Supporting Information for "Coupled Atmosphere–Ocean Reconstruction of the Last Millennium Using Online Data Assimilation"

W. A. Perkins[1], G. J. Hakim[1]

[1]University of Washington

[1]Seattle, WA

Contents of this file

1. Text S1
2. Text S2
3. Figures S1 to S9
4. Tables S1 to S3

Additional Supporting Information (Files uploaded separately)

1. Captions for Movies S1 to S2

Introduction

The provided supporting information covers the process by which we determine the production LIM calibration parameters, and describes the details of sensitivity testing regarding the cooler temperatures near the beginning of the reconstruction (∼1000–1400

Corresponding author: W. A. Perkins, Department of Atmospheric Sciences, University of Washington (wperkins@uw.edu)

C.E.). Additionally, we provide figures and two movies as supporting information to the main text.

S1. LIM Calibration Testing

To test the ensemble forecast characteristics of the LIM relevant for data assimilation, we investigate ensemble calibration ratios (ECRs; e.g., as in Perkins & Hakim, 2020) for a number of multivariate-EOF (mvarEOF) component truncations (ℓ = 10, 15, 20, 21, 22, ..., 29, 30). For each test, we calibrate the LIM (Eq. 4) with the specified number of retained multivariate-EOF components and perform a 1-year ensemble forecast (100 ensemble members) initialized from every available year of the calibration data. ECRs are based on comparison to the reference calibration data coincident with the forecast time. The ECR measure, defined as

$$\mathrm{ECR} = \frac{1}{T}\sum_{i=1}^{T}\frac{\mathrm{SE}_t}{\sigma_t^2}, \tag{1}$$

represents the time-average (over all times, T) ratio between squared errors (SE_t) and ensemble variance (σ_t^2) calculated on the forecast ensemble ($\mathbf{g}_t$) and reference data (v_t). The squared error at time t, is defined as

$$\mathrm{SE}_t = (\overline{\mathbf{g}}_t - v_t)^2,$$

where the overline (e.g., $\overline{\mathbf{g}}_t$) denotes the ensemble average. The ensemble variance is given by

$$\sigma_t^2 = \frac{1}{N}\sum_{i=1}^{N}(g_{ti} - \overline{\mathbf{g}})^2$$

where g_{ti} represents the i^{th} of N total ensemble members. Well calibrated ensemble forecasts have a value near 1.0, while values less than 1.0 are considered overdispersive (errors are smaller than ensemble spread) and values greater than 1.0 are considered underdispersive (errors are larger than ensemble spread).

After performing the ensemble forecast experiments, we compare ensemble characteristics between them by aggregating the relative distance from 1.0 for groups of ECR quantities (for global averages, ENSO, and PDO-related measures). The function, $f(\mathbf{g}, v)$ we use to calculate distance from being well-calibrated is as follows:

$$f(\mathbf{g}, v) = \begin{cases} ln(\mathrm{ECR}(\mathbf{g}, v)) & 0 < \mathrm{ECR}(\mathbf{g}, v) < 1 \\ (\mathrm{ECR}(\mathbf{g}, v) - 1)^2 & \mathrm{ECR}(\mathbf{g}, v) \geq 1 \end{cases}. \tag{2}$$

For each reconstruction experiment, we select the multivariate-EOF truncation that displays the lowest total ECR distance from 1.0 (Tables S2 and S3). For the CCSM4-LIM, the minimum aggregate ECR occurs at a truncation of 20 modes ($f = 0.79$), and for the MPI-LIM, the minimum occurs at 27 modes ($f = 1.25$).

S2. Pre-industrial Control LIM Test

Our reconstructed global-average temperature estimates during the early period of the last millennium are cooler than many previous reconstructions (Fig. 6). Here we investigate whether the cooler average temperature during the early period of the last millennium is caused by the LIM model formulation. Specifically, we test whether the inclusion of forcing (especially volcanic response) plays a role in the cool temperatures and lack of millennium-scale cooling trend, by training another LIM based on the CCSM4 pre-industrial control simulation (no-forcing) and using it for an online DA reconstruction.

To train the pre-industrial control (piControl) LIM, we use the same fields and procedures as in the past1000 CCSM4-LIM (see Sections 3 and S1), searching for a "well-calibrated" ensemble forecast. In general, the lower variance of the piControl simulation produces underdispersive forecast ensembles when testing forecasts against piControl data (minimum ECR at 26 modes retained, $f = 2.1$) and severely underdispersive results when

testing forecasts against past1000 data (minimum at 22 modes retained, $f = 2550$). We select $l = 26$ for the number of retained multivariate-EOFs based on the forecast tests against piControl data. The $ECR >> 1$ values shown in the piControl calibration experiments indicate that variance inflation is necessary to produce a representative ensemble for data assimilation purposes. To find an appropriate inflation ratio, we perform three reconstructions (with 5 iterations each) with inflation factors of 1.0, 2.0, and 4.0, to find an inflation parameter that produces at least similar global-average TAS verification scores as in the offline case. The experiment with an inflation factor of 4.0 produced the closest verification correlations (SST ∼0.8, Niño 3.4 ∼0.7, PDO ∼0.6, OHC700m ∼0.1–0.5). Therefore, we use the inflation factor 4.0 experiment to assess whether the LIM calibrated on the past1000 data is solely responsible for the colder temperatures.

Results of the piControl-LIM experiment are compared against the offline case in Fig. S6. The global-average temperatures for the piControl-LIM (Fig. S6a) still show cooler conditions are prevalent during the early period (1000–1400 C.E.) with no clear millennium-scale cooling trend. The global-average SST and OHC700m also show cooler conditions compared to the offline case, albeit less consistently cool than in the CCSM4-LIM experiment from the main text. In general, the multi-decadal variations and uncertainty bounds are larger in the piControl-LIM experiment, which is related to our use of inflation. However, even without any information on forced-response of the climate system, the piControl-LIM reconstruction still produces colder temperatures. This strongly suggests the behavior during the early part of the reconstruction is a byproduct of the system memory introduced by the LIM and consistent reinforcement from proxy information during that time.

Movie S1.

A video showing the grand-ensemble mean (taken over the 50 × 100 Monte-Carlo iterations and ensemble members) spatial field results (TAS, SST, PR, SLP, ZG500) for the LMR Online reconstruction using the CCSM4-LIM calibration. Fields are centered about the 1000–1850 mean values. The spatial distribution of the proxy network available for assimilation in each year is provided in the lower right panel.

