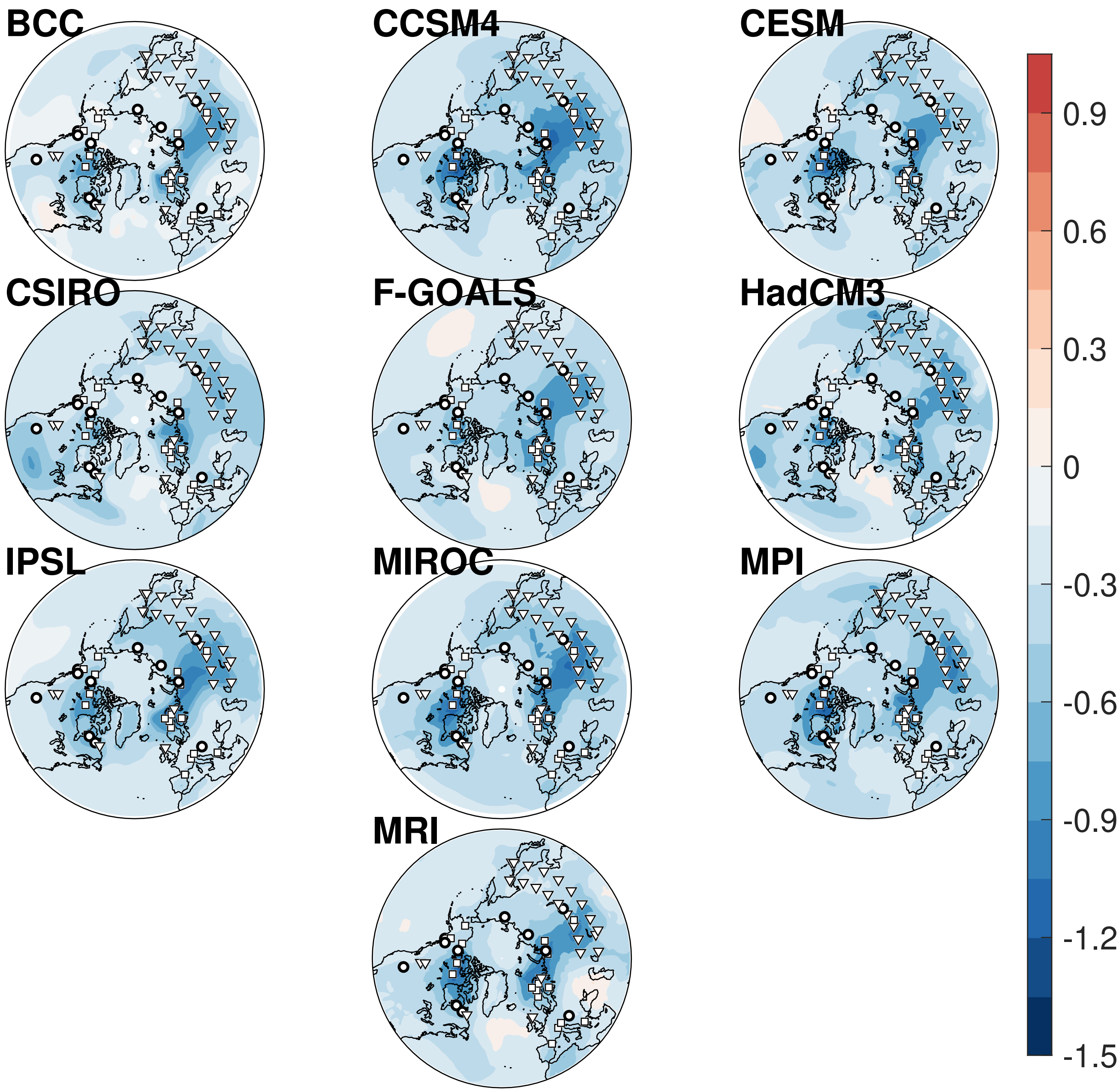
