Do multi-model ensembles improve reconstruction skill in paleoclimate data assimilation?

Luke A Parsons[1], Daniel E. Amrhein[2], Sara Sanchez[3], Robert Tardif[3], M. Kathleen Brennan[3], and Gregory J. Hakim[3]

[1]Duke University
[2]NCAR
[3]University of Washington

July 10, 2023

AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

RESEARCH ARTICLE
10.1029/2020EA001467

Do Multi-Model Ensembles Improve Reconstruction Skill in Paleoclimate Data Assimilation?

Luke A. Parsons[1,2], Daniel E. Amrhein[3], Sara C. Sanchez[4,5], Robert Tardif[1], M. Kathleen Brennan[1], and Gregory J. Hakim[1]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, [2]Nicholas School of the Environment, Duke University, Durham, NC, USA, [3]National Center for Atmospheric Research, Boulder, CO, USA, [4]Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Seattle, WA, USA, [5]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA

Key Points:

- Ensembles drawn from single climate models and combinations of multiple models are used to reconstruct surface air temperature variability
- Reconstructions from multi-model ensembles produce lower error than reconstructions from single-model ensembles
- Reconstructions from multi-model ensembles show the largest decreases in error in regions with few observations such as high-latitude oceans

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. A. Parsons,
luke.parsons@duke.edu

Citation:

Parsons, L. A., Amrhein, D. E., Sanchez, S. C., Tardif, R., Brennan, M. K., & Hakim, G. J. (2021). Do multi-model ensembles improve reconstruction skill in paleoclimate data assimilation? *Earth and Space Science*, *8*, e2020EA001467. https://doi.org/10.1029/2020EA001467

Received 16 SEP 2020
Accepted 25 FEB 2021

Abstract Reconstructing past climates remains a difficult task because pre-instrumental observational networks are composed of geographically sparse and noisy paleoclimate proxy records that require statistical techniques to inform complete climate fields. Traditionally, instrumental or climate model statistical relationships are used to spread information from proxy measurements to other locations and to other climate variables. Here ensembles drawn from single climate models and from combinations of multiple climate models are used to reconstruct temperature variability over the last millennium in idealized experiments. We find that reconstructions derived from multi-model ensembles produce lower error than reconstructions from single-model ensembles when reconstructing independent model and instrumental data. Specifically, we find the largest decreases in error over regions far from proxy locations that are often associated with large uncertainties in model physics, such as mid- and high-latitude ocean and sea-ice regions. Furthermore, we find that multi-model ensemble reconstructions outperform single-model reconstructions that use covariance localization. We propose that multi-model ensembles could be used to improve paleoclimate reconstructions in time periods beyond the last millennium and for climate variables other than air temperature, such as drought metrics or sea ice variables.

Plain Language Summary Understanding past climate variability is important for contextualizing climate change as well as for testing the ability of climate models to simulate the climate system before global warming. However, reconstructing past climate variability remains a complex task because pre-instrumental paleoclimate proxy records, such as tree rings, corals, ice cores, and sediment cores, are geographically sparse and are not perfect recorders of climate information. Exactly how to extrapolate information from paleoclimate proxies to other locations and climate variables remains an outstanding issue. Traditionally, information about how one location varies with another location or variable (covariance) is derived from one climate model or from one instrumental data source. Here we find that reconstructions using covariance estimated from combinations of multiple climate models produce less error than reconstructions that use just one climate model.

1. Introduction

Large uncertainties in pre-instrumental climate reconstructions are expected to arise from sparse geographic observational coverage and noise in paleoclimate proxy (e.g., tree ring, coral, ice core, lake/ocean sediment) observations. Traditionally, the covariance information (how one location varies with other locations or variables) from one global climate model or instrumental-based data source is used to infer climate fields from proxy information in last millennium climate reconstructions (e.g., Barriopedro et al., 2011; Cook et al., 2004; 2010; Evans et al., 2002; Hakim et al., 2016; Luterbacher et al., 2004; Mann et al., 1998, 1999, 2009; Rutherford et al., 2005; Singh et al., 2018; Smerdon et al., 2016; Steiger et al., 2014, 2018; Tardif et al., 2019). Climate field reconstruction methods can generally be placed into three categories: (1) a regression-based approach that uses a multivariate regression of climate variables, often from the instrumental record, onto proxy records, (2) Bayesian Hierarchical Models that invert the proxy-climate relation to obtain a probabilistic analysis of the predicted climate field, and (3) data assimilation methods that update model-simulated climate states with novel information from observation data and spatially spread that information via the relevant climate variables (see Hakim et al., 2016 for a more in-depth description and comparison among

methods). Here we focus on an offline (no temporal cycling) data assimilation framework (e.g., Hakim et al., 2016; Oke et al., 2002; Steiger et al., 2014; 2018), in which large ensemble draws of climate "states" from climate model simulations are used to infer covariances. Local and regional covariances estimated from different climate models are often quite similar among models, but spatially remote covariances can vary substantially (see Figure S1 in Parsons & Hakim, 2019; see also Weare, 2013). Specifically, a recent study focused on errors in paleoclimate reconstructions due solely to the choice of climate model used in these reconstructions finds that up to 50% of error can be attributed to uncertainties in climate model physics (Amrhein et al., 2020).

To limit the effects of observations influencing remote locations over large distances, the covariance estimate is often "localized" (e.g., Liu et al., 2017; Steiger et al., 2014; Tardif et al., 2019) using a spatially isotropic function (Gaspari & Cohn, 1999). Although restricting the impact of proxy information equally in all directions has been shown to improve reconstruction skill (e.g., Bhend et al., 2012), anisotropic approaches are preferred since they can retain useful information. Here we test the hypothesis that estimating covariances from samples derived from multiple models (hereafter multi-model ensembles, or MMEs) can reduce reconstruction error in offline ensemble Kalman filter reconstructions. We also test if estimating covariances from an MME can reduce reconstruction error as effectively as isotropic localization in offline ensemble Kalman filter reconstructions.

The use of MMEs has a long history in weather and climate forecasting. In an early study on accounting for uncertainty in model physics, Houtekamer et al. (1996) show that perturbing the physics parameterizations of a single model does not produce sufficient spread in the resulting ensemble weather forecasts, suggesting that MMEs might be necessary to capture full uncertainty. Multi-model ensembles for seasonal forecasts (e.g., Doblas-Reyes et al., 2000; Krishnamurti et al., 1999) have been shown to improve forecasts over individual model ensembles in both the skill of the ensemble mean and the reliability of the ensemble statistics (e.g., Hagedorn et al., 2005). Even when there is a range of skill in the contributing ensemble members, the MME can be more reliable than the best member if the less skillful members contribute to ensemble variance without sacrificing skill in the ensemble mean (Weigel et al., 2008). Weighting the ensemble members can further improve forecast skill (e.g., Krishnamurti et al., 1999; Robertson et al., 2004), although accounting for the lack of independence among members in the weighting scheme remains an unresolved problem (e.g., Bishop & Abramowitz, 2013; Christiansen, 2020; Stephenson et al., 2005).

We conduct a series of idealized experiments using model ensembles and pseudo-observations drawn from Coupled Modeling Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012) models or from instrumental-era analyses at paleoclimate proxy locations of the Common Era (CE; e.g., Mann & Rutherford, 2002; Smerdon et al., 2016; Steiger et al., 2014; see Smerdon, 2012 for a review of pseudoproxy experiments) to examine: (1) If using MMEs improves reconstruction skill as compared to single-model ensembles when reconstructing a withheld climate model or instrumental surface air temperature variability, (2) If using a MME provides comparable or improved results as compared to the use of isotropic localization (e.g., Bhend et al., 2012; Gaspari & Cohn, 1999) in single-model ensembles. Finally, we explore why MMEs reduce reconstruction error and suggest this improved reconstruction skill may be related to the greater degrees of spatial freedom (Bretherton et al., 1999) and shorter spatial covariance length scale from observation locations in MMEs as compared to single-model ensembles.

2. Data and Methods

2.1. CMIP5 past1000 Climate Model Data

In the construction of single and MMEs, we use the 2-m air temperature output from 10 CMIP5 past1000 ("Last Millennium") simulations spanning 850–1850 Common Era (Schmidt et al., 2011; Taylor et al., 2012): BCC-CSM1-1, NCAR CCSM4, CSIRO-Mk3L-1-2, FGOALS-s2, GISS-E2-R, HadCM3, IPSL-CM5A-LR, MIROC-ESM, MPI-ESM-P, and MRI-CGCM3. The atmospheric variables from these models are provided at various spatial resolutions; to combine and analyze data from different models, we first regrid the data to a common latitude-longitude grid from HadCM3 (2.5° × 3.75°; Collins et al., 2001). We then remove the linear trend from each grid point over the last millennium time period (850–1849) to minimize model drift from spin up (e.g., Gupta et al., 2013). Specifically, the GISS-E2-R and MIROC-ESM past1000 simulations

are often excluded from analyses of the last millennium due to model drift (Bothe et al., 2013; Sueyoshi et al., 2013). Here we choose to include these model simulations to test the ability of single and MMEs to reconstruct features of the climate system that other models may not share. We treat the contributing ensemble members with equal weighting (Section 2.2), and leave unequal weighting strategies for future research. In all reconstructions, annual-mean, 2-m air temperature is used, and draws of years are selected after the linear trend has been removed from the full past1000 simulation.

2.2. Construction of Single and Multi-Model Ensembles

To construct the 9-model multi-model ensemble (MME) prior, we randomly draw 100 years from each re-gridded CMIP5 past1000 model after removing the linear 2-m air temperature trend from each grid point (Section 2.1). Although there are 10 CMIP5 past1000 models, here we withhold one model to conduct a "jackknife" experiment (Section 2.3) in which we use the covariance from the MME (900 members in total) to reconstruct the one withheld "target" model. We have tested the sensitivity of our results by randomly drawing a new selection of 100 years from each model to construct the MME, and we find no noticeable differences in our main results. Each single-model ensemble consists of 900 years randomly drawn from a regridded and detrended CMIP5 past1000 model simulation.

