Supporting Online Material

This file includes:
Materials and Methods
SOM Text
Figures S1-S7
Tables S1-S3

Materials and Methods

Lowpass filtering
Multidecadal lowpass filtering was performed using the method of Mann (2008) (*53*), which determines optimal boundary constraints based on the combination of boundary constraints (minimization of norm, slope, or roughness at edges) that minimizes the mean-square-error with respect to the raw series.

Estimates of the forced component
The CMIP5-All ensemble mean North Atlantic (NA), North Pacific (NP), and Northern Hemisphere (NH) series (shown in Fig. 1) for the estimation of observed AMO, PMO and NMO (i.e., results in Fig. 3) were determined by 1) calculating a mean series for each model realization for the respective target regions, 2) mean centering the series for each realization, 3) averaging the realizations for each model to produce a model mean, then 4) averaging the model means at each time step. This ensured that each model was equally represented in the ensemble mean and that models with a large number of realizations (e.g. GISSE2-R/H; Table S1) did not have a relatively larger influence on the CMIP5-All temperature series. The GISS-E2-R mean series (24 realizations) were determined by averaging the target region means of all realizations at each time step. The CMIP5-All ensemble mean for the estimation of AMO, PMO, and NMO in each model realization (i.e., results in Figs. 2, S2–S4) was determined by averaging all realizations. We extended the CMIP5 mean series (which spans AD 1850–2005) to 2012 using the slope of the 30 year trend (AD 1975–2005). This reflects the assumption of statistical persistence of the multidecadal timescale variations during the most recent decade. All CMIP5 model simulation data were regridded at 5° spatial resolution prior to analysis.

In the NMO analyses, to allow for direct comparison with instrumental NH mean series, which are based on Surface Air Temperature (SAT) over land and Sea Surface Temperature (SST) over ocean regions, we calculated the mean (latitude weighted) SAT over land by masking ocean grid cells, calculated the mean (latitude weighted) SST over the ocean, and combined the two series using a weighted average based on a land coverage value of 39% and an ocean coverage value of 61% for the northern hemisphere (Figs. 1,S1).

Regression Method
To calculate the AMO, PMO, and NMO we 1) regressed the observed mean temperature series onto the model derived estimate of the forced component, 2) estimated the forced

component of observed variability using the linear model from step 1, then 3) subtracted the forced component from the observations to isolate the internal variability component.

Uncertainty
To estimate confidence limits for the CMIP5-All AMO, PMO, and NMO, we repeated the target region regression analysis using 1000 surrogates (produced using bootstrap resampling) of the North Atlantic, North Pacific, and Northern Hemisphere model mean temperature series, and used the 2-sigma range of the resulting AMO, PMO, and NMO surrogates to define the uncertainty range.

Assessment of the internal variability in each model realization
To assess the statistical independence of internal variability in the model realizations, we used the known formulation for the standard deviation of the mean of a sample of *N* independent, identically distributed series. The n^{th} value of the time series for any individual surrogate is assumed to be a single realization of a larger family of possible values resulting from the process producing the internal variability. Under that assumption, the standard deviation of the mean over a number of *N* realizations (i.e., "samples") can be defined as the single sample standard deviation "*s*" (which can be estimated from the actual series themselves) divided by the square root of the number of samples (i.e., number of realizations used, *N*). We show that this successfully predicts the decrease in variance when truly independent realizations (as estimated by the regional regression method) are used, but that the variance is substantially greater than this value when the series are not independent (i.e., in the case of the detrending method) because the various "surrogates" actually contain a common forced signal.

SOM Text

Definition of the AMO, PMO, and NMO
The observed low-frequency internal variability may represent the sum of true oscillatory (i.e., 50–70 year) signals and simple low-frequency red noise. For our purposes the distinction is unimportant, and we use AMO, PMO, and NMO in a loose sense to denote the multidecadal (>40 year timescale) internal variability in the respective series, regardless of whether it constitutes a true oscillation or not.

Summary of past work on the AMO
Mann and Emanuel (*47*) showed that what is commonly termed the "AMO" (e.g., *17–19*) may substantially be an artifact of the misidentification of forced trends as internal variability. Focusing on tropical Atlantic SST, they showed that a commonly employed method of estimating the AMO—simple linear detrending followed by low-frequency filtering of the residual series—results in an artificial, apparent low-frequency "oscillation", since the true forced signal is not linear in time. The cooling from the 1950s–1970s AD associated with a substantial increase in anthropogenic aerosols in the Northern Hemisphere and a subsequent warming due to the decrease in aerosols commencing in the late 1970s, masquerades as a low-frequency "oscillation". Other more recent work has supported that finding (*54–58*).

Ting et al. (*22*) analyzed Coupled Model Intercomparison Project Phase 3 (CMIP3) simulations, finding, as did Mann and Emanuel (*48*), that the simple detrending approach fails to correctly isolate the AMO signal. They (as well as Trenberth and Shea (*23*)) argued for an alternative method where the forced trend is estimated based on regression of North Atlantic SST against global mean SST, and removed to yield an estimate of the internal variability component. Mann and Emanuel (*48*) had considered that approach, but found it does not account for the full impact of anthropogenic aerosol cooling, which has a greater relative influence on the North Atlantic than is captured by its global mean SST projection. Hence, they favored a bivariate regression using both global mean SST and the anthropogenic aerosol series to define the forced temperature response.