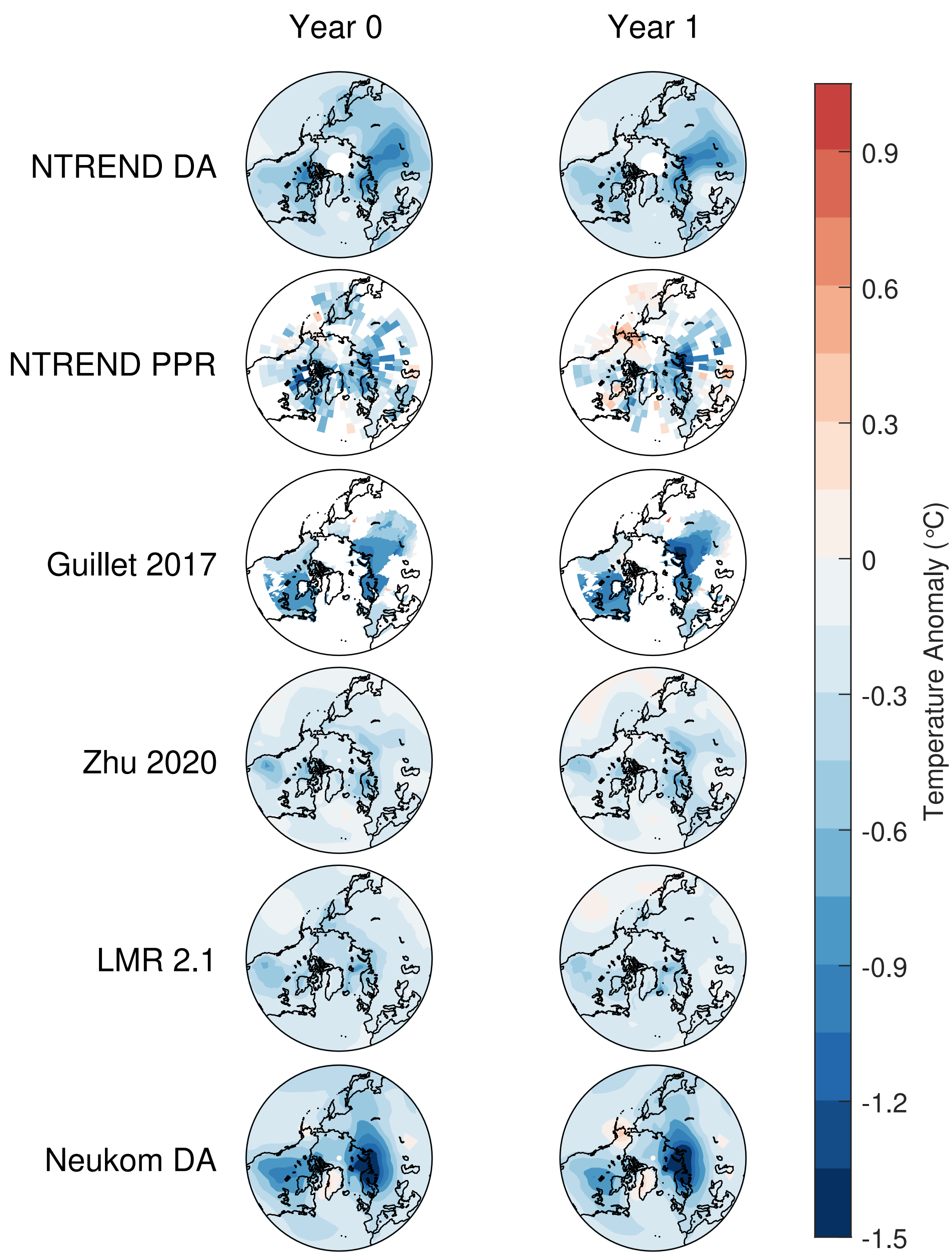