We also use single-model ensembles and MMEs to reconstruct instrumental-era temperature variability (Section 2.6). In this experiment, we combine a draw of 100 years from each of the 10 CMIP5 past1000 simulations into one 1,000 member MME. We compare reconstruction error in this 1,000 member MME to error from 10 reconstructions, each using 1,000 ensemble members (e.g., all years 850–1849) from each CMIP5 past1000 simulation.

2.3. Idealized Jackknife Experiments

Following Amrhein et al. (2020), here we assess the impact of using MMEs in an offline ensemble Kalman filter in reconstructions of annual-mean, 2-m air temperature. We conduct a series of idealized jackknife experiments in which we construct prior ensembles (Section 2.2) then perform reconstructions in which the covariance from the different 900-member single or MMEs are used to reconstruct 100 years of variability from the withheld model. As a "control" reference, we also use a 900-member ensemble from the withheld target model to reconstruct itself. We have tested the sensitivity of our results to the choice of target years used in the withheld model by re-conducting the experiment 10 times using different random draws of 100 years from the withheld target model and find that our main results remain unchanged. Although the relative performance of the MME as compared to the single-model ensembles remains unchanged, the absolute error can shift depending on the choice of target years. In all experiments, except those discussed in Section 2.5, we use localization covariance length scales (e.g., Gaspari & Cohn, 1999) of infinity ("no localization").

For each jackknife reconstruction, we assimilate a sparse network of "pseudo-proxy" observations derived from annual-mean, 2-m air temperature output from the withheld target model (e.g., Amrhein et al., 2020; Smerdon et al., 2010; Steiger et al., 2014; von Storch et al., 2004). Observations are sampled from the regridded climate model grid point (Section 2.2) closest to the locations corresponding to the proxy network of sub-annual and annually resolved paleoclimate records from the Past Global Changes (PAGES) Consortium collection (Emile-Geay et al., 2017). We use annual-mean air temperature at all proxy locations, even if the actual proxies are sub-annually resolved at some of these locations. Figure 1 and Figure S1 show the geographic locations of these 542 sub-annual and annually resolved proxy records, which include tree rings, ice cores, corals, and lake sediments. Several proxy locations are situated within one latitude-longitude grid point at the resolution of analysis (2.5° × 3.75°, Section 2.1), so although there are 542 proxies in the PAGES network, there are only 220 unique grid points from which observations are drawn. We then add zero-mean, Gaussian distributed, random error to each pseudo-observation from the withheld model to create a signal-to-noise ratio of 0.4 (by standard deviation), which is a typical signal-to-noise ratio in many paleoclimate proxy records (Smerdon et al., 2016). The number and location of available PAGES proxies is temporally variable, but here we assimilate all 542 PAGES sub-annual and annually resolved proxy network locations (220 unique grid points) for each year reconstructed. In one set of experiments, we test the sensitivity of our

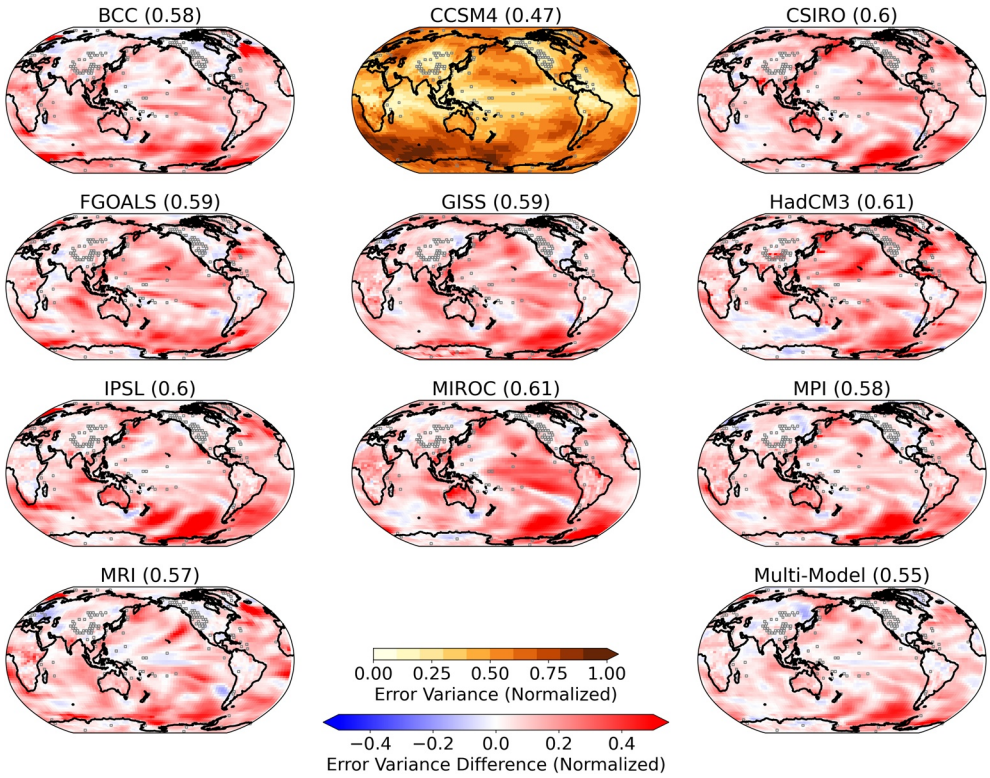

Figure 1. Local, normalized error variance differences for each single-model ensemble and the multi-model ensemble using observations drawn from CCSM4 at 542 PAGES Consortium paleoclimate proxy locations (gray squares on maps). Normalized error variance differences (blue-white-red divergent colormap) are relative to error from reconstructions of CCSM4 when both observations and the single-model ensemble are drawn from the CCSM4 target (center top, yellow-orange-brown color map). Numbers in parentheses show the global-mean, area-weighted, normalized error variance. Local error is normalized to local 2-m air temperature variance in CCSM4 (Figure S2). Note that given the sampling locations and "noise" added to the pseudo-observations, some amount of reconstruction error is inevitable, even in the case in which observations and the single-model ensemble are both drawn from the CCSM4 target (center top).

results to the location of these 542 paleoclimate proxy locations by limiting the pseudo-observation network to the 127 proxy locations (57 unique grid points) that have annual resolution data for the entire 1500–2000 CE time period, which eliminates most observation locations in the tropics (bottom map in Figure S1).

2.4. Data Assimilation Methodology

Data assimilation produces paleoclimate reconstructions originating from models and measurements; this methodology offers useful machinery to quantify uncertainties in paleoclimate reconstructions arising from uncertain climate physics. Here we use a least squares approach of fitting observations to ensembles of model output without a forecasting step, referred to as the offline ensemble Kalman filter (e.g., Oke et al., 2007).

We represent paleoclimate pseudo-proxies $\mathbf{y}_j$ as linear functions of a climate state anomaly,

$$\mathbf{y}_j = \mathbf{H}\mathbf{x}_j^t + \mathbf{n} \tag{1}$$

where $\mathbf{x}_j^t$ denotes the zero-mean, annual-average surface air temperature in the jth year of a withheld target simulation (denoted with superscript t). Here $\mathbf{H}$ is a linear operator that selects data from the regridded climate model (Section 2.2) at the locations closest to the PAGES network proxy locations, and the vector $\mathbf{n}$ is a vector of Gaussian noise (Section 2.3). We do not consider offsets (time mean biases) between models and data, which can also contribute to errors, and we assume observational errors are uncorrelated in space and time.

This work evaluates error variances computed by comparing ensemble mean reconstructed 2-m air temperature $\underline{\mathbf{X}}_j^a$ (the "best guess" solution afforded by offline ensemble Kalman filter (EnKF)) and the target values drawn from a climate model or instrumental data set. Reconstructions of target fields are informed by a set of "prior" zero-mean surface air temperature fields $\mathbf{x}^p$ drawn from a different model simulation. These individual prior ensemble members (indexed by i) are updated to yield an ensemble of posterior estimates as

$$\mathbf{x}_{ij}^a = \mathbf{x}_i^p + \mathbf{K}\left(\mathbf{y}_j - \mathbf{H}\mathbf{x}_i^p\right) \tag{2}$$

where the gain matrix $\mathbf{K}$ is

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T\left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{3}$$

the diagonal matrix $\mathbf{R}$ contains the proxy error variances, and the prior covariance, $\mathbf{B}$, is computed as

$$\mathbf{B} = \left(N_{ens} - 1\right)^{-1}\mathbf{X}^p\mathbf{X}^{pT}. \tag{4}$$

where each column of $\mathbf{X}^p$ is an annually averaged 2-m air temperature field drawn from the prior model and N_{ens} is the number of prior ensemble members. Since we have defined the prior to have zero mean and H does not contribute a mean value, the ensemble mean solution (denoted by an underbar) takes the simple form

$$\underline{\mathbf{x}}_j^a = \mathbf{K}\mathbf{y}_j. \tag{5}$$

2.5. Multi-Model Ensembles Compared to Single-Model Ensemble Localization

Restricting the impact of proxy information equally in all directions ("isotropic localization," see Section 1) has been shown to improve reconstruction skill (e.g., Bhend et al., 2012), but anisotropic approaches are preferred because they can retain meaningful climate information that may not be spatially uniform. Covariance relationships in a MME are attenuated whenever the underlying models disagree about their sign or amplitude, leading to spatially anisotropic changes from the covariance relationships in any one model. We assess how MME reconstruction error compares to solutions derived from use of isotropic localization by conducting a series of experiments with covariance localization length scales of infinity (no localization), 25,000, 10,000, 5,000, and 2,000 km in single and MMEs.