Movie S2.

As in Movie S1, but depicting fields of RLUT, RSUT, SST, OHC700m, SSS, and ZOS. Provided as a separate video for clarity.

References

Biondi, F., Gershunov, A., & Cayan, D. R. (2001). North Pacific Decadal Climate Variability since 1661. *Journal of Climate*, *14*(1), 5–10. doi: 10.1175/1520-0442(2001)014⟨0005:NPDCVS⟩2.0.CO;2

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. doi: 10.1175/1520-0442(1999)012⟨1990:TENOSD⟩2.0.CO;2

Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., & Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Science Advances*, *3*(3), 1–11. doi: 10.1126/sciadv.1601545

D'Arrigo, R., Villalba, R., & Wiles, G. (2001). Tree-ring estimates of Pacific decadal climate variability. *Climate Dynamics*, *18*(3-4), 219–224. doi: 10.1007/s003820100177

D'Arrigo, R., & Wilson, R. (2006). On the Asian expression of the PDO. *International*

Journal of Climatology, *26*(12), 1607–1617. doi: 10.1002/joc.1326

MacDonald, G. M., & Case, R. A. (2005). Variations in the Pacific Decadal Oscillation over the past millennium. *Geophysical Research Letters*, *32*(8), 1–4. doi: 10.1029/2005GL022478

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., & Francis, R. C. (1997). A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bulletin of the American Meteorological Society*, *78*(6), 1069–1079. doi: 10.1175/1520-0477(1997)078⟨1069:APICOW⟩2.0.CO;2

Perkins, W. A., & Hakim, G. (2020). Linear inverse modeling for coupled atmosphere-ocean ensemble climate prediction. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001778. (e2019MS001778 10.1029/2019MS001778) doi: 10.1029/2019MS001778

Shen, C., Wang, W. C., Gong, W., & Hao, Z. (2006). A Pacific Decadal Oscillation record since 1470 AD reconstructed from proxy data of summer rainfall over eastern China. *Geophysical Research Letters*, *33*(3), 1–4. doi: 10.1029/2005GL024804

Zanna, L., Khatiwala, S., Gregory, J. M., Ison, J., & Heimbach, P. (2019, jan). Global reconstruction of historical ocean heat storage and transport. *Proceedings of the National Academy of Sciences*, *116*(4), 1126–1131. doi: 10.1073/pnas.1808838115

Table S1. LMR Online (MPI-LIM) reconstruction scalar correlations with instrumental products.

Product	Glob. Avg. SST	Glob. Avg. OHC700m	Nino 3.4	PDO
HadleyEN4	0.79	0.37	0.72	0.42
GFDLECDA	0.88	0.56	0.78	0.24
ORA20C	0.91	0.85	0.72	0.55
ESRL	–	–	0.77	–
Mantua	–	–	–	0.42
Cheng17	–	0.90	–	–
Zanna19	–	0.98	–	–

Table S2. Aggregate ensemble calibration ratio (ECR) distances from the ideal ECR of 1.0 as calculated using Eq. S2. Values are shown for each test using different mvarEOF truncations during LIM calibration on CCSM4 last millennium data. The global average (Glob Avg) measure includes ECRs from TAS, SST, OHC700m, RSUT, and RLUT. The ENSO measure includes ECRs from Niño 3, 3.4, and 4 indices and the Southern Oscillation Index (SOI). The PDO measure only includes the PDO index ECR. The Total column shows the sum of Glob Avg, ENSO, and PDO aggregate ECR distances.

num mvarEOFs	Glob Avg	ENSO	PDO	**Total**
15	2.98	0.20	0.01	3.18
20	0.14	0.00	0.65	0.79
21	0.42	0.01	0.71	1.15
22	0.21	0.01	0.88	1.11
23	0.32	0.03	0.82	1.17
24	1.75	0.26	0.21	2.22
25	0.01	0.51	2.70	3.22
26	3.27	0.24	0.03	3.55
27	0.54	0.08	0.46	1.09
28	0.84	0.22	0.12	1.17
29	1.33	0.11	0.52	1.96
30	1.60	0.03	0.09	1.73

