Evaluation of Leading Modes of Climate Variability in the CMIP Archives

JOHN T. FASULLO, A. S. PHILLIPS, AND C. DESER

National Center for Atmospheric Research, Boulder, Colorado

ABSTRACT

The adequate simulation of internal climate variability is key for our understanding of climate as it underpins efforts to attribute historical events, predict on seasonal and decadal time scales, and isolate the effects of climate change. Here the skill of models in reproducing observed modes of climate variability is assessed, both across and within the CMIP3, CMIP5, and CMIP6 archives, in order to document model capabilities, progress across ensembles, and persisting biases. A focus is given to the well-observed tropical and extratropical modes that exhibit small intrinsic variability relative to model structural uncertainty. These include El Niño–Southern Oscillation (ENSO), the Pacific decadal oscillation (PDO), the North Atlantic Oscillation (NAO), and the northern and southern annular modes (NAM and SAM). Significant improvements are identified in models' representation of many modes. Canonical biases, which involve both amplitudes and patterns, are generally reduced across model generations. For example, biases in ENSO-related equatorial Pacific sea surface temperature, which extend too far westward, and associated atmospheric teleconnections, which are too weak, are reduced. Stronger tropical expression of the PDO in successive CMIP generations has characterized their improvement, with some CMIP6 models generating patterns that lie within the range of observed estimates. For the NAO, NAM, and SAM, pattern correlations with observations are generally higher than for other modes and slight improvements are identified across successive model generations. For ENSO and PDO spectra and extratropical modes, changes are small compared to internal variability, precluding definitive statements regarding improvement.

1. Introduction

The adequate simulation of internal climate variability is vital for efforts involving historical attribution (Santer et al. 2009; Stott et al. 2010; Schurer et al. 2013; Imbers et al. 2014; Deser et al. 2016; Wallace et al. 2016; McKinnon and Deser 2018), seasonal and decadal prediction (Robertson et al. 2015; Thoma et al. 2015; Meehl et al. 2016; Vitart et al. 2017; Simpson et al. 2019), and multidecadal projection (Deser et al. 2012a, 2014, 2017b; Kumar and Ganguly 2018; Dai and Bloecker 2019). Internal fluctuations have the ability to fully obscure or amplify the underlying climate-change signal in many fields for decades (Trenberth and Fasullo 2013; Deser et al. 2014; Lehner et al. 2018; Guo et al. 2019) and correctly accounting for their influence is necessary to understand past changes and estimate both uncertainty and ensemble spread in predictions across a range of time scales, from seasonal to decadal and beyond.

Climate models have historically been deficient in their simulation of internal climate variability (Stoner et al. 2009; Lienert et al. 2011; Bellenger et al. 2014; Capotondi et al. 2015; McKinnon and Deser 2018), with a broad range of performance across models. High-frequency atmospheric modes in the extratropics include the North Atlantic Oscillation (NAO; Hurrell 1995; Hurrell and Deser 2009) and the southern and northern annular modes [SAM (Thompson and Wallace 2000) and NAM (Cattiaux and Cassou 2013; Gillett and Fyfe 2013)]. These modes exhibit a peak in their spectra on seasonal (Lee and Black 2015) time scales and model bias has been characterized primarily by amplitude and secondarily by pattern (Lee et al. 2019). In the tropics, the dominant mode of internal variability is El Niño–Southern Oscillation (ENSO; Trenberth 1997; Guilyardi et al. 2012), which involves strong couplings between the atmosphere and ocean and significant variability across a broad spectrum, spanning from a few years to over a decade (Vimont 2005; Deser et al. 2012b). In simulating ENSO, models have suffered from biases in their amplitude, patterns, and transient structure (e.g., Guilyardi et al. 2012; Bellenger et al. 2014). At lower frequencies and connected to ENSO, the Pacific decadal oscillation (PDO; Mantua and Hare 2002; Newman

Corresponding author: John T. Fasullo, fasullo@ucar.edu

et al. 2016) and the related interdecadal Pacific oscillation (IPO; Power et al. 1999) represent a particularly pronounced mode of variability, with teleconnections that emanate globally, well beyond the region used for its definition in the North Pacific Ocean. Historically, climate models have had difficulty in adequately simulating connections between the tropics and extratropics associated with the PDO, and with reproducing its observed magnitude (Oshima and Tanimoto 2009; Furtado et al. 2011; Newman et al. 2016). Additional sources of model error impacting their representation of internal variability are known to exist. These include biases arising from the limited resolution of global models (e.g., Bojariu and Giorgi 2005) and errors in the simulated base state, particularly in the tropics (Capotondi et al. 2015) where coupled feedbacks are influenced by biases in winds and ocean structure (Vannière et al. 2013) and clouds (Li and Xie 2012). The propagation of model bias across modes has also been identified (Yang et al. 2018).