2.6. Instrumental Era Reconstructions

The ability of models to reconstruct one another is not a guarantee of performance because models can share common physical biases (e.g., Tebaldi & Knutti, 2007). As such, we also compare the ability of single and MMEs to reconstruct instrumental-era temperature variability. We use the same methodology described in Section 2.3, but we draw 2-m air temperature observations from 20th Century Reanalysis version 3 (20CRv3; Slivinski et al., 2019) (100 years: 1916–2015) in place of a withheld target CMIP5 past1000 simulation. The 20CRv3 has the advantage of a relatively long-duration output and spatially complete coverage, but itself uses a model in the data assimilation process, which could hypothetically be subject to some of the same biases as the CMIP5 models. Motivated by this concern, we perform another test of MME reconstruction skill by reconstructing Hadley Centre Climate Research Unit Temperature version 4.6 (HadCRUTv46; Morice et al., 2012). The HadCRUTv46 data set is created using temperature observations on a regular latitude/longitude grid, without using a model or other interpolation method to spread information from observations to data-poor regions (Morice et al., 2012). For the experiment using HadCRUTv46, we draw observations (60 years: 1960–2019) from HadCRUTv46 grid points nearest to the paleoclimate proxy locations. Even in the 1960–2019 time period, the HadCRUTv46 data set has incomplete spatial coverage, so we compare single-model and MME reconstruction skill for reconstructions of both 20CRv3 and HadCRUTv46. Similar to our treatment of the CMIP5 past1000 data, we first regrid 20CRv3 and HadCRUTv46 data to a common

latitude-longitude grid from HadCM3 (2.5° × 3.75°) then remove the linear trend from each grid point before use in climate reconstructions. In these experiments, we use localization covariance length scales (e.g., Gaspari & Cohn, 1999) of infinity ("no localization").

2.7. Reconstruction Error Quantification

We compare the output from ensemble-mean single and MME experiments to the original 100 target years drawn from the withheld CMIP5 model (Section 2.3) or the instrumental-based observations (100 years for 20CRv3, 60 years for HadCRUTv46, Section 2.6). For a given reconstructed scalar quantity x^a_j (e.g., 2-m air temperature in a model grid box, or averaged over a region), error is computed as

$$\eta^2 = \frac{1}{M_{\text{ens}} - 1} \sum_{j=1}^{M_{\text{ens}}} \left(x^a{}_j - x^t{}_j \right)^2 \tag{6}$$

where x^t_j is the corresponding scalar quantity from the withheld target model and M_{ens} is the number of target years. M_{ens} is 100 years for the withheld target models and 20CRv3 and 60 years for HadCRUTv46 targets.

We show maps of local error variance and plots of global-mean, local error variance to compare results from single and MMEs. Error variance is sensitive to the variance of the original data, so we also show normalized error variance (Figure 1); error variance is normalized to the local 2-m air temperature variance in the withheld target CMIP5 model data (Figure S2).

We also create regional-mean time series for the PAGES continental regions (Ahmed et al., 2013) to compare in the withheld model and reconstructions. Specifically, we highlight the importance of remote covariance by comparing reconstruction error from the regional-mean time series from the PAGES Consortium North America region, a geographic location with numerous paleoclimate records of the Common Era, to the regional mean time series from the Niño3.4 region (5°S–5°N, 190°E–240°E), and the South Pacific region (60°S–40°S, 180°E–280°E), two locations with few local, annually resolved paleoclimate proxy observations.

2.8. Examining Impact of Combinations of Single-Model Ensembles

We illustrate the influence of combining information from multiple models by analyzing the effective number of spatial degrees of freedom (ESDOF; Bretheron et al., 1999). We also compute the centroid distance of isotropic radial covariance functions at all 542 PAGES Consortium proxy locations. To do this, we calculate the global map of covariances for each proxy location, then the area-weighted average a_i in 100 bins $(a_1, a_2, \ldots a_{100})$ as a function of spherical distance d_i. The centroid distance (here referred to as "length scale") is then computed as

$$_c = \left(\sum_{i=1}^{100} a_i d_i \right) \Big/ \left(\sum_{i=1}^{100} a_i \right) \tag{7}$$

For these calculations, we use 1000 years from each of the past1000 models for the single-model ensembles. For the MME, we use 1000 years, which includes 100 years from each of the 10 CMIP5 past1000 models. We test the sensitivity of the MME ESDOF to the selection of years drawn from the CMIP5 past1000 simulations by randomly drawing new selections of 100 years from each model to construct the MME 100 different times and show the spread in ESDOF results for the MME ESDOF as a boxplot.

3. Results

3.1. Single-Model Ensembles Versus Multi-Model Ensembles Using No Localization

Here we are interested in quantifying the reconstruction error in paleoclimate data assimilation associated with uncertainties in climate model physics and in determining if these errors can be reduced by combining information from multiple climate models into MMEs. Given the typical signal-to-noise ratio in proxy re-

cords and the sparse geographic coverage of sub-annual and annually resolved paleoclimate proxy locations from the PAGES (Emile-Geay et al., 2017) database (Figure S1), some amount of error in reconstructions is inevitable (e.g., Amrhein et al., 2020; Smerdon et al., 2016). For example, regions of relatively large error in reconstructions of CCSM4 (both when CCSM4 itself is used in the single-model ensemble and when other CMIP5 past1000 models are used) are found over the Southern Ocean, eastern tropical Pacific, and Arctic (Figure S3). The error metric is sensitive to local variability, so we also normalize local error variance to the variability in the withheld CCSM4 model (Figure S2 shows maps of 2-m air temperature variance in CMIP5 past1000 models), yielding a noise-to-signal ratio in reconstructed values (here referred to as "normalized error variance"). The top center panel in Figure 1 shows the normalized error variance produced when the CCSM4 "target" is used to reconstruct itself. The other panels in Figure 1 show normalized error variance anomalies from this CCSM4 reference. Normalized error variance is highest over mid-latitude ocean regions, particularly in the Southern Ocean. The MME shows lower normalized error variance compared to the single-model ensembles when reconstructing CCSM4 across most of the globe, particularly over much of the mid-latitude Southern Ocean.

The MME generally produces lower reconstruction error when reconstructing CCSM4 (Figure 1, Figure S3), but CCSM4 is only one of many model representations of the Earth system. Therefore, we test the robustness of these results in a series of jackknife experiments, in which we compare reconstruction error from single and MMEs when reconstructing 100 randomly drawn years from each of the 10 CMIP5 past1000 simulations (Section 2.3). Zonal-mean normalized error variance relative to the MME error is shown in Figure 2 for each of these reconstructions. Almost all red lines in Figure 2 show positive values, which indicate that the single-model ensembles produce larger zonal mean error than the MME. Specifically, for several target models (CSIRO, MPI), there is largest improvement over the mid- and high-latitude Southern Ocean. In these regions there are few nearby observations and models tend to disagree on locations of frontal regions (e.g., Beadling et al., 2019; Holmes et al., 2019). For reconstructions of other target models, the lower error in the MME reconstructions is less focused over the high-latitude southern hemisphere (Figure 2).

We summarize this single and multi-model comparison using global-mean error variance values (Figure 3). Although the overall skill of reconstructions is dependent on the model target, the MMEs (black dots in Figure 3) almost always produce lower global-mean error than the single-model ensembles (colored dots in Figure 3). We test the sensitivity of the relative reconstruction error by reconstructing each of the target models 10 different times, each using a new random draw of years from the target model. We find that the MME consistently outperforms all of the single-model ensembles (Table S1 shows mean ranking of reconstruction skill for all single and multi-model ensembles). On average, the MME produces the lowest error in these experiments, followed by MPI, IPSL, and MRI. HadCM3, CSIRO, and BCC on average show the least reconstruction skill (Table S1). Limited cases and regions where individual models show lower error variance could be due to the particular observational network, or to finite ensemble sizes.

Geographic variations in reconstruction skill can provide more useful information than global-mean error variance (Figure S4). The MME typically produces lower error in the tropical Pacific (with the exception of reconstructions of MIROC, GISS) and large portions of the Southern Ocean. However, the MME does not outperform the median single prior in the Weddell and Scotia Seas (reconstructions of BCC, CSIRO, GISS, HadCM3) and near the Amundsen and Ross Seas (reconstructions of MIROC, MRI).

The MME shows decreases in error over many ocean regions (Figure S4) and in the global mean (Figure 3), but continental regions are often targeted in paleoclimate reconstructions (e.g., Bothe et al., 2015). Therefore, in Figure S5 we also show reconstruction error of the regional mean time series over the PAGES Consortium (Ahmed et al., 2013) continental regions not shown in the main text (Arctic, Europe, Asia, South America, Australasia, and Antarctica). Given that most sub-annual and annual paleoclimate proxy records are located on or near continents in these regions, we expect less consistent improvement in the MME results as compared to single-model ensembles because the influence of the prior covariances in these regions is lower.

To further highlight this result, we contrast reconstruction errors in two types of regions: regions with few or no local paleoclimate observations (e.g., Niño3.4, South Pacific) and a region with a relatively large number of paleoclimate observations (PAGES North America region; paleoclimate locations shown in Figure S1).