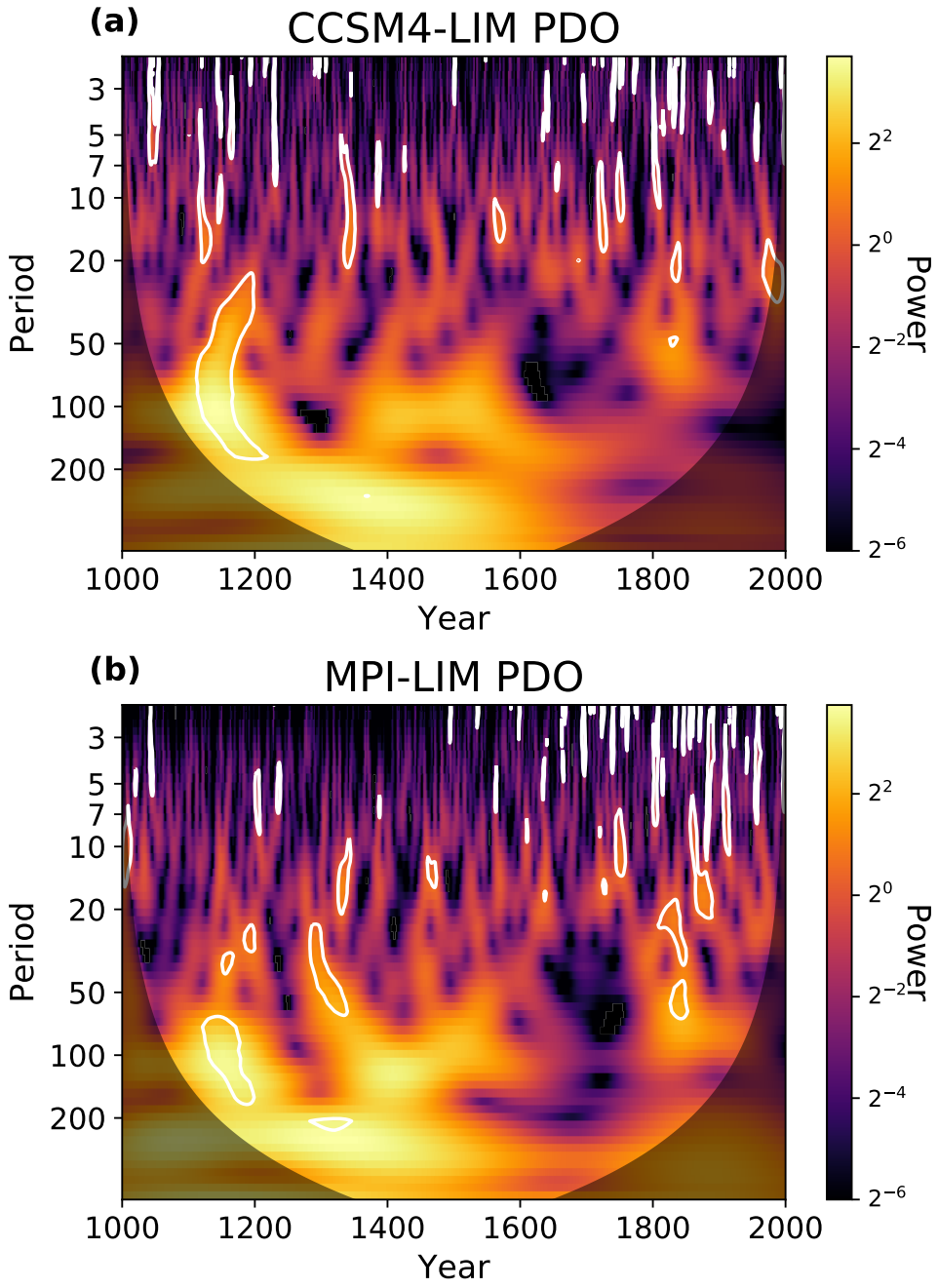

Figure S1. Complex Morlet wavelet analysis of the grand-ensemble average (taken over 50 × 100 members) Pacific Decadal Oscillation (PDO) index reconstructed in the (a) CCSM4-LIM and (b) MPI-LIM experiments. The displayed power is normalized by dividing by the input timeseries variance in each case. White contours indicate a power exceedance of a 95% confidence interval generated using 1000 integrations of a red-noise model fit to the PDO timeseries.

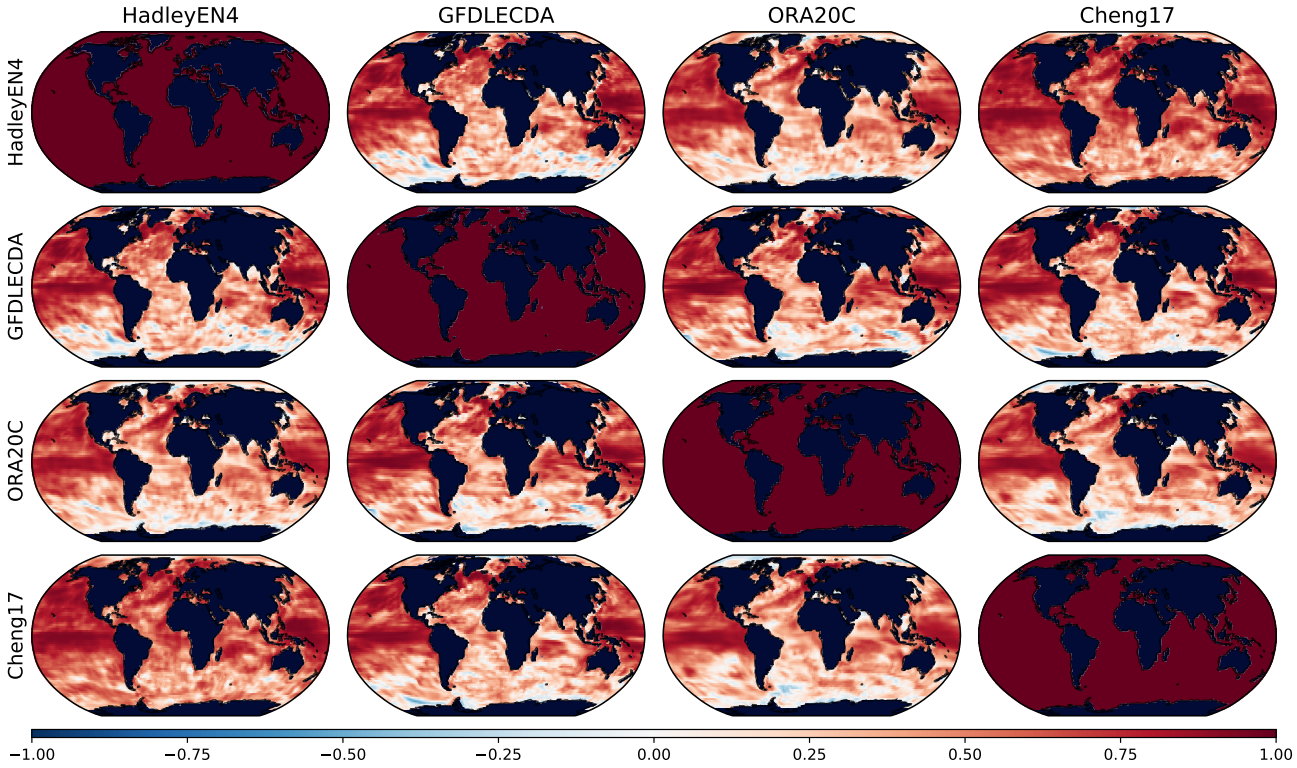

Figure S2. A gridpoint correlation comparison of the OHC700m spatial validation products with each other including HadleyEN4 (1900–2000), GFDLECDA (1961–2000), ORA-20C (1900–2000), and Cheng2017 (1940–2000)