FIG. 8. Composite mean maps of the reconstructed temperature response in years containing a major tropical volcanic event. Events (N=20) are selected as tropical eruptions with a global forcing magnitude equal or larger than the 1884 Krakatoa eruption: this set consists of 916, 1108, 1171, 1191, 1230, 1258, 1276, 1286, 1345, 1453, 1458, 1595, 1601, 1641, 1695, 1809, 1815, 1832, 1836, and 1884 CE (Sigl et al. 2015; Anchukaitis et al. 2017). Temperature anomalies (in Celsius) are determined relative to the mean temperature of the five years preceding each volcanic event. Each map shows the results for a particular model prior.



1099 FIG. 9. As in Figure 8, but for the temperature CFRs summarized in Table 3 (rows). We only show grid points
1100 with reconstructed values for at least 6 eruptions. Maps show the composite mean response in years with a major
1101 tropical eruption (left), and in the year following a major eruption (right).

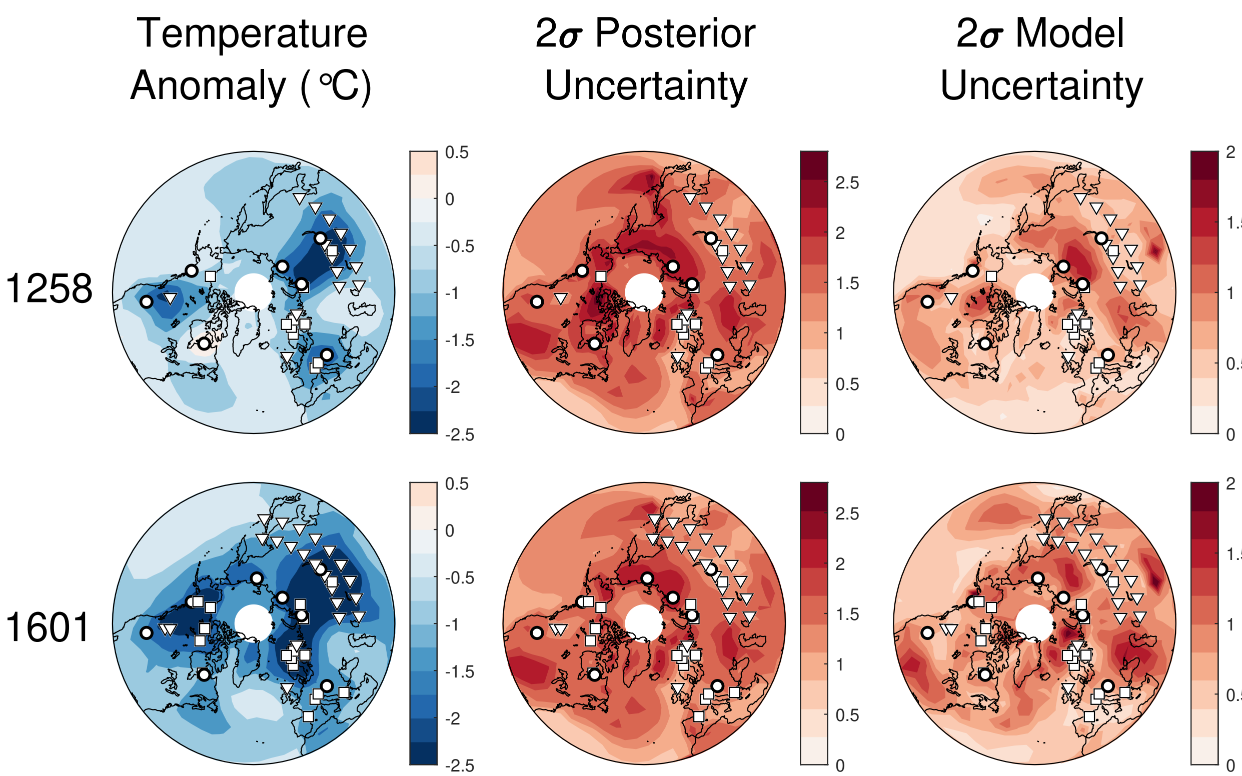


FIG. 10. Spatial characteristics in the year following volcanic eruptions in 1257 (top) and 1600 (bottom) (De Silva and Zielinski 1998; Lavigne et al. 2013) in the multi-model mean reconstruction. The left column displays temperature anomalies relative to the five preceding years in Celsius. The middle column shows the average 2σ width of the 10 posterior ensembles, and the right column shows the 2σ width of the multi-model ensemble. White markers show the proxy network for each event. Marker symbols follow the convention in Figure 1.

Supplemental Information for 'A data assimilation approach to last millennium temperature field reconstruction using a limited high-sensitivity proxy network'

Jonathan M. King

Department of Geosciences and Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona

Kevin J. Anchukaitis

School of Geography, Development, and Environment and Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona

Jessica E. Tierney

Department of Geosciences, University of Arizona, Tucson, Arizona

Gregory J. Hakim

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

Julien Emile-Geay and Feng Zhu

Department of Earth Sciences, University of Southern California, Los Angeles, California

Rob Wilson

School of Earth and Environmental Sciences, University of St Andrews, St Andrews, UK

LIST OF TABLES

TABLE S1. Pseudo-proxy localization radii and split-sample validation metrics. As in Table 2, but using climate model output as the target field.