Knight (*24*) and Terray (*25*) combined climate model simulations with observational data to estimate the internal AMO variability. The forced component of the AMO was defined as the mean of North Atlantic SST in an ensemble of simulations from a modest number of models (11 and 12 for Knight (*24*) and Terray (*25*), respectively). The AMO was defined as the difference between the observed SST series and the multi-model mean (in which modeled internal variability components canceled leaving only the forced component of variability). Knight (*24*) argued that the AMO calculated in this way retained properties comparable to the detrended AMO, with similar phasing but with a smaller overall amplitude.

Mann et al. (*42*) used a semi-empirical method combining climate model simulations with observational data to estimate the component of forced Northern Hemisphere mean temperature change. The internal variability component was estimated as the difference between actual NH mean temperatures and the estimated forced component. The AMO projection onto NH mean temperature was defined by smoothing (i.e., low-pass filtering) this series at a multidecadal (>40 year) timescale. With a peak NH mean amplitude of ~0.1^{o}C, the series was found to have similar characteristics to model-simulated AMO variability (*3*). Mann et al. (*42*) noted that a decreasing trend in the thusly-defined AMO has contributed to the recent "slowdown" in warming. In their analysis, the AMO was essentially equated with the NMO, assuming little additional role of Pacific multidecadal variability (PMO) at the hemispheric scale.

Supplemental results and discussion

We applied each of the four methods to each of the target regions for all model realizations, yielding an ensemble of AMO, PMO, and NMO series. If the method in question were performing properly, the estimated internal variability component should be statistically independent for each realization of each model. Averaging them should thus lead to a reduction in amplitude of roughly $s/\sqrt{(N-1)}$ where s is the sample standard deviation (which we can take to be the mean standard deviation across N-1 realizations) and N is the number of ensemble members. The results of all three analyses, CMIP5-All (Figs. 2,S2), CMIP5-GISS (Fig. S3), CMIP5-AIE (Fig. S4) show that this *only* holds for our regional regression method (Table S2). For both the detrending and global regression methods, the amplitude of the mean series considerably exceeds the bounds that would be expected for an ensemble of independent realizations of internal variability. Moreover, the residual structure clearly indicates that forced variability is leaking into the putative

estimates of internal variability. Especially noteworthy are the systematic, sharp apparent increases in the inferred AMO/PMO/NMO toward the end of the series and the systematic positive peaks in the 1930/1940s. These artifacts are particularly pronounced for the detrending method, but are also present for the global SST regression method. The regional differencing approach (*24*) also appears to perform properly upon initial inspection, but that depends strictly on the model-estimated amplitude of forced variability being correct, which is insured in this case due only to the self-consistent experimental design. If the mean $2xCO_2$ equilibrium climate sensitivity (ECS) of the model ensemble were, for example, substantially different from the real-world ECS, that will no longer hold. We test this scenario by repeating the analysis but rescaling the ensemble mean series by a factor of 0.78 to simulate the plausible case where the real-world ECS is 2.5°C while the mean ECS of the model simulations used to define the forced series is a higher value of 3.2°C (the actual CMIP5 mean ECS). The results obtained via our regional regression method are unaltered in this case, but the results using regional differencing are impacted dramatically, with the method performing even *worse* than the global regression approach. We conclude that of the four methods considered for defining the AMO, PMO, and NMO, only our regional regression method performs adequately.

The AMO, PMO, and NMO amplitudes are seen to be unusually large with the detrending approach (Fig. S5A). Particularly striking are the very large positive trends in the AMO and NMO at the end of the series, which were indeed predicted (Figs. 2,S2–S4) as structural artifacts of the method. The root mean square (RMS) amplitude of the NMO is 0.14°C, more than twice the simulated amplitude of the hemispheric multidecadal variability from Knight et al. (*3*). The AMO and PMO have estimated amplitudes of 0.15°C and 0.09°C, respectively, and show high levels of apparent correlation with each other (R^2=0.563, lag = 0, statistically significant at p=0.05 level for a one-sided test—see next section for details about the associated calculation). The AMO, PMO, and NMO collectively give the appearance of a "stadium wave" pattern (*18*,*19*), wherein each varies coherently but at variable relative lag.

Our regional regression approach yields AMO, PMO, and NMO series that are dramatically different from those obtained with the detrending approach. Absent now are the very large positive trends in the AMO and NMO near the end of the series. The amplitude of the NMO (0.07°C using CMIP5-All) is half that inferred from the detrending approach. Unlike with the detrending approach, the maximum lagged correlation between the AMO and PMO (R^2=0.334 lag = 3) is no longer statistically significant.

The two other alternative approaches (global SST regression and regional differencing) yield somewhat lesser, but still non-trivial, biases in comparison with the detrending approach. Using instead the global SST regression approach (Fig. S5), the estimated AMO series is seen to have a modestly inflated amplitude with a larger positive recent AMO peak, features once again consistent with the structural artifacts that were predicted (Figs. 2,S2–S4) for the method. Using the regional differencing approach, we also see the artifacts predicted for that method, most notably a rather acute sensitivity of the results to

the precise estimate of the forced series, with the AMO series currently declining using CMIP5-GISS but peaking using CMIP5-AIE and CMIP5-All (Fig. S5). Anomalous negative spikes in the PMO at the end of the series too are likely artifacts of that method.