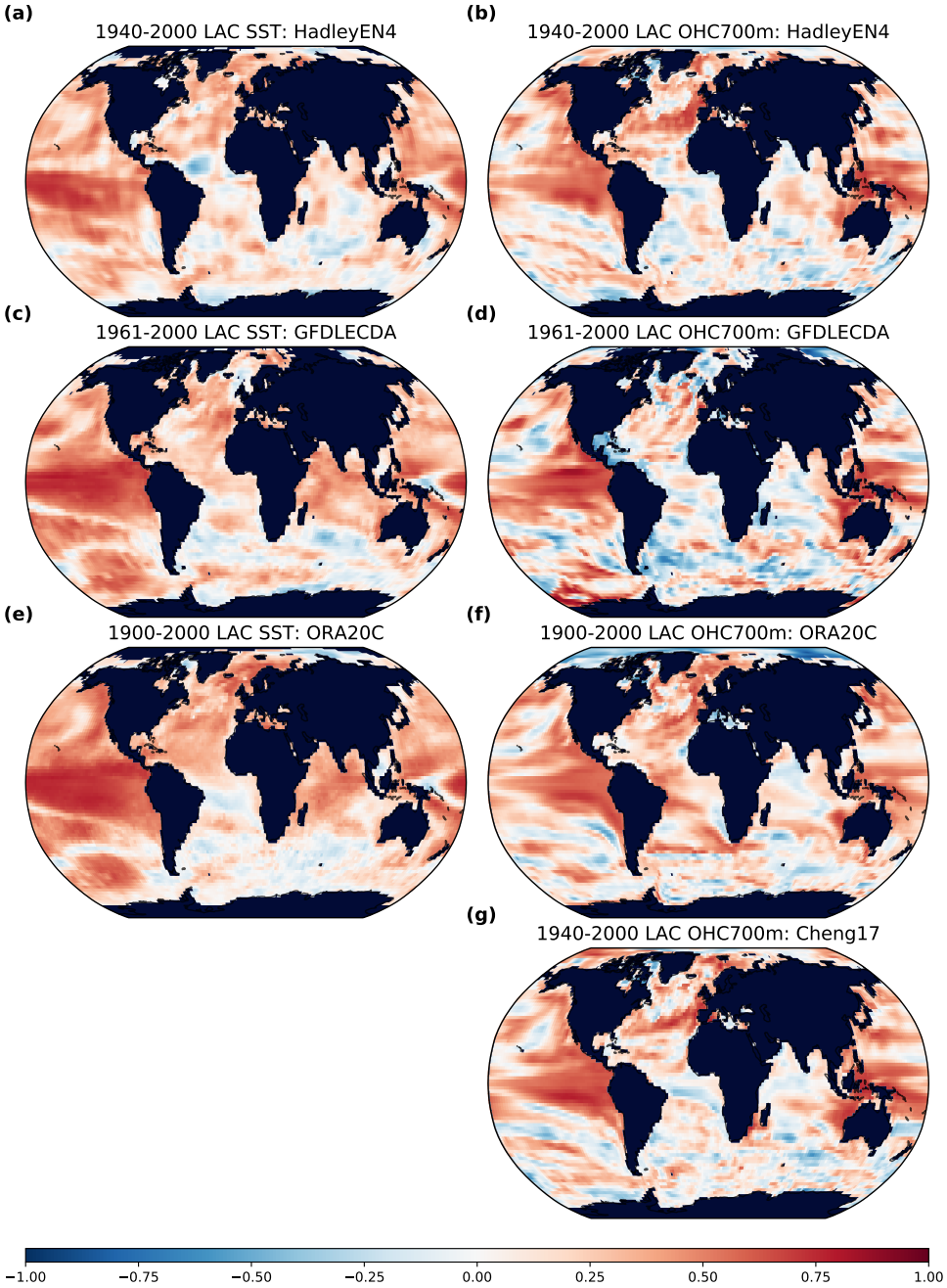

Figure S3. Detrended spatial field gridpoint correlations of the LMR MPI-LIM with Instrumental Era observational and reanalysis products for SSTs (column a) and OHC700m (column b). Spatial correlations are calculated against HadleyEN4 data (a, b; 1950-2000), GFDLECDA (c, d; 1961–2000), ORA-20C (e, f; 1900–2000), and Cheng2017 (OHC only, g; 1940–2000)