Target	Prior	Localization Radius (km)	Correlation	RMSE (°C)	σ Ratio	Mean Bias (°C)
CESM	CESM	∞	0.73	0.18	0.76	0.02
CESM	MPI	∞	0.72	0.19	0.91	0.02
MPI	CESM	∞	0.74	0.21	0.62	0.09
MPI	MPI	∞	0.75	0.20	0.75	0.07

TABLE S2. Skill metrics for pseudo-proxy reconstructions of mean extratropical May-August time series. DA reconstructions use the realistic biased-model, noisy-proxy, time-attrition experimental design. PPR time series and target time series are calculated using only the grid cells for which RE>0 in each reconstructed time step.

Target Field	Reconstruction Method	Correlation	RMSE (°C)	σ Ratio	Mean Bias (°C)
CESM	DA, MPI Prior	0.67	0.20	0.84	-0.03
	PPR	0.68	0.25	0.96	0.03
MPI	DA, CESM Prior	0.74	0.41	0.66	0.35
	PPR	0.73	0.46	0.84	0.37

LIST OF FIGURES

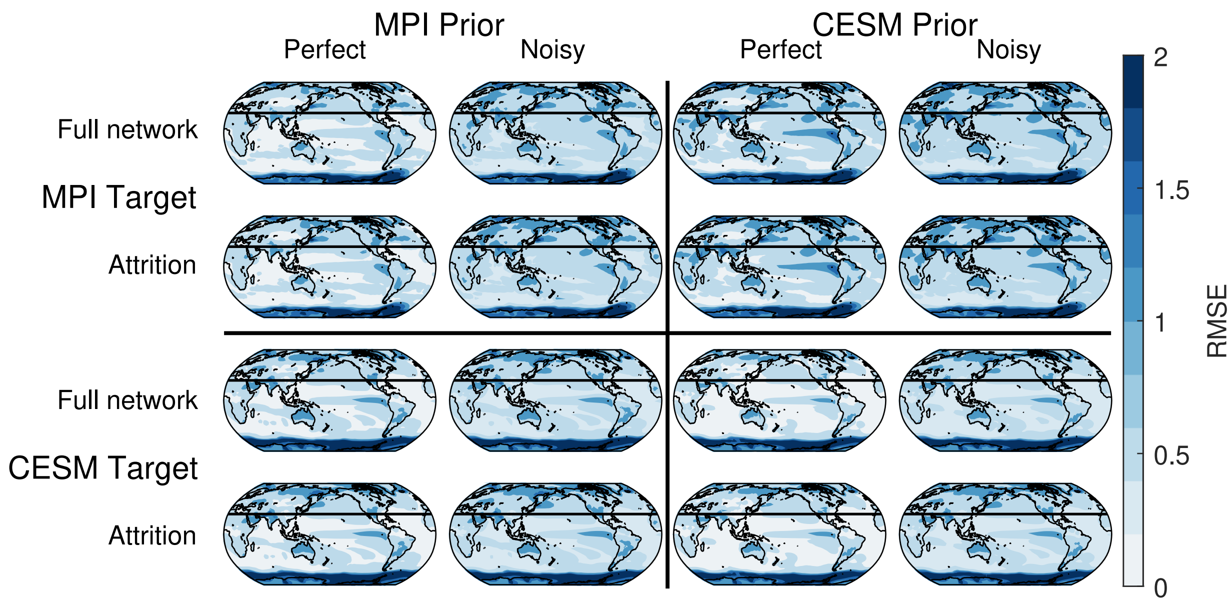


FIG. S1. As in Figure 2, but for RMSE (°C).