Assessment of AMO-PMO correlation

To assess the significance of lagged correlations between the AMO and PMO determined using the detrending and target region regression methods, we produced 17,000 uncorrelated AR(1) surrogates (100 for each model realization) based on the lag-1 autocorrelation of the actual North Atlantic and North Pacific series and computed the AMO and PMO for each using both the detrending and target region regression methods. We then determined the maximum lagged correlation for each surrogate using a window of up to ± 20 years (to allow for the variable lag between series that is intrinsic to the "stadium wave" hypothesis) and compared the maximum lagged correlation between the observed AMO and PMO for each method with this null distribution of correlations (Fig. S7A). A one-sided hypothesis test is employed since we require a positive lagged correlation. The null distributions for both methods are centered on a correlation value of r~0.4. The observed 'detrending' method correlation r=0.75 (corresponding to the R^2=0.563 value discussed above), is highly inconsistent with the null distribution (p=0.034 for a one-sided test is statistically significant employing a critical test value of α=0.05). The observed 'target regression' correlation r=0.58 (corresponding to the R^2=0.334 value discussed above) by contrast is consistent with the null distribution (p=0.114 for a one-sided test is *not* statistically significant employing a critical test value of α=0.05). In other words, using the target regression approach there is no evidence of a statistically significant linkage between multidecadal internal SST variability centered in the Atlantic and Pacific basins, in contrast with the detrending approach.

Using the actual AMO and PMO series computed for the CMIP5-All simulations (170 realizations) we again calculated the predicted distribution of maximum lagged correlation values based on the two methods and compared against the observed values. This comparison (Fig. S7B) demonstrates that 1) the observed correlation values for each method are consistent (p=0.382 for target regression and p=0.429 for detrending, i.e. $p>>0.05$ in both cases) with the corresponding distributions predicted from the CMIP5-All ensemble and 2) the detrending method produces a distribution that is substantially positively skewed, biased toward large apparent levels of positive AMO/PMO correlation.

These results, in summary, indicate that the detrending method does indeed produce artificially high (and statistically significant) apparent lagged correlations between the AMO and PMO and the illusion of a "stadium wave" pattern of coherently linked SST variability in distinct basins.

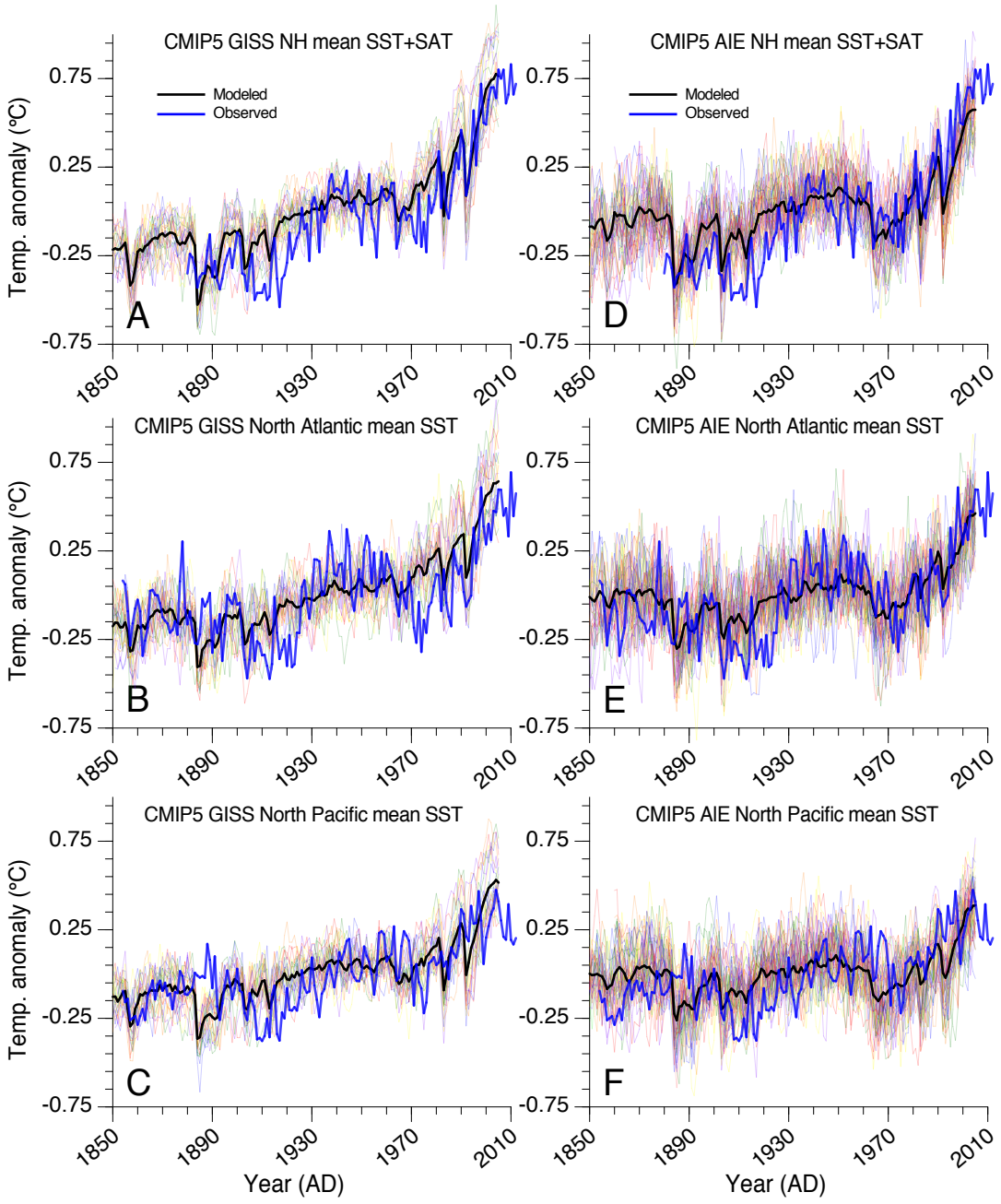

Fig. S1. CMIP5-AIE (aerosol indirect effects) and GISS-E2-R ensemble mean of Northern Hemisphere SST+SAT, North Atlantic SST, and North Pacific SST (black curves) shown with individual realizations (colored curves). Blue line depicts observed temperatures.