The evaluation of climate models and their simulation of variability poses various challenges (Phillips et al. 2014; Eyring et al. 2016a; Gleckler et al. 2016). The observational record is sparse in time and space prior to the satellite era and errors in the record lead to nontrivial differences in observational estimates for many internal modes. The limited duration of some records also imparts an inherent uncertainty to identified patterns, transient features, and spectra, uncertainties that have yet to be broadly quantified in importance relative to structural climate model bias. A major uncertainty also exists regarding the role of external climate forcing in the interpretation of some modes, such as the Atlantic multidecadal oscillation (AMO, e.g., Otterå et al. 2010; Booth et al. 2012). Are model biases large relative to these uncertainties, as would be necessary for a meaningful evaluation of models, or are some modes inherently noisy and therefore difficult to evaluate? These issues are particularly pronounced for low-frequency global climate modes, which contain limited temporal degrees of freedom, and uncertainty is therefore large.

In recent years, as climate modeling groups have released simulations from their newest model versions as part of phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016b), the opportunity has arisen to take stock of the simulation of variability across model generations (Eyring et al. 2016c). In our study, the representation of selected leading modes of internal variability in CMIP6 and its predecessors [CMIP3 and CMIP5, described in Meehl et al. (2007) and Taylor et al. (2012), respectively] is assessed by applying the Climate Variability Diagnostics Package (CVDP; Phillips et al. 2014) to over 500 model simulations

and numerous observational datasets. In particular, we use the complete set of historical simulations across all three generations of CMIP in conjunction with a 40-member initial-condition ensemble performed with Community Earth System Model version 1 (CESM-LE: Kay et al. 2015) and 40 members of the Max Planck Institute for Meteorology (MPI-M) Grand Ensemble (MPI-GE; Maher et al. 2019) to discriminate between structural versus sampling uncertainty in models' representation of these modes. In these respects, our study thus goes beyond previous model evaluation efforts (Stoner et al. 2009; Gillett and Fyfe 2013; Bellenger et al. 2014).

The modes of internal variability selected for consideration and their metrics are motivated and discussed in section 2. To focus our examination of specific modes, the magnitude of sampling variability and observational uncertainty in a range of mode metrics is compared against structural model spread in this section. An assessment of ENSO, including its large-scale teleconnections and space–time evolution in the tropical Pacific Ocean, is presented in section 3. In section 4 the teleconnections of the PDO are assessed, while in section 5 model skill in reproducing observed patterns of selected extratropical atmospheric modes (NAO, NAM, and SAM) is examined. In section 5, a focus is given to examining performance across the CMIP archives, a topic that is also touched on in earlier sections. A summary, discussion, and conclusions are presented in section 6.

2. Methods, data, and mode selection

a. The Climate Variability Diagnostics Package

The CVDP is an automated software package that computes a broad suite of modes of climate variability in the atmosphere, oceans, and cryosphere (sea ice and snow cover) based on techniques widely accepted by the climate science community (Phillips et al. 2014). The CVDP can be applied to any number of model simulations and observational datasets over any time period as specified by the user (see examples provided on the CVDP homepage: http://www.cesm.ucar.edu/working_groups/CVC/cvdp/ and the data repository: http://www.cesm.ucar.edu/working_groups/CVC/cvdp/data-repository.html). Output from the CVDP is provided in both graphical and NetCDF formats, facilitating additional comparisons and analyses such as those undertaken in this study. Here, we make use of CVDP output of historical simulations from the CMIP3, CMIP5, and CMIP6 archives in addition to the CESM-LE and MPI-GE, as well as observations based on the common period 1900–2018. In particular, we examine a subset of modes

of variability from the CVDP; all modes are defined in the CVDP methodology link (see, e.g., http://webext.cgd.ucar.edu/Multi-Case/CVDP_ex/cesm1.lens_1920-2018/methodology.html). These include ENSO, defined using a one standard deviation threshold of the detrended Niño-3.4 SST anomaly index in December (Trenberth and Hoar 1997; Deser et al. 2012b), and the PDO, based on methods outlined in Mantua et al. (1997). The NAO and NAM are defined based on the approach described in Hurrell and Deser (2009) while the SAM index is based on methods developed in Thompson and Wallace (2000), although in that work 850-hPa geopotential height was used rather than SLP to deal with orographic effects. In the CVDP, to avoid the data storage and processing demands of using pressure level data, SLP is used. The Pacific–North American (PNA) pattern is based on methods described in Barnston and Livezey (1987) while the AMO is defined using methods described in Trenberth and Shea (2006). These indices are then used to generate regression patterns with surface meteorological fields (see below), both globally and over regions used for mode definitions.