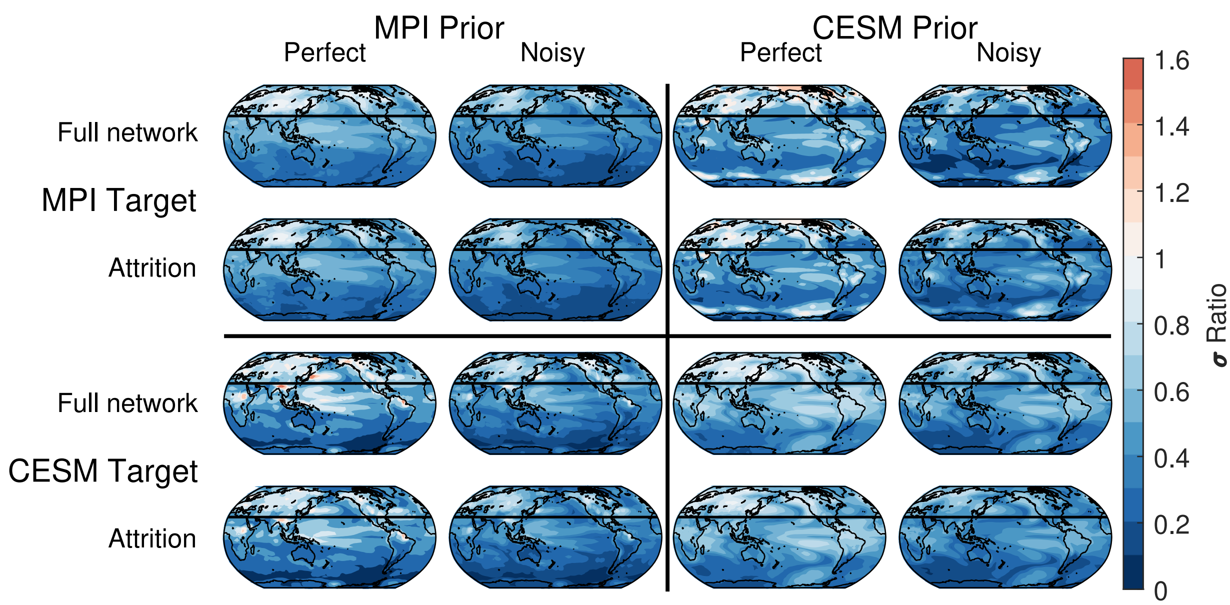


FIG. S2. As in Figure 2, but for σ ratios.

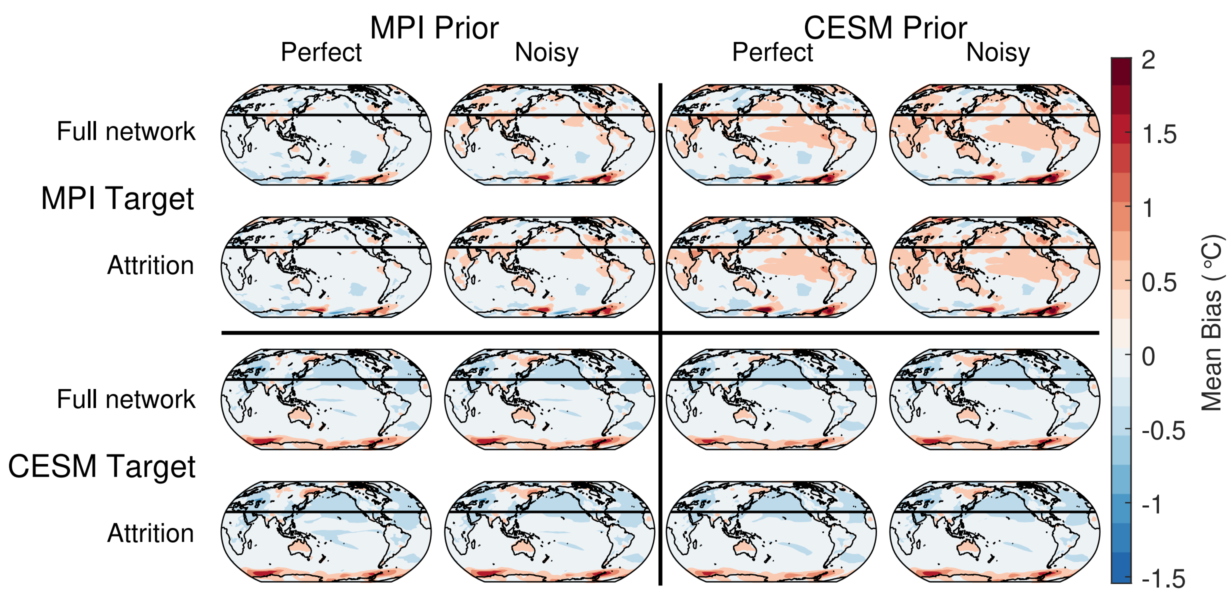


FIG. S3. As in Figure 2, but for mean biases (°C).