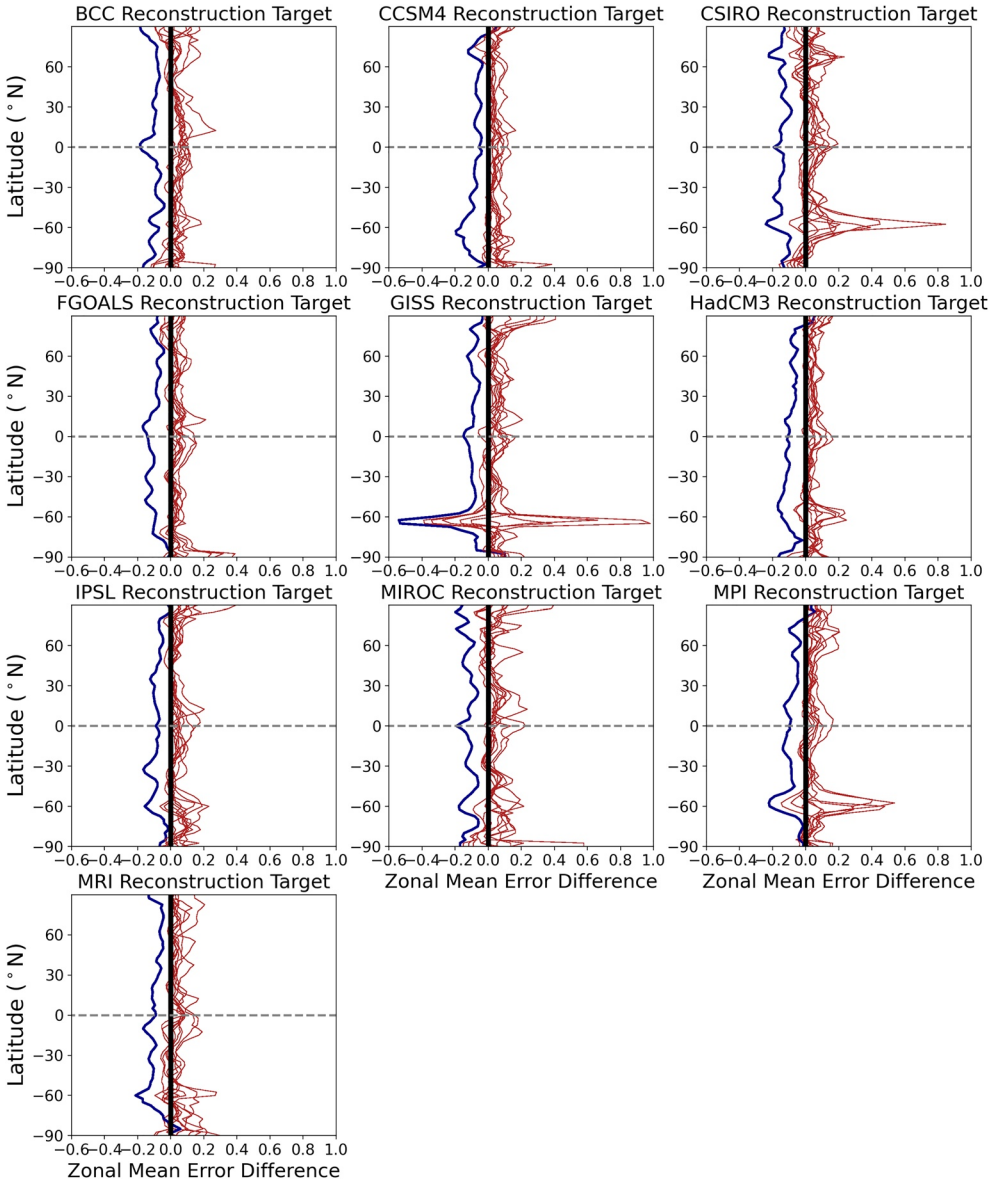

Figure 2. Zonal-mean, normalized error variance anomalies for reconstructions of each CMIP5 past1000 model. Reconstruction errors for single-model ensembles are shown as anomalies relative to the multi-model ensemble (MME) reconstruction errors. The MME zonal mean errors are shown as a vertical black line at zero on the *x*-axis, with both absolute error and error relative to the MME diminishing to the left on the x-axis. The blue line shows results for the single-model ensemble drawn from the same model used to generate observations (the "target" model), and the red lines show results for the other single-model ensembles when they are used to reconstruct this target model. The name of the target model from which observations are drawn is listed above each plot. Local error variance is normalized to the local variance in the model from which observations are drawn (Figure S2) before anomalies and zonal means are calculated. Horizontal dashed line marks the equator. Here positive values indicate larger errors than the MME and negative values indicate smaller errors than the MME.

The MME shows no consistent improvement over single-model ensembles for the PAGES North America continental region. By contrast, use of a MME usually results in lower reconstruction error for the Niño3.4 region (Figure 4). To summarize the information in Figure 4, we rank the normalized error variance from 1 to 10, with 1 indicating lowest error variance after excluding the withheld CMIP5 past1000 target model that is used to reconstruct itself. For the PAGES North America continental region, the MME mean rank is 3.8, but this ranking can be quite variable (range of 1–9), and FGOALS shows an average improvement over the MME. For the Niño3.4 region, the MME shows the lowest average error as compared to the single-model

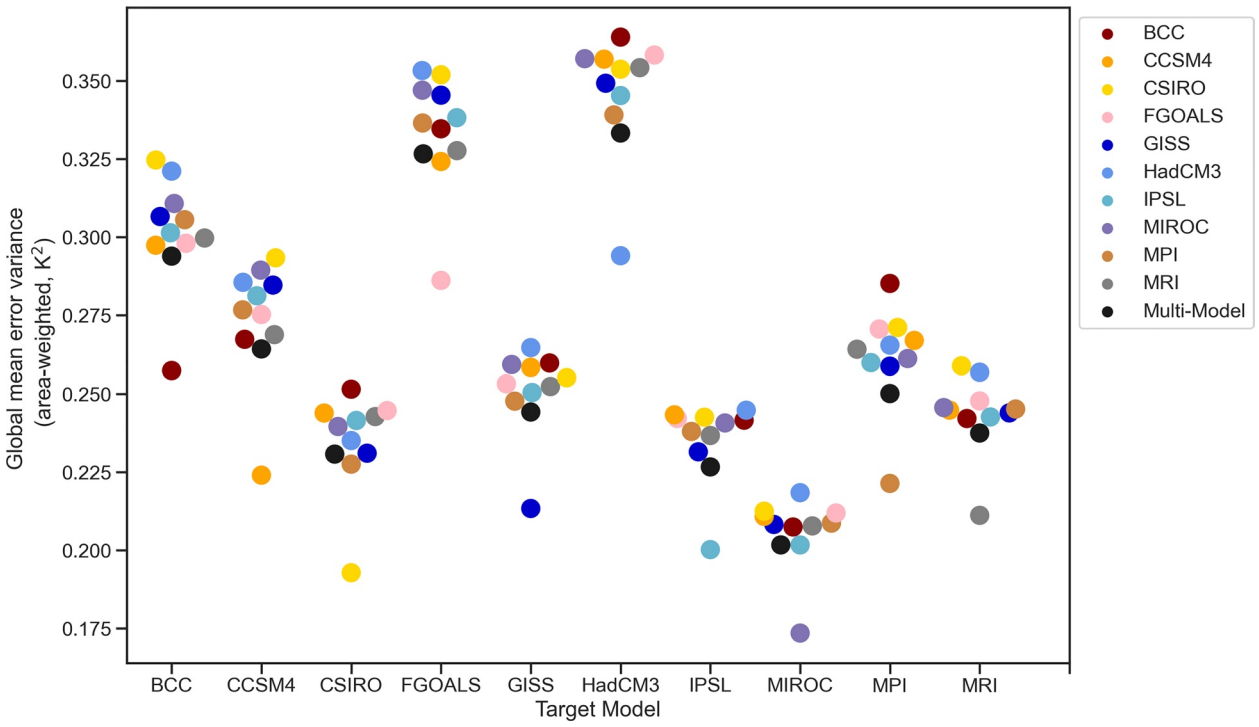

Figure 3. Global-mean, area-weighted local error variance for reconstructions of each CMIP5 past1000 model target using observations drawn from 542 PAGES Consortium paleoclimate proxy locations. Black dot marks error variance of multi-model ensemble and colored dots show single-model ensemble reconstruction error. Lowest dot in each column shows reconstruction error variance for the target model from which observations have been drawn.

ensemble reconstructions (MME mean rank of 2.8, range of 1–6), but the MME shows no improvement over the median single-model ensembles for reconstructions of FGOALS and MIROC. The MME also shows the most consistent lowest reconstruction error for the South Pacific, another ocean region with no local annual paleoclimate observation locations (MME mean rank of 3.5, not shown).

3.2. Multi-Model Ensemble Versus Single-Model Ensemble Isotropic Localization

Isotropic localization has been used to limit geographically distant covariance relationships when using single model finite ensemble sizes in paleoclimate data assimilation (e.g., Tardif et al., 2019). However, spatial covariance does not necessarily vary evenly as a function of geographic distance across the globe, so isotropic localization remains an imperfect method for filtering covariance relationships that are geographically distant from proxy locations. We test if use of the MME can more effectively reduce reconstruction error than isotropic covariance localization by reconstructing a single target model using single and MMEs that do not use localization and that use localization with cut-off distances of 25,000, 10,000, 5,000, and 2,000 km.

The upper panel in Figure 5 shows the global-mean error variance from these experiments for reconstructions of CCSM4. The non-localized MME error variance tends to show lower error than the localized single-model ensembles. Furthermore, localized MMEs show minimal/no improvement over non-localized MMEs. Therefore, isotropic localization does not improve MME reconstruction skill, but it does improve single-model ensemble reconstruction skill. We test if these results are robust for reconstructions of MPI-ESM (a CMIP5 model with different mean state and variability than the CCSM4) and find similar results (lower panel Figure 5).