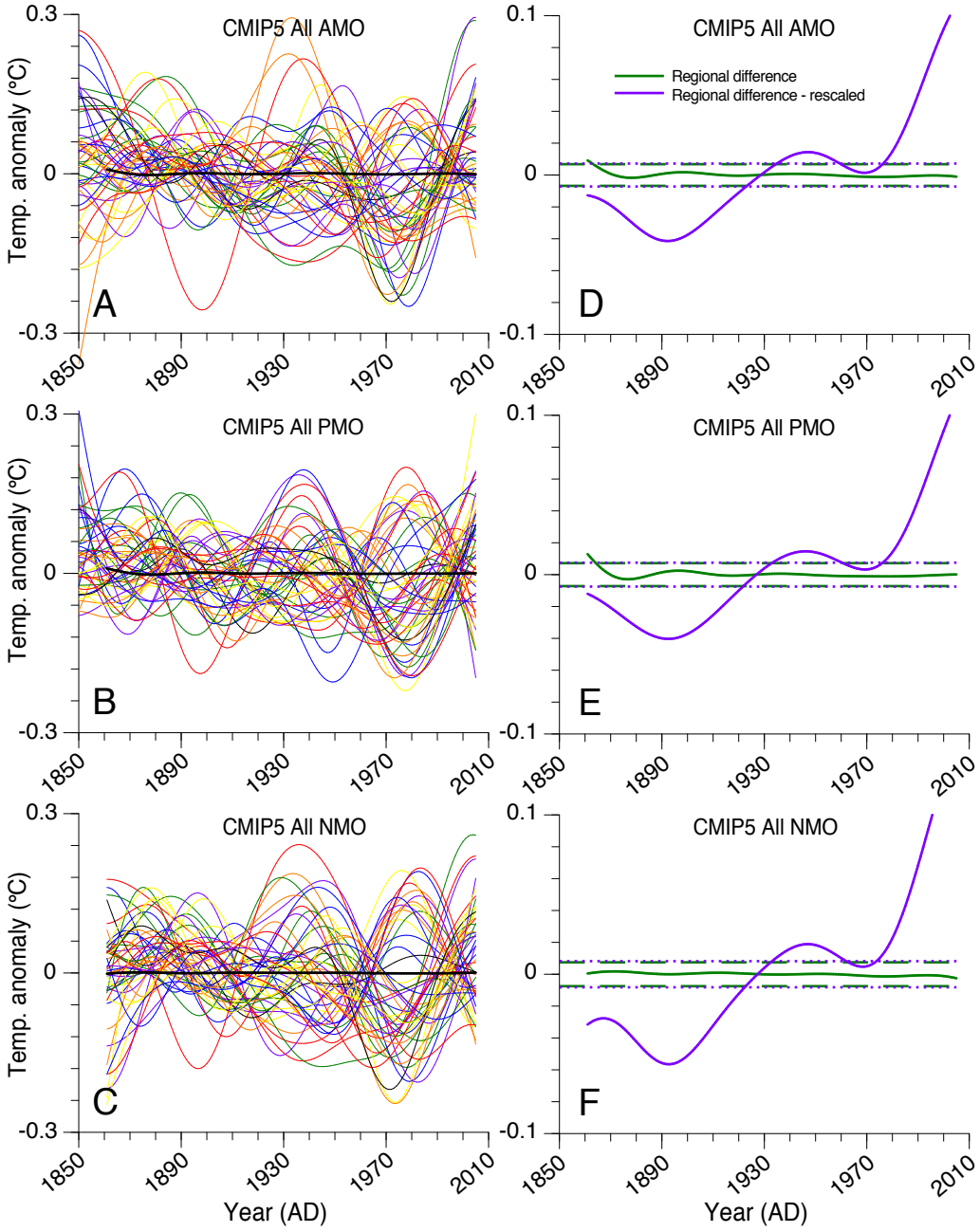

Fig. S2. **(A-C)** CMIP5-All mean (N-1) (black lines) and 24 individual realizations (colored lines) of AMO, PMO and NMO determined using target region regression. **(D-F)** Mean series (i.e., the mean of N-1 realizations; solid lines) and estimated 1-sigma bounds (dashed lines) for mean series under the assumption of statistical independence of internal variability among ensemble members determined using target region differencing (green), and rescaled target region differencing (purple).