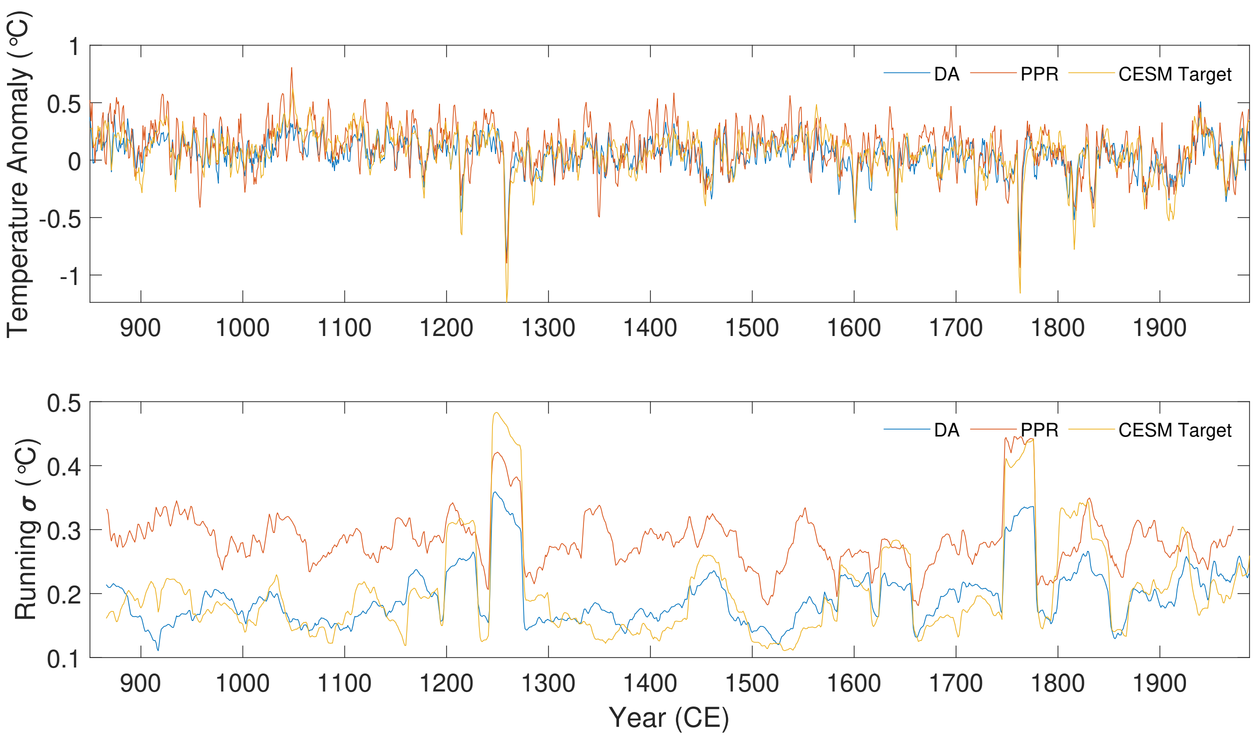


FIG. S4. Extratropical MJJA time series for the pseudo-proxy experiments with a CESM target. Reconstructed temperature anomalies are shown in Celsius (top) for the DA reconstruction (blue) and PPR reconstruction (red) along with the reconstruction target (yellow). The bottom panel displays a 31 year running standard deviation for each time series. A three year moving average has been applied to all time series.