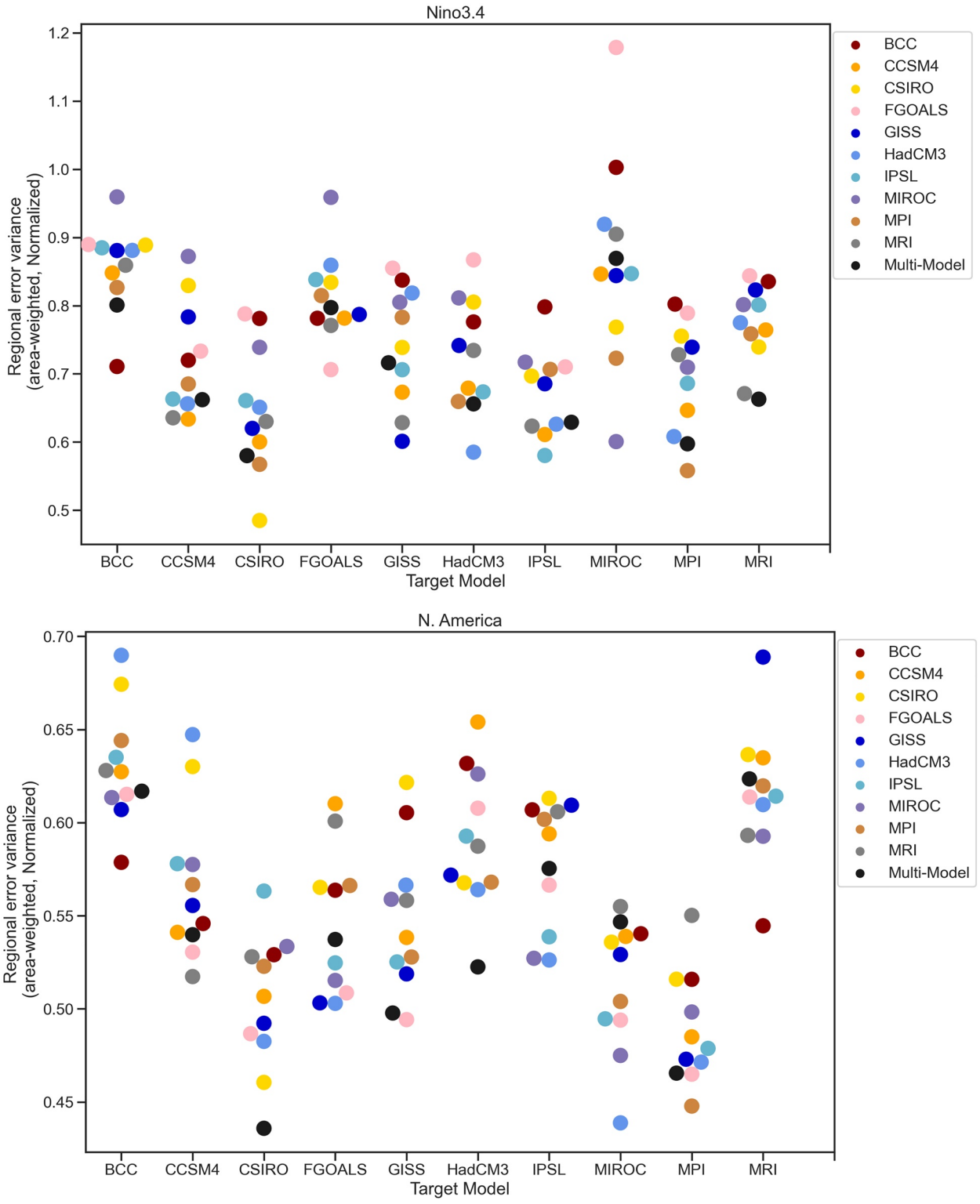

Figure 4. Normalized error variance for regional-mean time series for the Niño3.4 region (top) and for the PAGES Consortium North America region (bottom). Black dot marks error variance of multi-model ensemble and colored dots show single-model ensemble prior reconstructions.

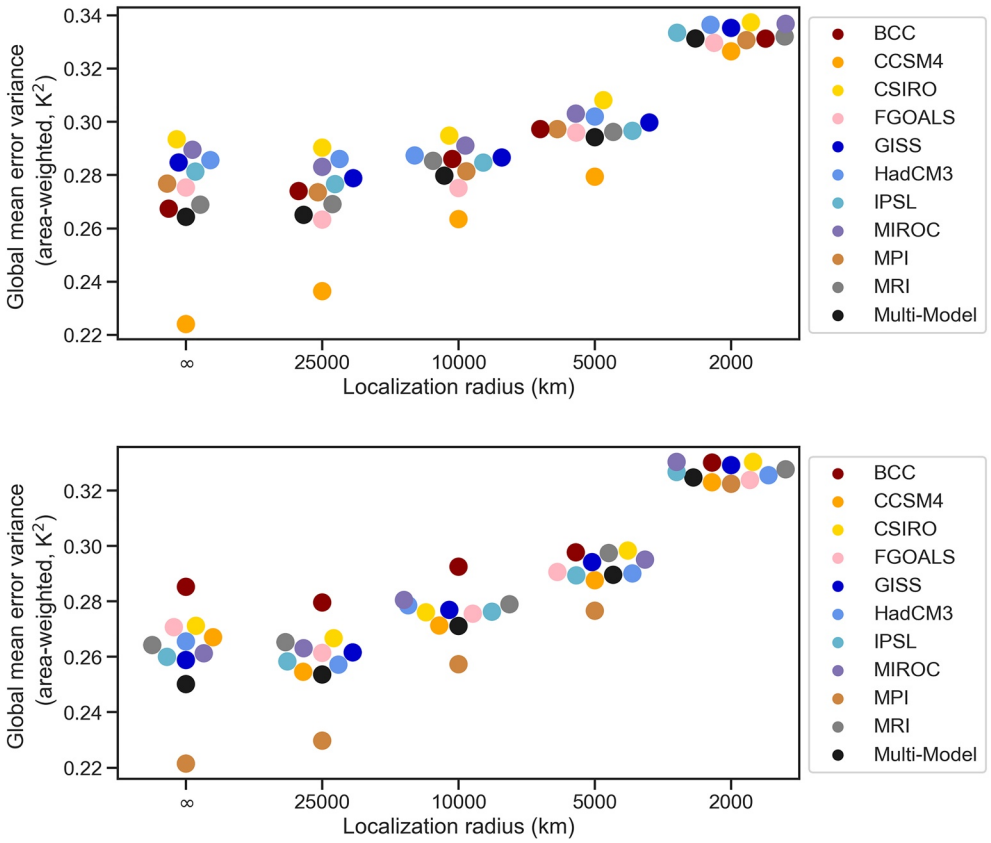

Figure 5. Global-mean, area-weighted local error variance for various localization radii for reconstructions of the CCSM4 target (top) and MPI target (bottom) using observations drawn from 542 PAGES Consortium paleoclimate proxy locations. Black circles mark error variance of multi-model ensemble and colored dots mark single-model ensemble reconstructions. Note the non-localized multi-model ensemble tends to show lower error than the localized single-model ensembles, and localized multi-model ensembles show minimal/no improvement over non-localized multi-model ensembles.

3.3. Instrumental Era Reconstructions

In Section 3.1, we show that MMEs produce lower reconstruction error than single-model ensembles when used to reconstruct a withheld CMIP5 past1000 target. We extend this experimental framework to reconstruct reanalysis and observation-based gridded temperature products of the instrumental era to test if the MME produces lower reconstruction error if the reconstruction target is not drawn from a CMIP5 model simulation. First, we compare errors from single and MMEs used to reconstruct 20CRv3 annual-mean temperature variability (1916–2015). As in the climate model reconstruction experiments shown in Sections 3.1 and 3.2, we draw observations at the same proxy locations (Figure S1), but now from the 20CRv3 data. We find that the 10-model MME shows lower global-mean error variance than any of the single-model ensemble reconstructions of 20CRv3. Maps in Figure 6 show the difference in normalized error variance among the MME and each single-model ensemble for reconstructions of 20CRv3, and panel titles in Figure 6 show differences among the MME and each single-model ensemble's global-mean error variance. Regions of high (non-normalized) reconstruction error are concentrated in the high-latitude Southern Ocean and high-latitude northeastern Asia, and the MME tends to show decreases in error in these regions (not shown). Regions of high normalized error variance in individual reconstructions are concentrated over the mid-latitudes (not shown). The MME tends to reduce normalized error the most in the tropics and mid-latitudes, with some exceptions (Figure 6).

As mentioned in Section 2.6, the 20CRv3 uses a model in the data assimilation process to produce spatially complete fields. Therefore, we change our reconstruction target to the HadCRUTv46 data set, which uses no spatial interpolation or infilling, to test if the improved skill of the MME is dependent on targeting a

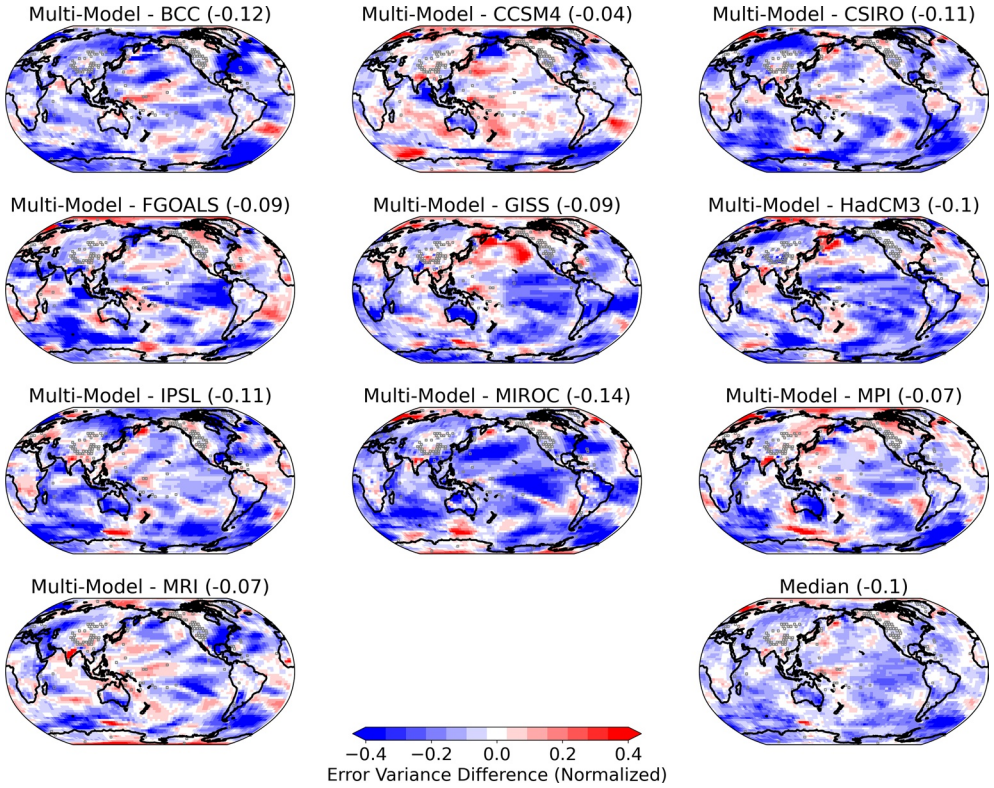

Figure 6. Maps of differences in normalized error variance among the multi-model ensemble and the single-model ensembles for reconstructions of the 20th Century Reanalysis v3 target (1916–2015) with observations drawn from 542 PAGES Consortium proxy locations (gray squares on maps). Numbers in parentheses in titles show the global-mean, area-weighted normalized error variance differences among the multi-model ensemble and the single-model ensemble (negative global-mean error values indicate lower error in multi-model ensemble). Blue colors show lower error in multi-model ensemble reconstructions, and red colors show lower error in the median single-model ensemble reconstructions of the 20th Century Reanalysis.

model in the reconstruction. We draw pseudo-observations from HadCRUTv46 temperature data at paleoclimate proxy locations that geographically overlap with the HadCRUTv46 grid points where observations are available every year 1960–2019 (Figure S1). Reconstructions of HadCRUTv46 show similar results to the 20CRv3 reconstructions: the MME produces lower global-mean error than the single-model ensembles, with the largest decreases in error over northern Asia and increases in error in several cases over northern Europe (Figure 7).