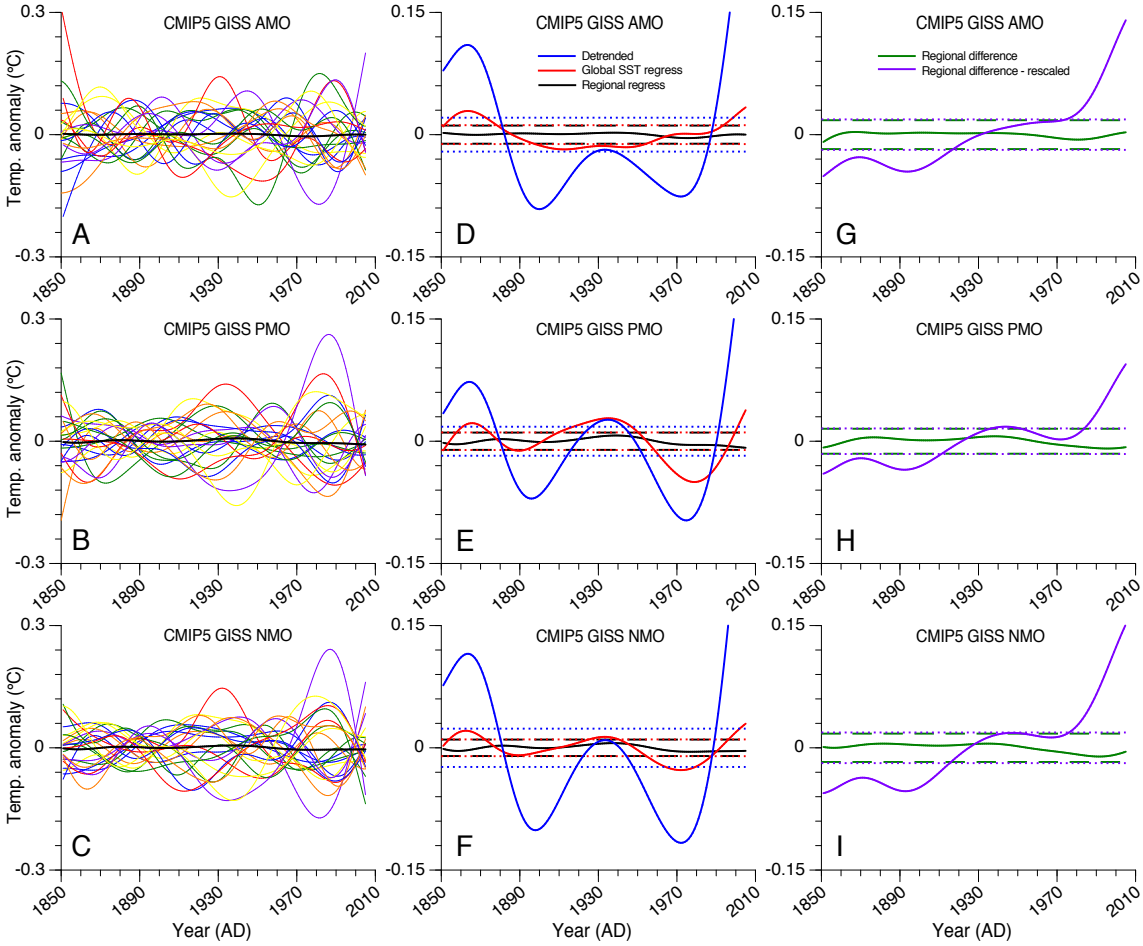

Fig. S3. **(A-C)** CMIP5-GISS mean (N-1) (black lines) and 24 individual realizations (colored lines) of AMO, PMO and NMO determined using target region regression. **(D-I)** Mean series (i.e., the mean of N-1 realizations; solid lines) and estimated 1-sigma bounds (dashed lines) for mean series under the assumption of statistical independence of internal variability among ensemble members determined using detrending (blue) global SST regression (red) target region regression (black), target region differencing (green), and rescaled target region differencing (purple).