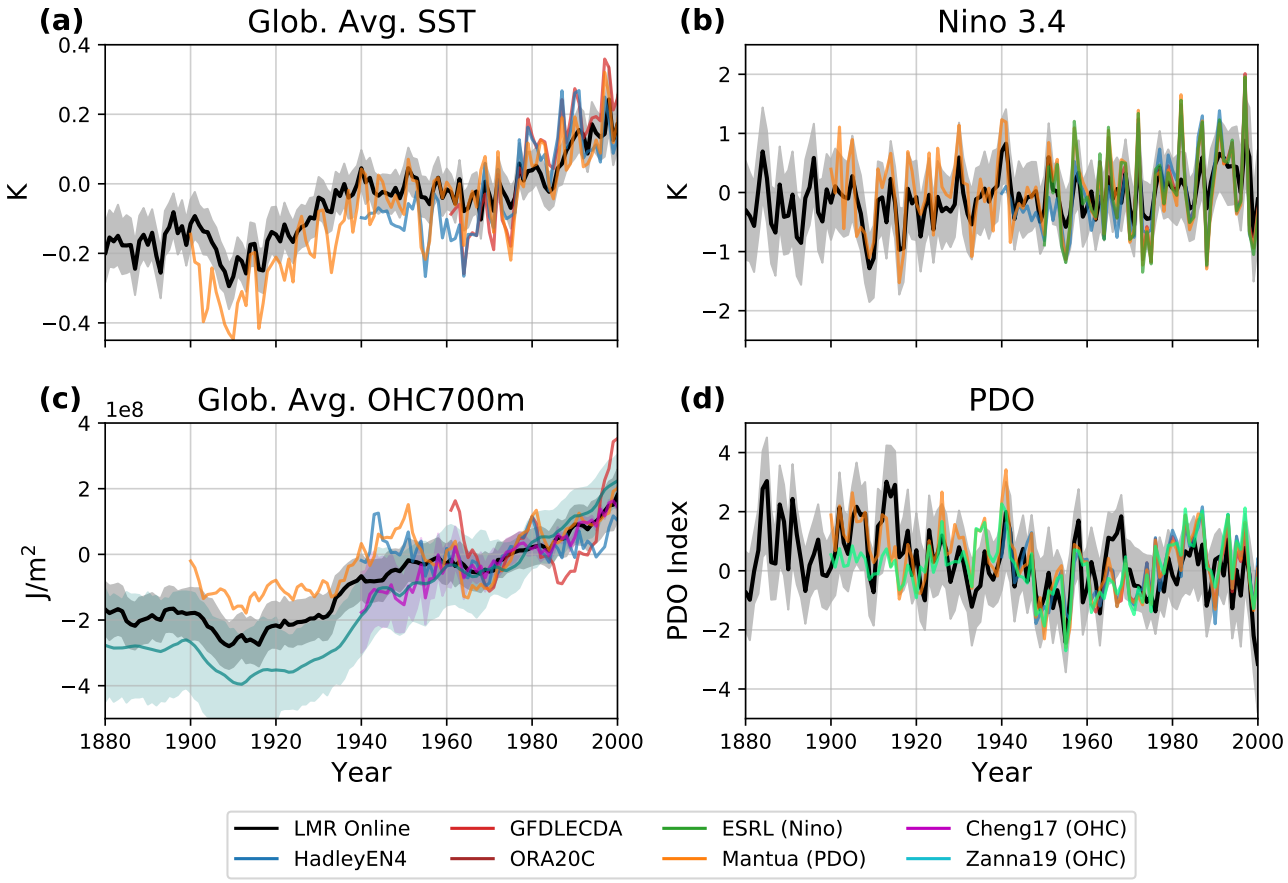

Figure S4. Scalar index comparison between the LMR Online (MPI-LIM; black with 95% confidence bounds in grey shading) reconstruction and instrumental products for (a) SST, (b) Niño 3.4, (c) OHC700m, and (d) PDO. The HadleyEN4, GFDLECDA, and ORA20C products are compared in all cases. Additionally, ESRL Niño 3.4 data, the Mantua et al. (1997) PDO index, and Cheng et al. (2017, Cheng2017) and Zanna et al. (2019, Zanna19) OHC data are compared. Error bounds ($\pm 2\sigma$) are shown for the Cheng2017 and Zanna19 OHC data.

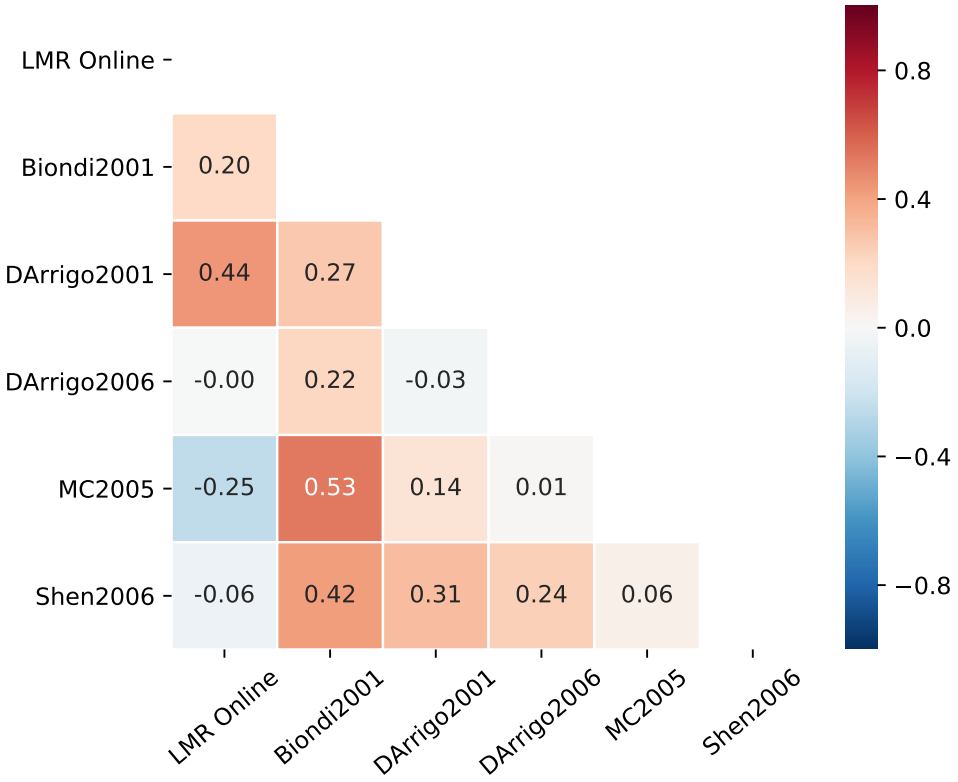

Figure S5. Correlations between the PDO index of the CCSM4-LIM reconstruction and previous PDO index reconstructions (Biondi et al., 2001; D'Arrigo et al., 2001; MacDonald & Case, 2005; D'Arrigo & Wilson, 2006; Shen et al., 2006).