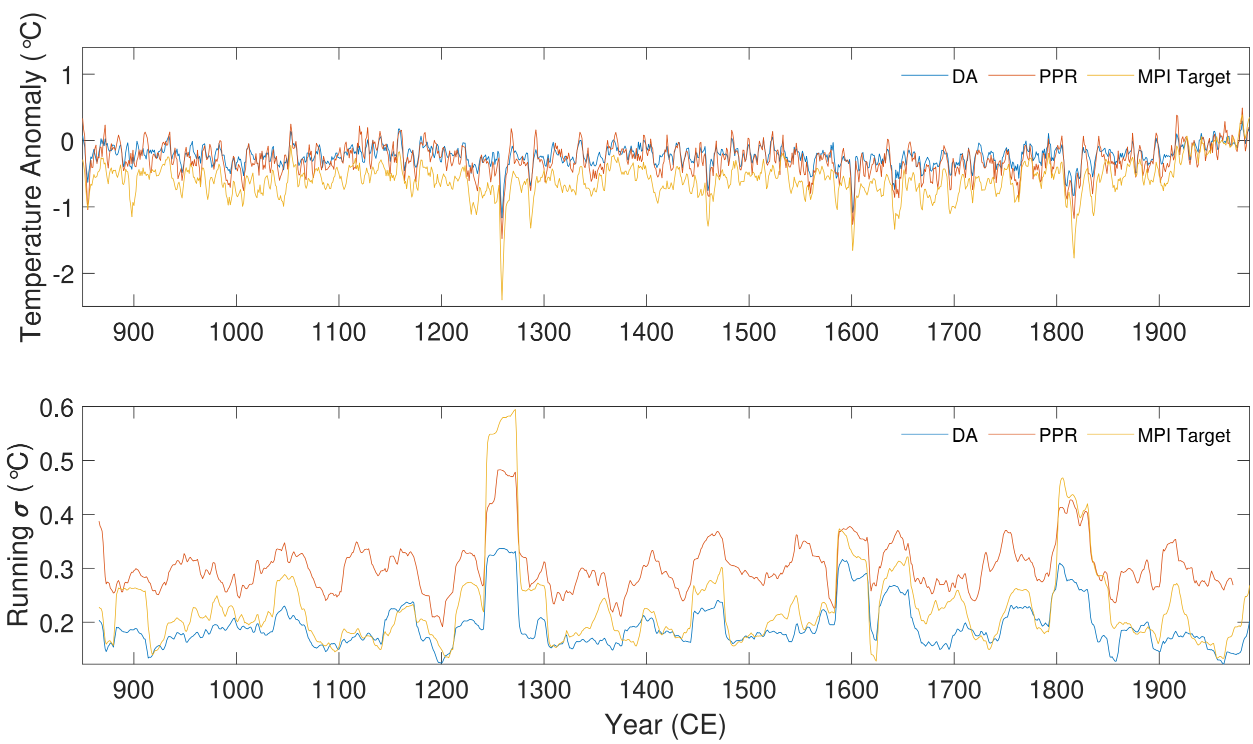


Fig. S5. As in Supplemental Figure 4, but for an MPI target.