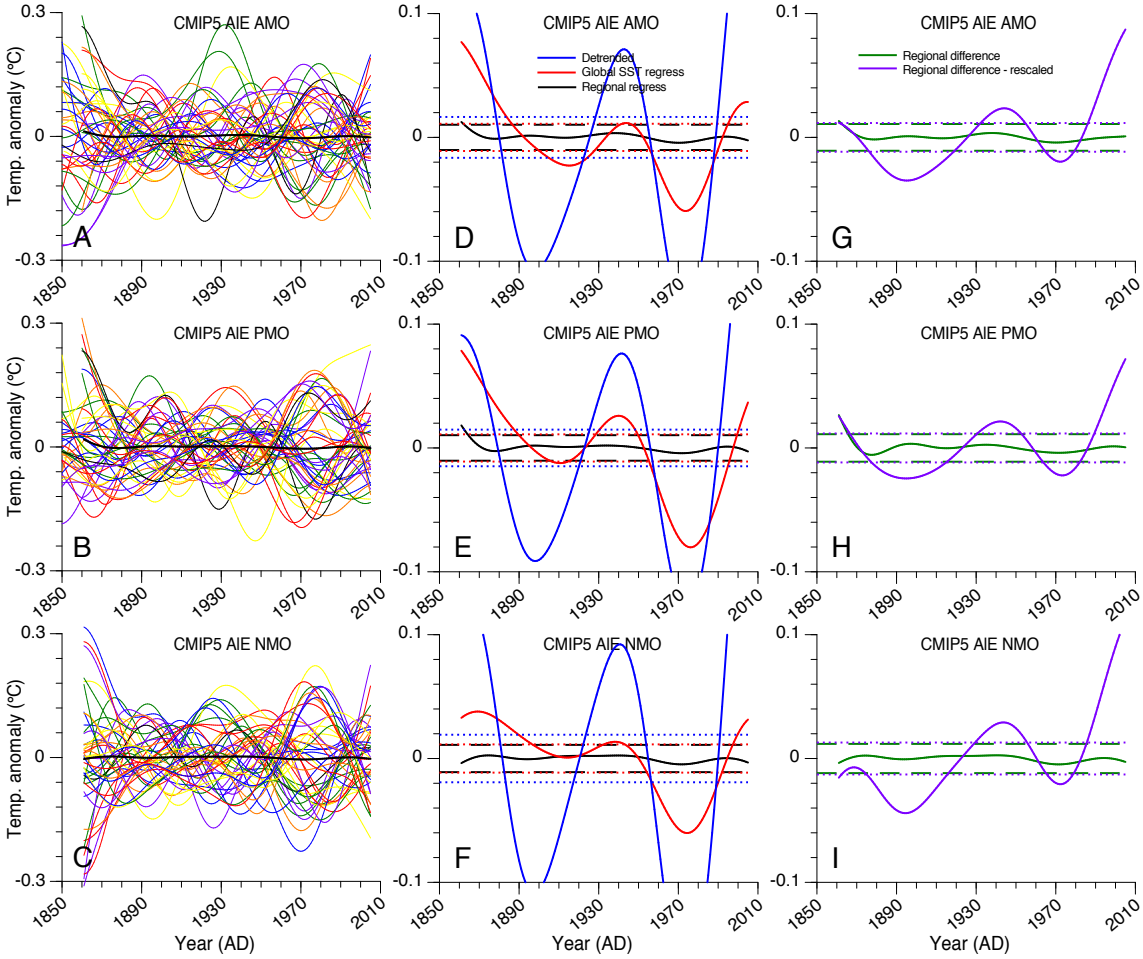

Fig. S4. **(A-C)** CMIP5-AIE mean (N-1) (black lines) and individual realizations (colored lines) of AMO, PMO and NMO determined using target region regression. **(D-I)** Mean series (i.e., the mean of N-1 realizations; solid lines) and estimated 1-sigma bounds (dashed lines) for mean series under the assumption of statistical independence of internal variability among ensemble members determined using detrending (blue) global SST regression (red) target region regression (black), target region differencing (green), and rescaled target region differencing (purple).

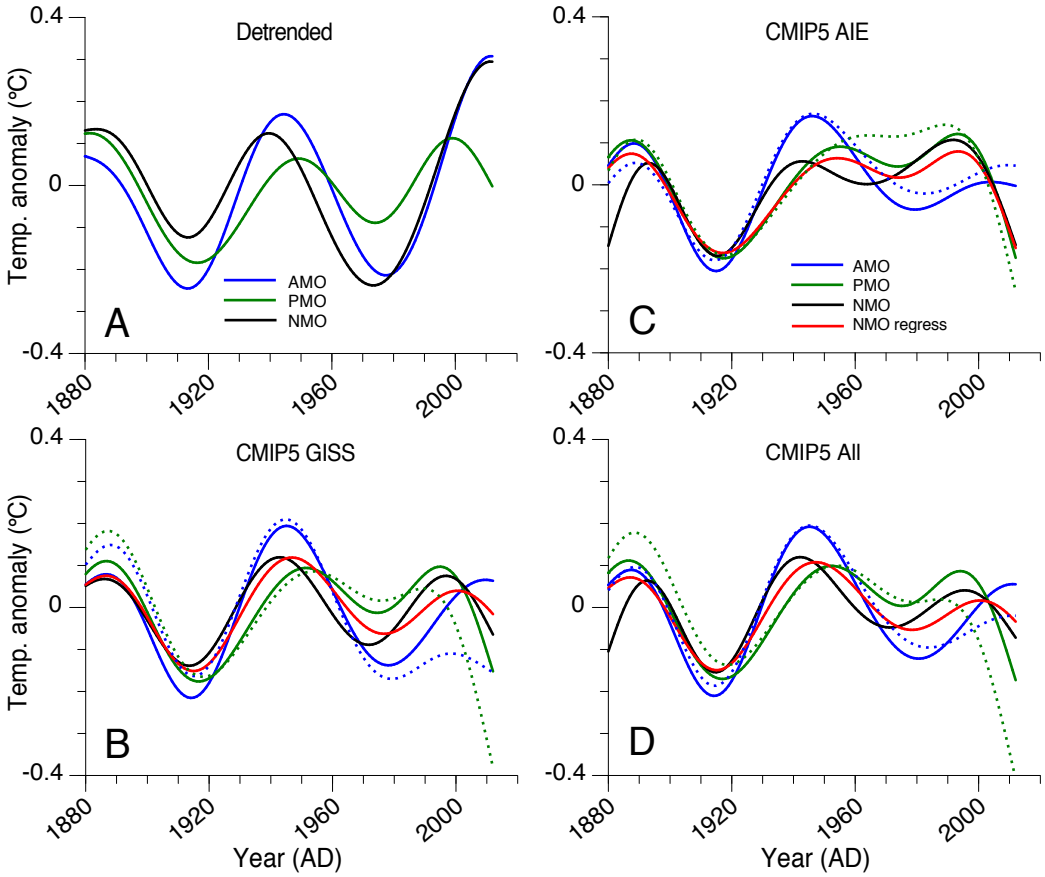

Fig. S5. Semi-empirical estimate of AMO (blue), PMO (green), and NMO (black) based on detrending **(A)**, global SST regression, and regional differencing using CMIP5-GISS **(B)** CMIP5-AIE **(C)** and CMIP5-All **(D)** historical climate model realizations. Solid lines are based on regression with global mean SST. Dashed curves are based on target region differencing (observed – model mean). Bivariate regression-based approximation of NMO (red).