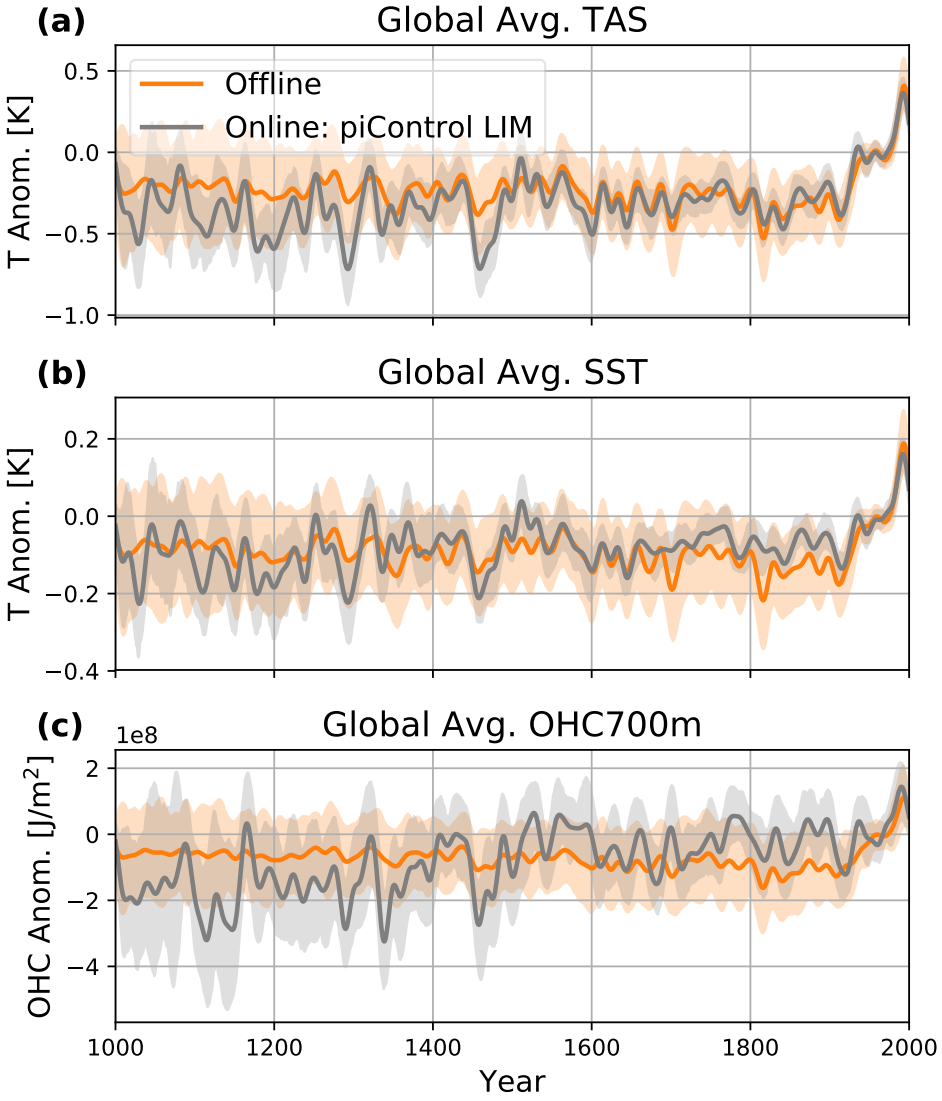

Figure S6. As in Fig. 7, but comparing the offline reconstruction against an online reconstruction using a LIM calibrated against CCSM4 preindustrial-control data. See Section S2 for details.

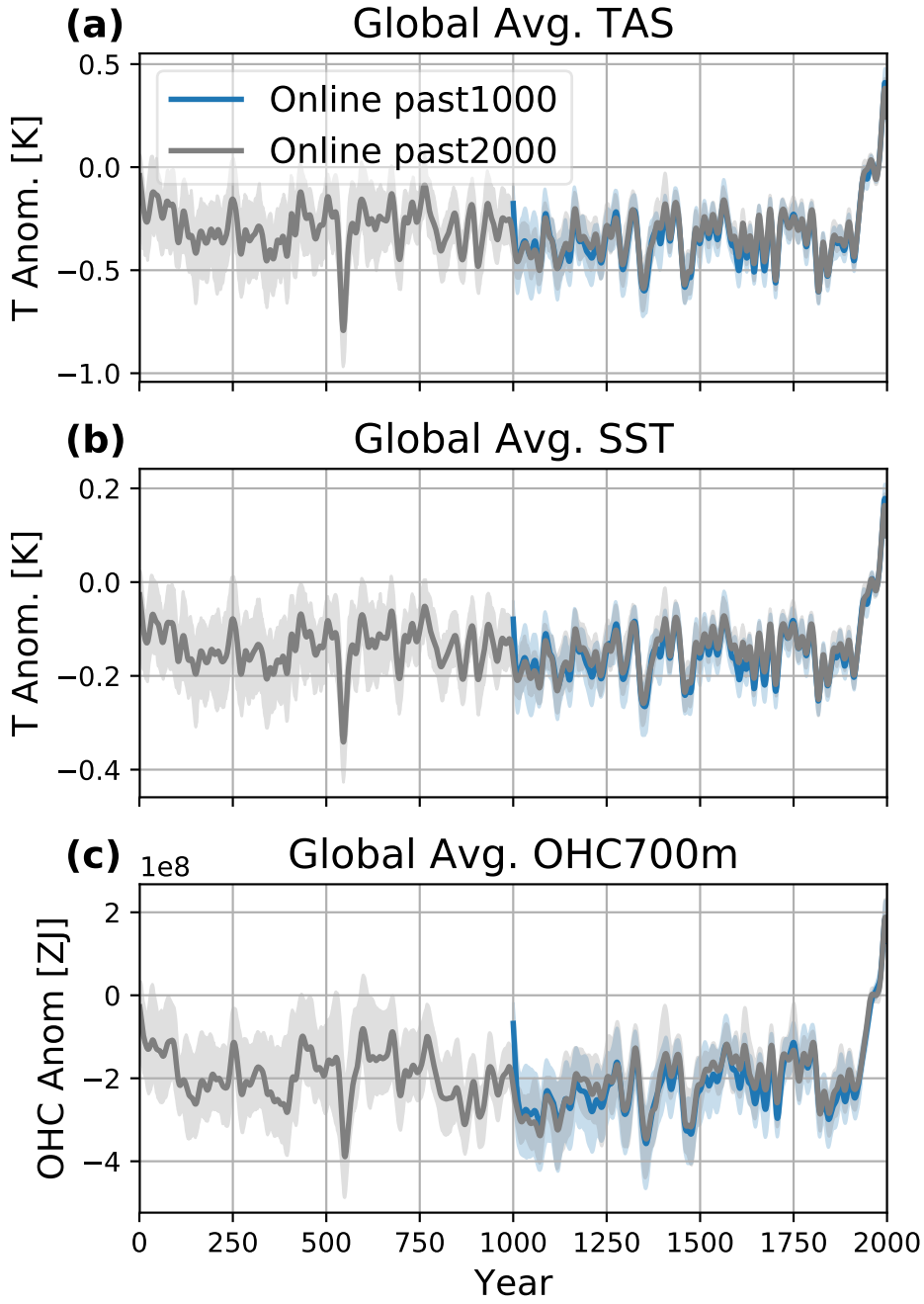

Figure S7. As in Fig. 7, but comparing online reconstructions intialized at 1000 C.E. (past1000) and 1 C.E. (past2000).