b. Observational datasets

The observations used in this work for mode identification and evaluation include sea surface temperature (SST), near-surface air temperature (TS), and sea level pressure (SLP). Multiple best-estimate datasets are considered so as to allow for estimation of observational uncertainty. These include SST estimates from the Extended Reconstructed Sea Surface Temperature (ERSSTv5; Huang et al. 2017) dataset and the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST; Rayner et al. 2003) dataset. For TS, the Berkeley Earth Surface Temperature (BEST; Rohde et al. 2013) and the Goddard Institute for Space Studies Surface Temperature (GISTEMP; Lenssen et al. 2019) datasets are used, which offer the advantage of infilling observational gaps. The NAO, NAM, and SAM are commonly defined using SLP data for which long records exist. Here we use SLP data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Twentieth Century Reanalysis (ERA20C; Poli et al. 2016), which are concatenated with SLP data from ERA-Interim after 1979 (Berrisford et al. 2009). Fields are also used from the coupled twentieth-century climate reanalysis (CERA-20C; Laloyaux et al. 2018), which assimilates only surface pressure and marine wind observations. Similar results in this work are obtained using the NOAA 20CR Reanalysis Product (Compo et al. 2011). For SST, SLP, and TS, the observational network is generally sufficient for sampling large-scale internal variability as the period 1920–2018 is used to
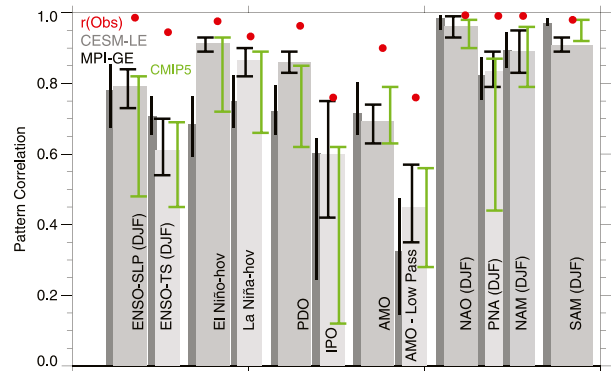

FIG. 1. Mean spatial and El Niño and La Niña Hovmöller (hov) pattern correlations (vertical bars) between simulations (1920–2005) and observations (1900–2018) for members of the CESM-LE along with their two standard deviation (2σ) range (black whiskers). The corresponding 2σ ranges are also shown for the MPI-GE (dark gray, with mean correlation; bar) and CMIP5 archive (green whiskers). Pattern correlations between observational estimates (1900–2018) are also shown (red dots). Instances in which the range of variability in the CESM-LE (i.e., noise N) is on par with the CMIP5 range, which comprises both model structural error S and internal variability N, suggests a weak S/N for mode evaluation.

document modes in observations using the CVDP, though sensitivity to the precise selection of timeframe is small, such as for example if data only from 1950 to 2018 are used.

c. Models and mode selection

Various challenges arise in evaluating internal modes in models. Structural model bias can pose a challenge due to "mode-swapping," whereby a biased and dissimilar leading mode obscures a lower-order mode that is nonetheless more relevant to the leading-order observed mode (Lee et al. 2019). Mode-swapping can occur with particular frequency in CMIP models for the PNA pattern (Chen et al. 2018; Lee et al. 2019). Other challenges also exist, such as in cases where sampling uncertainty (associated with the limited duration of the observational record; Deser et al. 2017a) in a measure of model skill is on par with the structural uncertainty across models, or in cases where disagreement among available observational datasets is large.