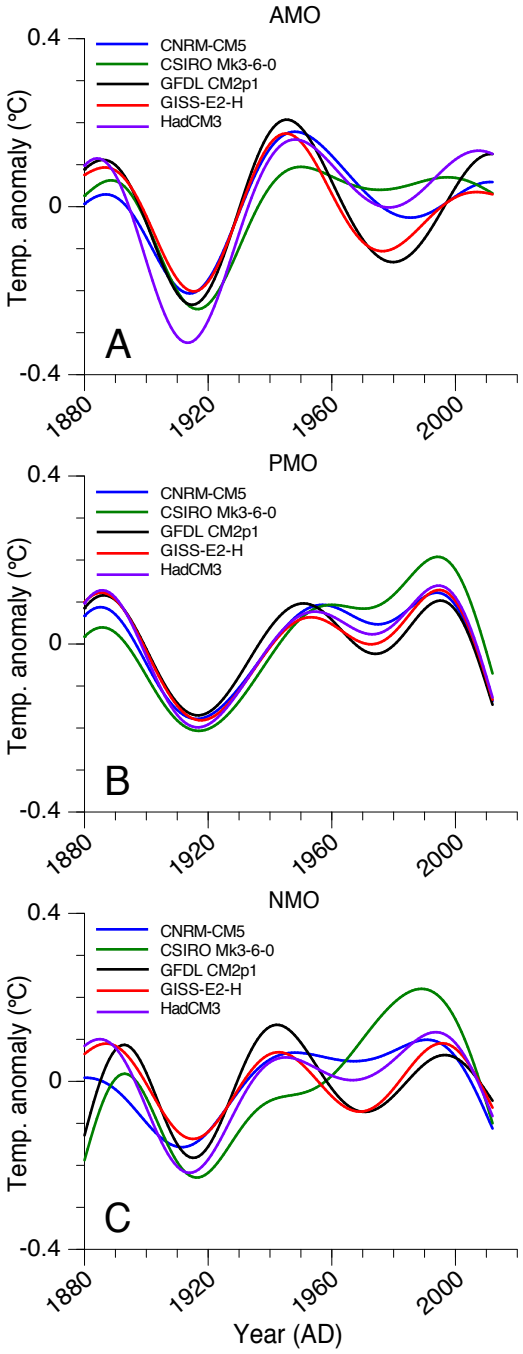

Fig. S6. Semi-empirical estimate of AMO **(A)**, PMO **(B)**, and NMO **(C)** based on target region regression using mean series from CMIP5 models with ten or more realizations.

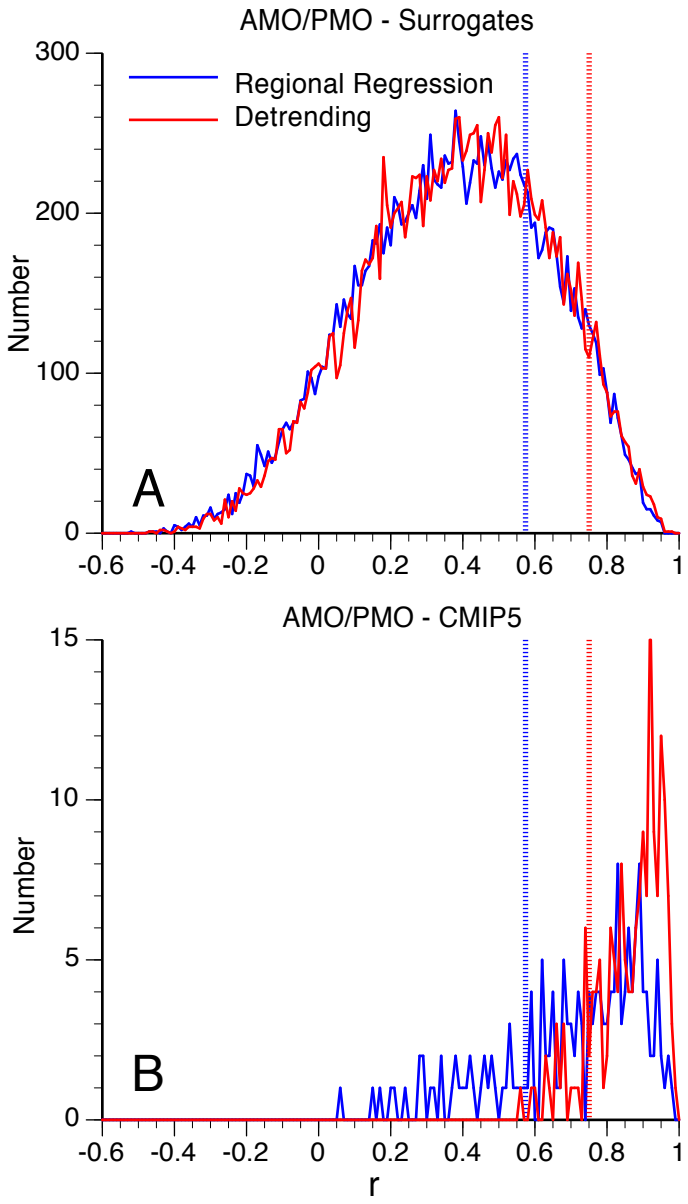

Fig. S7. **(A)** Null distributions of maximum lagged AMO-PMO correlations (*r* values) in an array of randomly generated, AR(1) surrogates (100 for each realization) produced using the actual CMIP5 North Atlantic and North Pacific series (170 total realizations) shown with the observed maximum AMO-PMO correlation (vertical dashed lines). **(B)** Distribution of AMO-PMO maximum correlations in the actual CMIP5 ensemble shown with the observed AMO-PMO correlations.