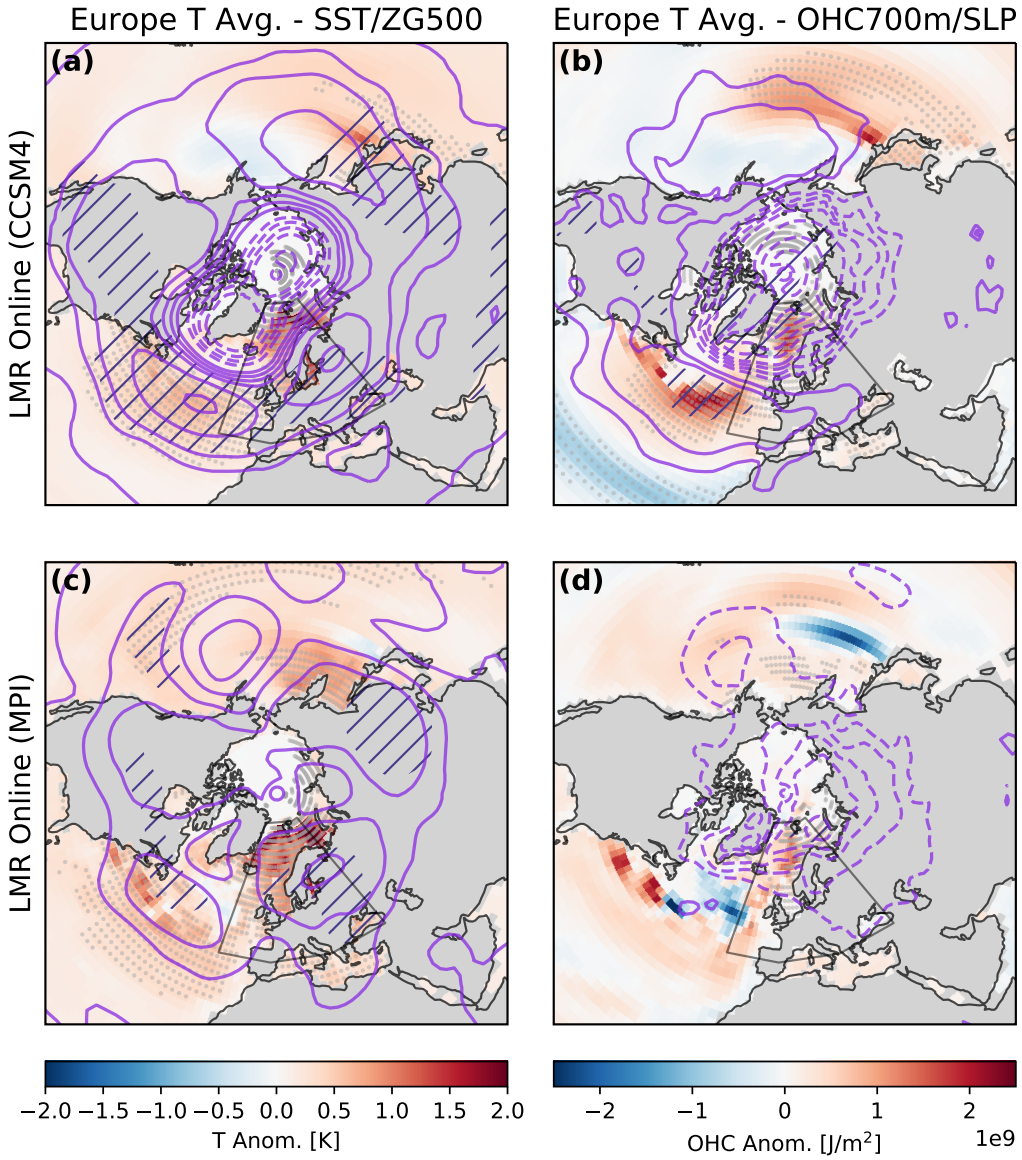

Figure S8. Regression of the reconstructed Europe average TAS (region denoted by black box) from 1000–1850 C.E. onto fields of SST/ZG500 (column a) and OHC700m/SLP (column b) for the CCSM4-LIM (row a) and MPI-LIM (row b) reconstructions. ZG500 field contour levels range from 7.5 m to 20 m incremented every 2.5 m for positive (solid) and negative (dashed) values. SLP field contour levels range from 0.25 hPa to 1.25 hPa incremented every 0.25 hPa. Regression coefficient significance (grey dots for SST/OHC700m, blue hatching for ZG500/SLP) determined using a two-tailed Student's t-test and the effective degrees of freedom (Bretherton et al., 1999).