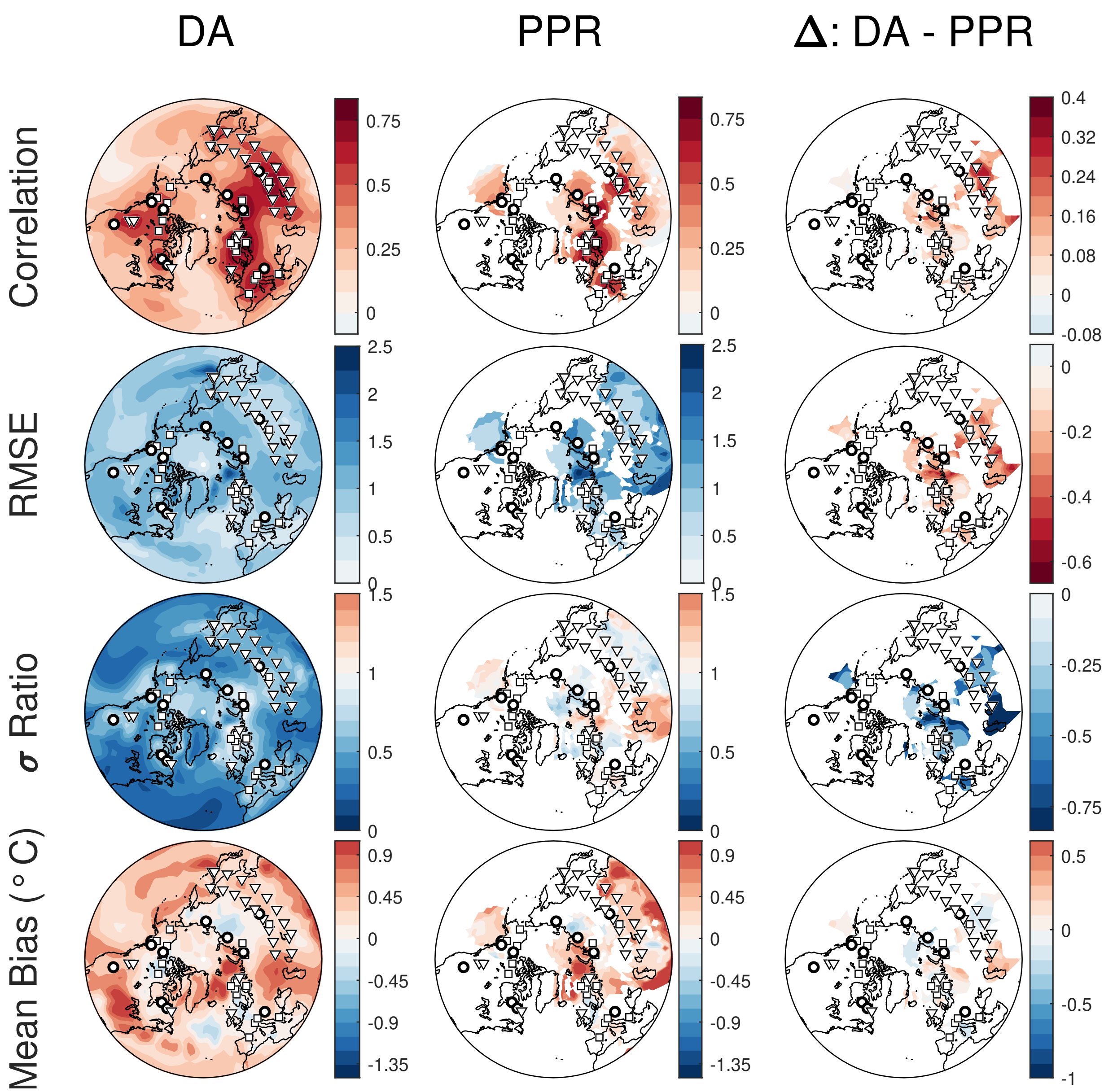


Fig. S6. As in Figure 3, but for a MPI target field. Here, the DA reconstructions use a CESM prior.

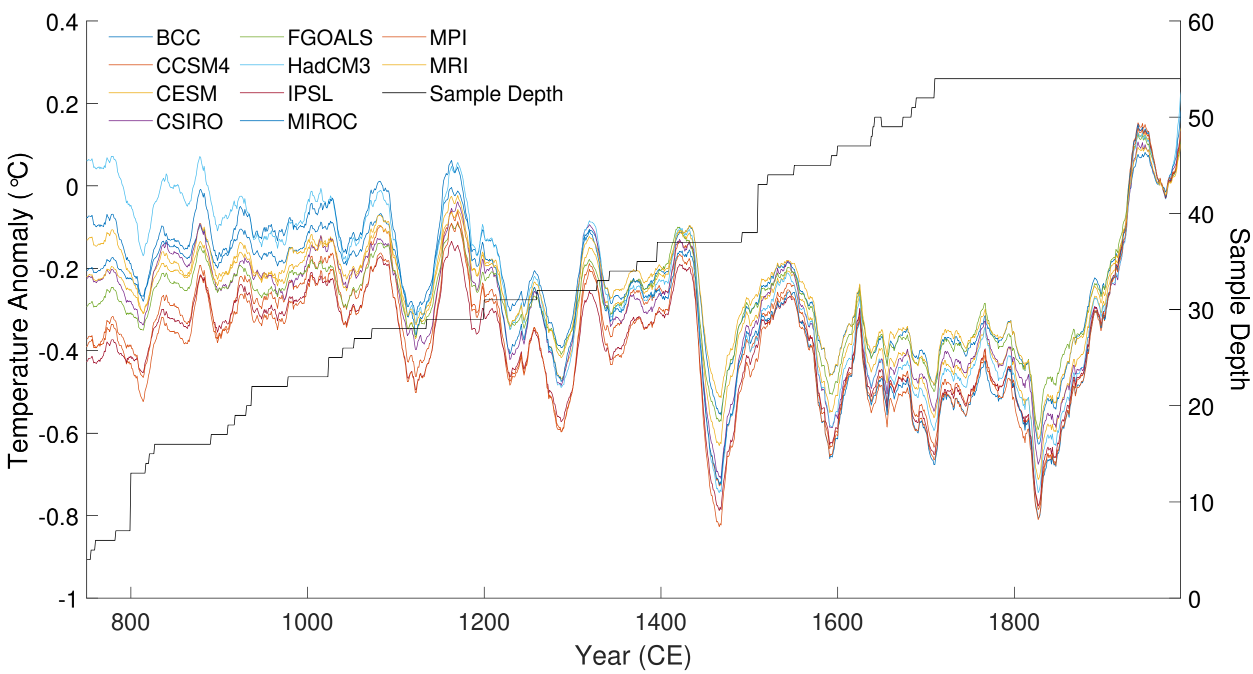


Fig. S7. Extratropical MJJA time series for the individual DA reconstructions. Each time series shows the results for a particular model prior. A 31 year moving average has been applied to each time series.