3.4. Mechanisms Underlying Improved MME Performance

Overall, the improved performance of the MME is consistent with the expectation that prior uncertainties in representing past climate states with simulated model fields originate from structural uncertainties in model representations of climate physics (as quantified by Amrhein et al., 2020) in addition to the internal and forced model variability in time that is represented by single-model ensembles. By including multiple models with different internal variability and responses to external forcing, MMEs incorporate an estimate of these structural effects and can describe uncertainty in the prior due to uncertain physical relationships in the form of weaker covariances in these regions. In these regions, covariance will be downweighted in a MME, leading to a smaller imprint of covariance error and a more accurate reconstruction.

Here we illustrate the differences among single-model and MME priors and connect them to improvements in reconstruction skill. The prior covariance affects the solution through its estimate of 1) covariance patterns between observational locations and global fields (the $\mathbf{BH}^T$ term in Equations 3) and 2) local variance at observation locations (the $\mathbf{HBH}^T$ term). Errors in either of these terms can contribute to reconstruction

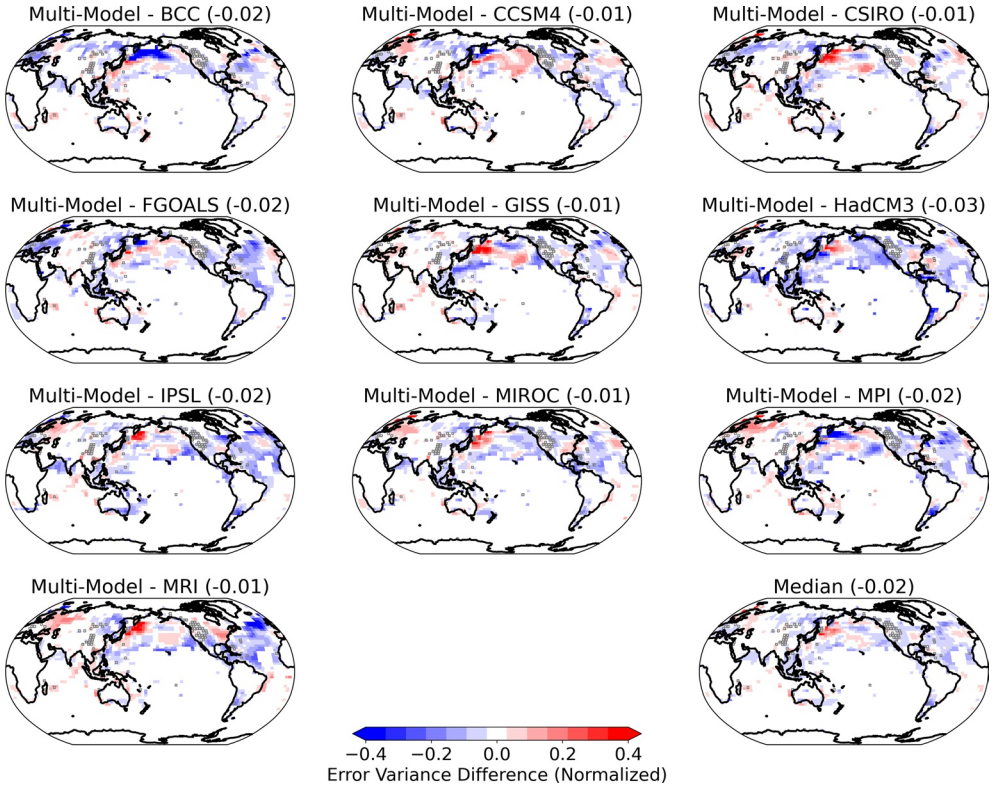

Figure 7. Maps of differences in normalized error variance among the multi-model and the single-model ensembles for reconstructions of instrumental data using HadCRUTv46 observations drawn from 542 PAGES Consortium proxy locations (gray squares on maps). Numbers in parentheses in titles show the global-mean, area-weighted local error variance differences among multi-model versus single-model ensemble reconstructions (negative global-mean normalized error variance values indicate lower error in multi-model ensemble reconstructions). Blue colors show lower error variance in multi-model ensemble reconstructions, and red colors show lower reconstruction error in the median single-model ensemble reconstructions.

errors; here we focus on covariance, motivated by Amrhein et al. (2020), who suggested this term may have a larger impact than local variance. One way to characterize these patterns is the isotropic covariance length scale (described in Section 2.8), which is a measure of the typical distance over which an observation can influence the solution. Comparing the covariance length scales in the MME to the length scales from the CCSM4 single-model ensemble reveals that the MME has reduced covariance length scales at nearly all PAGES Consortium proxy locations (Figure 8a). Figure 8b shows that the MME consistently has shorter length scales than nearly all single-model ensembles, particularly at mid- and high-latitude proxy locations. This result suggests that there is greater model agreement on covariance patterns at shorter length scales, and that attenuating uncertain longer-range teleconnections improves reconstruction skill when using an MME.

Covariance length scales are limited in their ability to characterize anisotropic relationships or distant climate teleconnections. A complementary analysis considers the number of independently varying spatial patterns in different prior ensembles by computing the estimated spatial degrees of freedom (ESDOFs; Bretherton et al., 1999) in different prior ensembles. We find that the MME tends to have more ESDOFs than any single-model ensemble except HadCM3 (Figure 8c). This increase in spatial degrees of freedom amounts to a greater number of unknown variables in the prior ensemble that must be constrained by observations.

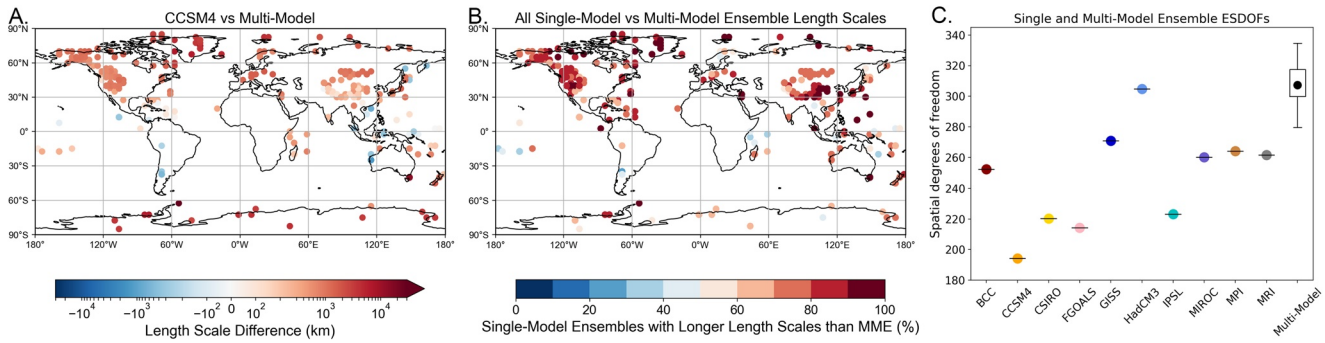

Figure 8. Differences in covariance length scales and degrees of freedom in single-model ensembles and multi-model ensemble (MME). (a) Map showing differences in covariance length scales between the CCSM4 single-model ensemble and MME, with red colors indicating longer length scales in the CCSM4 and blue colors indicating longer length scales in the MME. (b) Map showing the percent of single-model ensembles with longer covariance length scales than the MME, with red colors indicating >50% of single-model ensembles show longer length scales than the MME at that location. Covariance length scales are computed from interannual surface air temperature at 542 PAGES Consortium proxy locations. (c) Effective spatial degrees of freedom (ESDOF; Bretherton et al., 1999) for each single-model ensemble and the MME. Boxplot shows the median (black dot) and interquartile range (box) of the ESDOF calculated for 100 random draws of the MME.

4. Discussion and Conclusions

Here we compare reconstruction error from an offline ensemble Kalman filter using random draws from single CMIP5 past1000 models to errors from reconstructions conducted using the same size multi-model ensembles (MMEs). We consistently find lower global-mean reconstruction error in MMEs when using pseudo-observations drawn from the climate models at the 542 sub-annual and annually resolved PAGES Consortium paleoclimate proxy locations. Although the MME on average yields the most skillful reconstructions when global-mean error variance is used as the metric (Figure 3, Table S1), in certain cases single-model ensembles can produce lower global, regional, or local error (Figures 1–4). Absolute reconstruction skill in pseudo-proxy experiments may vary for a variety of reasons, including the nature of the spatial covariance in the target climate model, how the distribution of the observations samples that spatial covariance, and the temporal variability of the spatial covariance. We find the maximum possible skill in these reconstructions can vary depending on the target climate model in the reconstruction (Figure 3; see also Smerdon et al., 2016), but the MME consistently outperforms the single-model ensembles. We have tested the sensitivity of the MME results to temporal stability of spatial covariance by re-sampling 100 years from each CMIP5 past1000 model to create a new MME, and find no noticeable changes in our main results (Figure S6). Additionally, we have tested the sensitivity of our results by re-sampling 100 random target years from each target model 10 times and find the MME most consistently shows the highest reconstruction skill (Figure S7, Table S1), even though the absolute error can change depending on the target years of the reconstruction.

We also test the sensitivity of our results to the sparsity of the observation network because the 542 proxy locations used in most experiments here do not represent a realistic snapshot of geographic locations of available annually resolved observations before the 18th century. Therefore, we also consider reconstruction error in the single-model and MME reconstructions using observations drawn from locations corresponding to a sparser network of 127 locations concentrated in mid and high latitudes (middle panel Figure S1). Use of this more limited observation network also helps illustrate the importance of covariance in spatially propagating information from locations of observations to regions lacking local proxy information. The MME still consistently performs as well as the top several single-model ensembles if not better than all single-model ensembles (Figure S8).