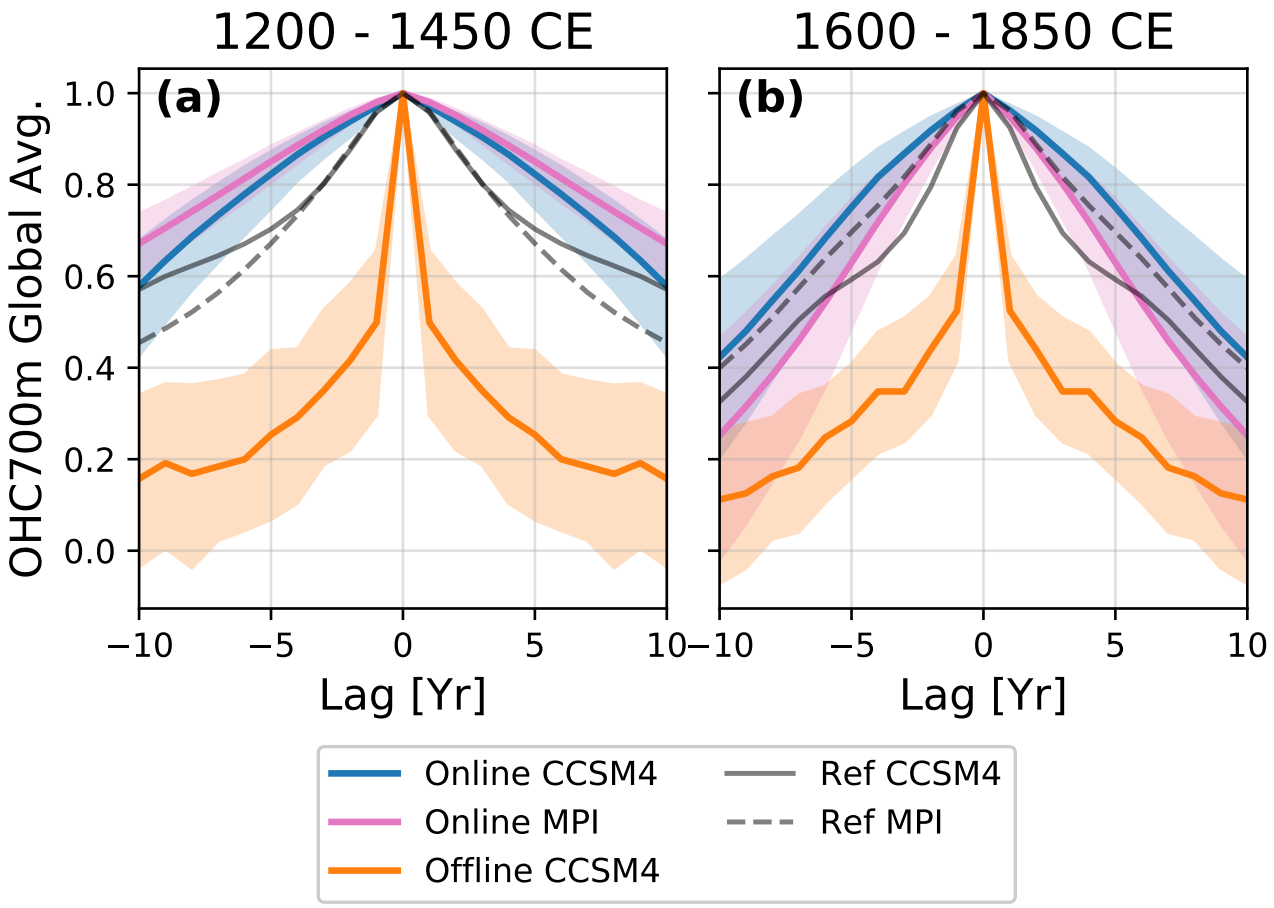

Figure S9. Autocorrelations of global average OHC700m for the time periods of 1200–1450 C.E. (a) and 1600–1850 C.E (b). Autocorrelations are shown for the Online CCSM4-LIM (blue) and MPI-LIM (pink) reconstructions, the CCSM4 Offline (orange) reconstruction, and reference Last Millennium Simulation data from the CCSM4 (black solid) and MPI (dashed black) models. Correlations are calculated at the specified lead/lag from the ensemble mean of each of the 50 reconstruction Monte-Carlo iterations. For reconstructions, solid lines denote the average correlation across Monte-Carlo iterations while shading shows the 95% confidence interval.

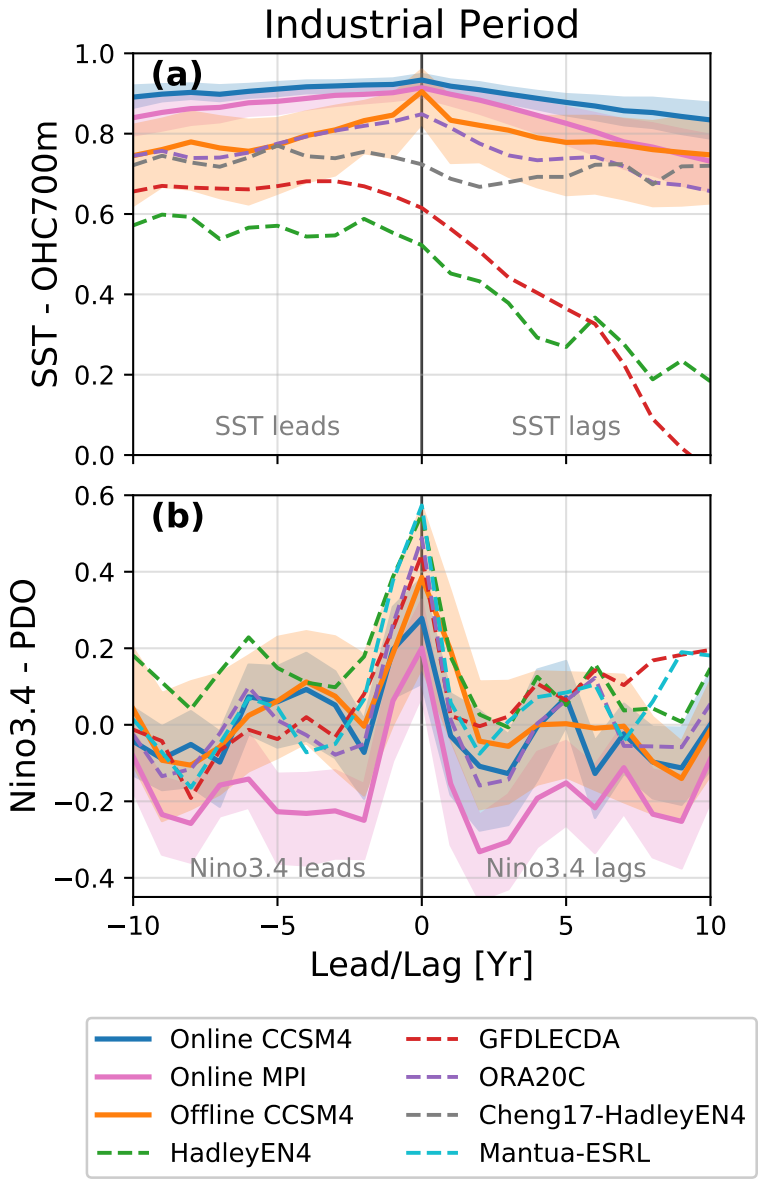

Figure S10. Lead–lag correlations for scalar indices of global average SST and OHC700m (a) and Niño 3.4 and PDO (c) during the instrumental period from 1900–2000 C.E. Correlations are shown for the Online CCSM4-LIM (blue) and MPI-LIM (pink) reconstructions, the CCSM4 Offline (orange) reconstruction, and instrumental products (dashed). Correlations are calculated at the specified lead/lag from the ensemble mean of each of the 50 reconstruction Monte-Carlo iterations. For reconstructions, solid lines denote the average correlation across Monte-Carlo iterations while shading shows the 95% confidence interval.

Table S3. As in Table S2, but for LIMs calibrated on MPI last millennium data. The global average (Glob Avg) measure includes ECRs from TAS, SST, OHC700m. RSUT and RLUT were omitted due to large ECR values dominating the selection process and the focus on atmosphere–ocean data in the present study. The ENSO measure includes ECRs from Niño 3, 3.4, and 4 indices and the Southern Oscillation Index (SOI). The PDO measure only includes the PDO index ECR. The Total column shows the sum of Glob Avg, ENSO, and PDO aggregate ECR distances.

num mvarEOFs	Glob Avg	ENSO	PDO	**Total**
15	1.07	0.12	0.51	1.69
20	3.14	0.13	0.18	3.45
21	1.96	0.35	0.09	2.40
22	1.47	0.11	0.18	1.76
23	1.95	0.02	0.53	2.50
24	0.25	0.43	0.75	1.42
25	2.24	0.14	0.26	2.64
26	1.49	0.15	0.11	1.75
27	1.03	0.20	0.02	1.25
28	1.56	0.05	0.82	2.43
29	1.94	0.28	0.12	2.34
30	5.22	0.84	0.18	6.24