Table S1. CMIP5 ensemble.

Model	Number of Realizations	Length of historical runs (yr)	Start year AD	End Year AD	1st *and* 2nd aerosol indirect effects
GISS-E2-R	24	156	1850	2005	N
GISS-E2-H	17	156	1850	2005	N
CNRM-CM5	10	156	1850	2005	N
CSIRO-Mk3.6.0	10	156	1850	2005	Y
GFDL-CM2.1	10	145	1861	2005	N
HadCM3	10	146	1860	2005	N
CCSM4	6	156	1850	2005	N
IPSL-CM5A-LR	6	156	1850	2005	N
CanESM2	5	156	1850	2005	N
GFDL-CM3*	5	146	1860	2005	Y
HadGEM2-ES	5	146	1860	2005	Y
MIROC5	4	163	1850	2012	Y
MRI-CGCM3	4	156	1850	2005	Y
ACCESS1.3	3	156	1850	2005	Y
bcc-csm1-1	3	163	1850	2012	N
bcc-csm1-1m	3	163	1850	2012	N
CESM1-CAM5	3	156	1850	2005	Y
CESM1-FASTCHEM	3	156	1850	2005	N
FIO-ESM	3	156	1850	2005	N
IPSL-CM5A-MR	3	156	1850	2005	N
MPI-ESM-MR**	3	156	1850	2005	N
MIROC-ESM	3	156	1850	2005	Y
MPI-ESM-LR*	3	156	1850	2005	N
NorESM1-M	3	156	1850	2005	Y
MPI-ESM-P**	2	156	1850	2005	N
CESM1-WACCM	1	156	1850	2005	N
HadGEM2-CC	1	146	1860	2005	Y
HadGEM2-AO**	1	146	1860	2005	Y
ACCESS1.0	1	156	1850	2005	Y
BNU-ESM	1	156	1850	2005	N
CESM1-BGC	1	156	1850	2005	N
CMCC-CESM	1	156	1850	2005	N
CMCC-CM	1	156	1850	2005	N
CMCC-CMS	1	156	1850	2005	N
CNRM-CM5-2	1	156	1850	2005	N
GFDL-ESM2G	1	145	1861	2005	N
GFDL-ESM2M	1	145	1861	2005	N
GISS-E2-H-CC	1	161	1850	2010	N
GISS-E2-R-CC	1	161	1850	2010	N
INM-CM4	1	156	1850	2005	N
IPSL-CM5B-LR	1	156	1850	2005	N
MRI-ESM1	1	155	1851	2005	Y
FGOALS-g2**	1	156	1850	2005	Y
NorESM1-ME	1	156	1850	2005	Y

*One realization from this model was not included in the NMO experiments.

** This model was not included in the NMO experiments.

Table S2. Observed and predicted standard deviations of the mean AMO, PMO, and NMO series for all realizations of CMIP5 and GISS-E2-R.

			Detrended	Global SST regress	Target SST difference	Target SST difference: rescaled	Target SST regress
AMO	CMIP5-All	Act.	0.0850	0.0183	0.0016	0.0350	0.0015
		Pred.	0.0080	0.0051	0.0067	0.0071	0.0052
	GISS-E2-R	Act.	0.0857	0.0154	0.0030	0.0429	0.0021
		Pred.	0.0208	0.0116	0.0180	0.0187	0.0112
	CMIP5-AIE	Act.	0.0943	0.0316	0.0029	0.0277	0.0027
		Pred.	0.0166	0.0110	0.0108	0.0115	0.0101
PMO	CMIP5-All	Act.	0.0822	0.0167	0.0022	0.0345	0.0016
		Pred.	0.0076	0.0049	0.0072	0.0075	0.0050
	GISS-E2-R	Act.	0.0697	0.0236	0.0050	0.0296	0.0040
		Pred.	0.0174	0.0107	0.0150	0.0152	0.0103
	CMIP5-AIE	Act.	0.0836	0.0391	0.0046	0.0221	0.0033
		Pred.	0.0149	0.0110	0.0111	0.0117	0.0103
NMO	CMIP5-All	Act.	0.1100	0.0188	0.0009	0.0477	0.0005
		Pred.	0.0098	0.0052	0.0075	0.0082	0.0053
	GISS-E2-R	Act.	0.0997	0.0142	0.0051	0.0483	0.0034
		Pred.	0.0228	0.0102	0.0171	0.0182	0.0100
	CMIP5-AIE	Act.	0.1123	0.0283	0.0022	0.0344	0.0022
		Pred.	0.0194	0.0114	0.0117	0.0129	0.0110

Table S3. Scaling factor ("beta") values for target region regression analysis.

	North Atlantic (scaling factor ± standard error)	North Pacific	Northern Hemisphere
CMIP5-GISS	0.7180 ± 0.0120	0.6202 ± 0.0181	0.9857 ± 0.0155
CMIP5-AIE	1.0913 ± 0.0185	0.7386 ± 0.0229	1.2773 ± 0.0187
CMIP5-All	0.9216 ± 0.0155	0.6286 ± 0.0182	1.0530 ± 0.0169