We also reconstruct instrumental-era variability by assimilating pseudo-observations drawn from 20CRv3 at paleoclimate proxy locations in the "full" (542 locations) and "reduced" (127 locations) networks. As before, we find lower reconstruction error in the MME than in the single-model ensembles in the "full network" experiment (Figure 6 shows 20CRv3 results, Figure 7 shows HadCRUTv46 results). If we limit the proxy network to the lower-density, mid and high latitude proxy network locations described above, the MME produces lower error for reconstructions of 20CRv3 than any of the single prior ensembles, with

particularly large decreases in normalized error in the tropics (Figure S9), a region with notably fewer observations in the sparse proxy network (middle panel, Figure S1). These results again emphasize the importance of improving spatial covariance estimates for the reconstruction of regions that are not well constrained by observations.

We attribute improved MME performance to the explicit representation of structural modeling uncertainties in the data assimilation prior ensemble. The MME shows a greater number of effective spatial degrees of freedom (Bretherton et al., 1999) and decreased average covariance length scales at nearly all data locations (Figure 8), leading to fewer a priori constraints and affording observations more power in determining climate covariance relationships, particularly at greater length scales. Based on these analyses, we suggest that MMEs improve reconstruction skill by giving observations greater sway in reconstructing long-range climate relationships by piecing together a larger set of shorter-range prior covariances that are robust among multiple climate models. Our results are intuitively similar to previous work showing that multi-model mean properties can provide better representations of climate evolution (e.g., Flato et al., 2013; Pierce et al., 2009; Reichler & Kim, 2008), although here we are considering model covariances. We also expect MMEs to yield more accurate representations of posterior uncertainty, which can suffer from inaccurate prior representations (Amrhein et al., 2020).

Covariance localization is often used to lower reconstruction error in single-model ensemble paleoclimate reconstructions (e.g., Liu et al., 2017; Steiger et al., 2014; Tardif et al., 2019). Since ensemble sample error (the usual motivator for localization in meteorological applications) can be made arbitrarily small with large ensembles in offline data assimilation, our results suggest that any error reduction is likely due to reducing the effects of spurious long-distance covariances due to model error. Moreover, while both localization and MME approaches tend to reduce covariance length scales, we show here that MMEs can reduce reconstruction error more effectively than isotropic covariance localization (Figure 5), as hypothesized in Amrhein et al. (2020).

Despite the consistent reduction in reconstruction error for the MME shown in these idealized experiments, some of these results should be interpreted with caution. Specifically, treating individual model simulations from a Coupled Modeling Intercomparison Project as independent can be problematic because models from different centers often share similar components (e.g., Bishop & Abramowitz, 2013; Knutti et al., 2013), and as noted above, can contain similar biases (Tebaldi & Knutti, 2007). Therefore, MME experiments that combine models from different modeling centers can produce results that appear to improve reconstruction skill, but this improvement could be based on the same systematic bias that appears in multiple models (Tierney et al., 2015). For example, most climate models show cold tongue bias in the tropical Pacific (Li et al., 2016), and a MME approach is unlikely to reduce the presence of those biases in reconstructions. However, the MME shows improved skill for reconstructions of the 20CRv3 or HadCRUTv46, so the decreased reconstruction error in the MME cannot be explained by shared model bias (Figures 6 and 7). Future work could test if model bias correction methods (e.g., Steiger et al., 2018) change reconstruction skill in the MME. However, several possible drawbacks to bias correction procedures should be considered, including distortion of covariance structure (Cannon, 2016; Rocheta et al., 2014), non-stationary bias in climate models (e.g., Reifen and Toumi, 2009), and non-stationarities in climate covariance in observations and models (e.g., Coats et al., 2013; Gallant et al., 2013; Cole and Cook, 1998). Finally, bias correction could impose a common covariance from an instrumental data product onto models used in the MME, thereby potentially reducing the diversity of patterns of variability from models that appears to underpin improved reconstruction skill in the MME.

Future work could also focus on weighted combinations of multiple climate models to create the MME. An ideal procedure would be to compose a MME based not on a small sample of different coupled model architectures, but instead on systematically varying unknown model parameters according to their likelihood. Combining varying numbers of climate models in the MME may reduce reconstruction error, but determining the ideal combination of models in the MME will likely depend on the metric used to measure this skill (e.g., the region or variable targeted). Additionally, an exercise focused on determining which combination of climate models best reconstructs a withheld model may be of limited utility because results may simply reveal similarities in related climate models from different modeling centers (e.g., Knutti et al., 2013). Therefore, we leave to future work testing the skill of weighting different models in multi-model ensembles.

In summary, we have found that MMEs decrease reconstruction errors relative to single-model prior ensembles. We have also illustrated that use of different CMIP5 simulations in a paleoclimate data assimilation context can result in a range of reconstruction skill, with notable spatial heterogeneity. Using just one climate model in these paleoclimate reconstructions does not fully account for structural uncertainties in the climate system. The combination of single climate models into MMEs shown here represents a first step in an attempt to account for sources of uncertainty in model physics. Future work could benefit from weighting different combinations of climate model priors (e.g., Krishnamurti et al., 1999; Robertson et al., 2004) to maximize reconstruction skill.

Although we have focused on 2-m air temperature reconstruction associated with Common Era paleoclimate observation locations, future work could also apply this method to other climate variables and reconstruction time periods. Variables such as precipitation, geopotential height, or sea ice concentration tend to covary with one another and with surface temperature quite differently among CMIP5 models (e.g., Coats et al., 2013; Parsons et al., 2018; Weare, 2013), so using MMEs could help account for some of this heterogeneity in model covariance. For example, CMIP5 models tend to simulate varied mean states and teleconnections associated with Arctic sea ice concentration (e.g., Bonan & Blanchard-Wrigglesworth, 2019; Li et al., 2017), so using a MME may help reduce uncertainty where models disagree on sea ice coverage. Multi-model ensembles could also be used to reconstruct time periods beyond the Common Era, in which paleoclimate observations are even more geographically sparse and covariance uncertainty is expected to play an even larger role in reconstructions.

Data Availability Statement

Monthly surface air temperature variable is used as input in all experiments. CMIP5 temperature data can be found at: https://esgf-node.llnl.gov/search/cmip5/. CMIP5 data from these simulations were selected using the following criteria on the ESGF website: CMIP5/PMIP3 "Project," past1000 "Experiment," mon "Time Frequency," "atmos" Realm, r1i1p1 (r1i1p121 for GISS) "Ensemble," and tas "Variable." Instrumental-based surface temperature data provided by the Climate Research Unit, University of East Anglia at: https://crudata.uea.ac.uk/cru/data/temperature/HadCRUT.4.6.0.0.median.nc. 20th Century Reanalysis V3 data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at https://psl.noaa.gov/, with a specific link to ensemble mean of the 2-m air temperature data at: ftp://ftp2.psl.noaa.gov/Datasets/20thC_ReanV3/Monthlies/2mSI-MO/air.2m.mon.mean.nc.

Acknowledgments
L. A. Parsons thanks the Washington Research Foundation (WRF) Postdoctoral Fellowship for funding support at the University of Washington and NASA GISS grant 80NSSC19M0138 for funding support at Duke. M. K. Brennan was supported by an NSF graduate research fellowship grant DGE-1762114. S. C. Sanchez was supported by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) Postdoctoral Fellowship. R. Tardif and G. J. Hakim were supported by NSF grant AGS-1702423 and Heising-Simons Foundation Grant 2016-014 awarded to the University of Washington. This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. We acknowledge the World Climate Research Program's Working Group on Coupled Modeling, which is responsible for the CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Code for installing and running the Last Millennium Reanalysis and associated ensemble Kalman filter available at: https://github.com/modons/LMR.

References

Ahmed, M., Anchukaitis, K. J., Asfawossen, A., Borgaonkar, H. P., Braida, M., Buckley, B. M., et al. (2013). Continental-scale temperature variability during the past two millennia. *Nature Geoscience*, *6*, 339–346.

Amrhein, D., Parsons, L., & Hakim, G. (2020). Quantifying structural uncertainty in paleoclimate data assimilation with an application to the last millennium? *Geophysical Research Letters*, *47*, e2020GL090485. https://doi.org/10.1029/2020GL090485

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., & García-Herrera, R. (2011). The hot summer of 2010: Redrawing the temperature record map of Europe. *Science*, *332*(6026), 220–224. https://doi.org/10.1126/science.1201224

Beadling, R. L., Russell, J. L., Stouffer, R. J., Goodman, P. J., & Mazloff, M. (2019). Assessing the quality of Southern Ocean circulation in CMIP5 AOGCM and earth system model simulations. *Journal of Climate*, *32*(18), 5915–5940. https://doi.org/10.1175/jcli-d-19-0263.1

Bhend, J., Franke, J., Folini, D., Wild, M., & Brönnimann, S. (2012). An ensemble-based approach to climate reconstructions. *Climate of the Past*, *8*(3), 963–976. https://doi.org/10.5194/cp-8-963-2012

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, *41*(3–4), 885–900. https://doi.org/10.1007/s00382-012-1610-y

Bonan, D. B., & Blanchard-Wrigglesworth, E. (2020). Nonstationary teleconnection between the Pacific Ocean and Arctic Sea Ice. *Geophysical Research Letters*, *47*(2), e2019GL085666. https://doi.org/10.1029/2019gl085666

Bothe, O., Evans, M., Fernandez Donado, L., & Elena Garcia Bustamante, (2015). Continental-scale temperature variability in PMIP3 simulations and PAGES 2k regional temperature reconstructions over the past millennium. *Climate of the Past*, *11*, 1673–1699.

Bothe, O., Jungclaus, J. H., & Zanchettin, D. (2013). Consistency of the multi-model CMIP5/PMIP3-past1000 ensemble. *Climate of the Past*, *9*, 2471–2487. https://doi.org/10.5194/cp-9-2471-2013

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., & Bladé, I. (1999). The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, *12*(7), 1990–2009. https://doi.org/10.1175/1520-0442(1999)012<1990:tenosd>2.0.co;2

Cannon, A. J. (2016). Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure. *Journal of Climate*, *29*(19), 7045–7064. https://doi.org/10.1175/jcli-d-15-0679.1

Christiansen, B. (2020). Understanding the distribution of multi-model ensembles. *Journal of Climate*, *33*(21), 1–52.

Coats, S., Smerdon, J. E., Cook, B. I., & Seager, R. (2013). Stationarity of the tropical pacific teleconnection to North America in CMIP5/PMIP3 model simulations. *Geophysical Research Letters*, *40*, 4927–4932.

Cole, J. E., Cook, E. R. (1998). The changing relationship between ENSO variability and moisture balance in the continental United States. *Geophysical Research Letters*, *25*(24), 4529–4532. https://doi.org/10.1029/1998gl900145
Collins, M., Tett, S. F. B., & Cooper, C. (2001). The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics*, *17*(1), 61–81. https://doi.org/10.1007/s003820000094
Cook, E. R., Anchukaitis, K. J., Buckley, B. M., D'Arrigo, R. D., Jacoby, G. C., & Wright, W. E. (2010). Asian monsoon failure and megadrought during the last millennium. *Science*, *328*(5977), 486–489. https://doi.org/10.1126/science.1185188
Cook, E. R., Woodhouse, C. A., Eakin, C. M., Meko, D. M., & Stahle, D. W. (2004). Long-term aridity changes in the western United States. *Science*, *306*(5698), 1015–1018. https://doi.org/10.1126/science.1102586
Doblas-Reyes, F. J., Déqué, M., & Piedelievre, J. (2000). Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quarterly Journal of the Royal Meteorological Society*, *126*(567), 2069–2087.
Emile-Geay, J., McKay, N. P., Kaufman, D. S., Von Gunten, L., Wang, J., Anchukaitis, K. J., et al. (2017). A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data*, *4*, 170088.
Evans, M. N., Kaplan, A., & Cane, M. A. (2002). Pacific sea surface temperature field reconstruction from coral δ18O data using reduced space objective analysis. *Paleoceanography*, *17*(1), 7–1. https://doi.org/10.1029/2000pa000590
Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W. J., et al. (2013). Evaluation of climate models. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Vol. 5, pp. 741–866).
Gallant, A. J. E., Phipps, S. J., Karoly, D. J., Mullan, A. B., Lorrey, A. M. (2013). Nonstationary Australasian Teleconnections and Implications for Paleoclimate Reconstructions. *Journal of Climate*, *26*(22), 8827–8849. https://doi.org/10.1175/jcli-d-12-00338.1
Gaspari, G. & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757.
Gupta, A. S., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate Drift in the CMIP5 Models*. *Journal of Climate*, *26*, 8597–8615. https://doi.org/10.1175/jcli-d-12-00521.1
Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting-I. Basic concept. *Tellus A*, *57*(3), 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x
Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., et al. (2016). The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres*, *121*, 6745–6764. https://doi.org/10.1002/2016jd024751
Holmes, C. R., Holland, P. R., & Bracegirdle, T. J. (2019). Compensating biases and a noteworthy success in the CMIP5 representation of Antarctic sea ice processes. *Geophysical Research Letters*, *46*(8), 4299–4307. https://doi.org/10.1029/2018gl081796
Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., & Mitchell, H. L. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, *124*(6), 1225–1242. https://doi.org/10.1175/1520-0493(1996)124<1225:assate>2.0.co;2
Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–1199. https://doi.org/10.1002/grl.50256
Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., et al. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, *285*(5433), 1548–1550. https://doi.org/10.1126/science.285.5433.1548
Li, D., Zhang, R., & Knutson, T. R. (2017). On the discrepancy between observed and CMIP5 multi-model simulated Barents Sea winter sea ice decline. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/ncomms14991
Li, G., Xie, S.-P., Du, Y., & Luo, Y. (2016). Effects of excessive equatorial cold tongue bias on the projections of tropical Pacific climate change. Part I: The warming pattern in CMIP5 multi-model ensemble. *Climate Dynamics*, *47*(12), 3817–3831. https://doi.org/10.1007/s00382-016-3043-5
Liu, W., Xie, S., Liu, Z., & Zhu, J. (2017). Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate. *Science Advances*, *3*(1), e1601666. https://doi.org/10.1126/sciadv.1601666
Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., & Wanner, H. (2004). European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, *303*(5663), 1499–1503. https://doi.org/10.1126/science.1093877
Mann, M. E., Bradley, R. S., & Hughes, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, *392*(6678), 779–787. https://doi.org/10.1038/33859
Mann, M. E., Bradley, R. S., & Hughes, M. K. (1999). Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters*, *26*(6), 759–762. https://doi.org/10.1029/1999gl900070
Mann, M. E., & Rutherford, S. (2002). Climate reconstruction using 'Pseudoproxies'. *Geophysical Research Letters*, *29*(10), 1391–1394. https://doi.org/10.1029/2001gl014554
Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., et al. (2009). Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science*, *326*(5957), 1256–1260. https://doi.org/10.1126/science.1177303
Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research*, *117*, D08101. https://doi.org/10.1029/2011JD017187
Oke, P. R., Allen, J. S., Miller, R. N., Egbert, G. D., & Kosro, P. M. (2002). Assimilation of surface velocity data into a primitive equation coastal ocean model. *Journal of Geophysical Research: Oceans*, *107*(C9), 5–25. https://doi.org/10.1029/2000jc000511
Oke, P. R., Sakov, P., & Corney, S. P. (2007). Impacts of localisation in the EnKF and EnOI: Experiments with a small model. *Ocean Dynamics*, *57*(1), 32–45. https://doi.org/10.1007/s10236-006-0088-8
Parsons, L. A., Coats, S., & Overpeck, J. T. (2018). The continuum of drought in Southwestern North America. *Journal of Climate*, *31*(20), 8627–8643. https://doi.org/10.9783/9780812295450
Parsons, L. A., & Hakim, G. J. (2019). Local regions associated with interdecadal global temperature variability in the Last Millennium Reanalysis and CMIP5 models. *Journal of Geophysical Research: Atmospheres*, *124*(1), 9905–9917. https://doi.org/10.1029/2019JD030426
Pierce, D. W., Barnett, T. P., Santer, B. D., & Gleckler, P. J. (2009). Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, *106*(21), 8441–8446. https://doi.org/10.1073/pnas.0900094106
Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–312. https://doi.org/10.1175/bams-89-3-303
Reifen, C., Toumi, R. (2009). Climate projections: Past performance no guarantee of future skill?. *Geophysical Research Letters*, *36*(13), https://doi.org/10.1029/2009gl038082
Robertson, A. W., Lall, U., Zebiak, S. E., & Goddard, L. (2004). Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Monthly Weather Review*, *132*(12), 2732–2744. https://doi.org/10.1175/mwr2818.1

Rocheta, E., Evans, J. P, Sharma, A. (2014). Assessing atmospheric bias correction for dynamical consistency using potential vorticity. *Environmental Research Letters*, *9*(12), 124010. https://doi.org/10.1088/1748-9326/9/12/124010

Rutherford, S., Mann, M. E., Osborn, T. J., Briffa, K. R., Jones, P. D., Bradley, R. S., & Hughes, M. K. (2005). Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain. *Journal of Climate*, *18*(13), 2308–2329. https://doi.org/10.1175/jcli3351.1

Schmidt, G. A., Jungclaus, J. H., Ammann, C. M., Bard, E., Braconnot, P., Crowley, T. J., et al. (2011). Climate forcing reconstructions for use in PMIP simulations of the last millennium (v1.0). *Geoscientific Model Development*, *4*, 33–45. https://doi.org/10.5194/gmd-4-33-2011

Singh, H. K. A., Hakim, G. J., Tardif, R., Emile-Geay, J., & Noone, D. C. (2018). Insights into Atlantic multidecadal variability using the Last Millennium Reanalysis framework. *Climate of the Past*, *14*, 157. https://doi.org/10.5194/cp-14-157-2018

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., et al. (2019). Toward a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908.

Smerdon, J. E. (2012). Climate models as a test bed for climate reconstruction methods: Pseudoproxy experiments. *WIREs Climate Change*, *3*(1), 63–77. https://doi.org/10.1002/wcc.149

Smerdon, J. E., Coats, S., & Ault, T. R. (2016). Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era. *Climate Dynamics*, *46*, 1921–1942. https://doi.org/10.1007/s00382-015-2684-0

Smerdon, J. E., Kaplan, A., Chang, D., & Evans, M. N. (2010). A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium*. *Journal of Climate*, *23*(18), 4856–4880. https://doi.org/10.1175/2010jcli3328.1

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., & Roe, G. H. (2014). Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, *27*, 426–441. https://doi.org/10.1175/jcli-d-12-00693.1

Steiger, N. J., Smerdon, J. E., Cook, E. R., & Cook, B. I. (2018). A reconstruction of global hydroclimate and dynamical variables over the Common Era. *Scientific Data*, *5*, 180086. https://doi.org/10.1038/sdata.2018.86

Stephenson, D. B., Coelho, C. A. S., Doblas-Reyes, F. J., & Balmaseda, M. (2005). Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 253–264. https://doi.org/10.3402/tellusa.v57i3.14664

Sueyoshi, T., Ohgaito, R., Yamamoto, A., Chikamoto, M. O., Hajima, T., Okajima, H., et al. (2013). Set-up of the PMIP3 paleoclimate experiments conducted using an Earth system model, MIROC-ESM. *Geoscientific Model Development*, *6*(3), 819–836. https://doi.org/10.5194/gmd-6-819-2013

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., et al. (2019). Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, *15*(4). https://doi.org/10.5194/cp-15-1251-2019

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of Cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Tierney, J. E., Ummenhofer, C. C., & deMenocal, P. B. (2015). Past and future rainfall in the Horn of Africa. *Science Advances*, *1*(9), e1500682.

Von Storch, H., Zorita, E., Jones, J. M., Dimitriev, Y., González-Rouco, F., & Tett, S. F. (2004). Reconstructing past climate from noisy data. *Science*, *306*(5696), 679–682. https://doi.org/10.1126/science.1096109

Weare, B. C. (2013). El Niño teleconnections in CMIP5 models. *Climate Dynamics*, *41*(7–8), 2165–2177. https://doi.org/10.1007/s00382-012-1537-3

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. https://doi.org/10.1002/qj.210