To briefly survey these issues in the CMIP archives, the ranges of pattern correlation scores estimated for modes in the CESM Large Ensemble and MPI Grand Ensemble (MPI-GE) are compared with the ranges across the CMIP5 multimodel ensemble in Fig. 1. Observational uncertainty in the modes is also indicated by the pattern correlations between observational datasets (Fig. 1, red dots). As only a single model is used to generate the CESM-LE and MPI-GE, the spread in scores within these ensembles is solely due to the

influence of internally generated variability, while for CMIP5 scores both model structural uncertainty and internal variability contribute to spread. Thus, in instances where the ranges of noise in the CESM-LE and MPI-GE are on par with the spread across models, constraints on model fidelity are not stringent. In this work, mode diagnostics used in model evaluation are chosen to maximize the ratio of structural model bias to sampling uncertainty. For example, for ENSO teleconnection patterns in surface temperature (ENSO-TS) the CMIP5 spread (S; range of green whiskers) is not significantly greater than the range of the CESM-LE (i.e., noise N; range of black whiskers). In contrast, the S/N for ENSO surface pressure (ENSO-SLP) teleconnections is considerably larger, and this index is therefore preferred as an index of ENSO teleconnections. Similarly, the S/N for the El Niño-SST Hovmöller is somewhat greater than for the La Niña-SST Hovmöller, a feature that may be rooted in ENSO's asymmetries and the inability of some models to simulate multiyear La Niña events (e.g., Okumura and Deser 2010). However, given the socioeconomic importance of the observed asymmetry in duration of El Niño versus La Niña, and the challenges faced by models in resolving it (DiNezio and Deser 2014; Zhang and Sun 2014), both measures of ENSO's spatiotemporal structure will be assessed. In the Pacific, the S/N intrinsic to the interdecadal Pacific oscillation (Meehl and Hu 2006) is low and observational uncertainty is high (as demonstrated by the low correlation between observational estimates) relative to the index of the PDO (Mantua et al. 1997). In the Atlantic, both raw and low-pass filtered measures of the AMO exhibit weak S/N and large observational uncertainty, and therefore neither index is considered in this analysis. For extratropical modes, such as the NAO, PNA, NAM, and SAM, model scores are high generally, with the exception of the PNA for some CMIP5 models, and observational uncertainty is small. In particular, the PNA will not be considered explicitly here as it is susceptible to mode swapping (Lee et al. 2019), whereby higher-order simulated modes are most similar to the leading mode observed in some models. The S/N for extratropical modes is also small generally and this will be a limiting factor in forming statements regarding model bias. An emphasis is therefore placed on evaluating these modes and the evolving structure of bias across model generations, rather than on their absolute fidelity.

Last, to differentiate the dominant spatial patterns of model bias, a covariance matrix based principal component (PC) analysis is used whereby the arrays of spatial pattern biases (longitude $\times$ latitude) across models is decomposed for its empirical orthogonal functions (EOFs), where the model biases are plotted as regressions against the normalized PC time series (and therefore have the same units as the raw fields). Both the first two EOFs and corresponding PC values, sorted by their values and averaged across terciles for each CMIP generation, are shown in summary plots, where terciles are labeled in ascending order based on their sorted PC values in the lower, middle, and upper thirds (terciles 1–3, respectively). In the PC analysis, two observational products are also included to provide context for model differences and their evolution across CMIP generations, as will be discussed further below. In the PC analysis, a single simulation from each modeling center for each CMIP generation is chosen (to avoid overweighting the biases of a particular model or modeling centers with multiple members). Where more than one model version is available, models not incorporating Earth system processes are chosen. The end result (Table 1) is a chosen set of 16 climate models for CMIP3, 20 models for CMIP5, and 20 models for CMIP6. The full temporal span of historical simulations available after 1900 is used for this analysis.

We have repeated the PC analysis of bias patterns using only the CESM-LE. In general, the magnitudes of the CESM-LE bias EOFs are found to be smaller than those of the analogous bias EOFs in the CMIP archives; however, in some regions and for some modes comparable magnitudes exist, suggesting a nonnegligible contribution of internally generated variability in computed patterns of bias in CMIP models.

3. ENSO teleconnections and space–time structure

As the dominant mode of tropical interannual climate variability, ENSO originates in the tropical Pacific Ocean and is associated with climate extremes and societal impacts worldwide (Trenberth 1997; McPhaden et al. 2006). ENSO involves feedbacks between various oceanic and atmospheric processes and accurately modeling these interactions with global coupled models has been a long-standing challenge (Wittenberg 2009; Bellenger et al. 2014). The main features of ENSO's large-scale teleconnections, expressed as composites of SLP during December through February (DJF), are shown for observations and models in Fig. 2. Observed features (Fig. 2a) include negative SLP anomalies in the eastern tropical Pacific that extend to high latitudes, and particularly in the Aleutian low in the North Pacific Ocean, and positive SLP anomalies in the western tropical Pacific Ocean that extend poleward and eastward in a horseshoe-shaped pattern. Negative SLP anomalies extend across much of the midlatitudes in both hemispheres while positive anomalies also exist in polar regions. Simulation biases can be summarized by

TABLE 1. Simulations considered in this study by CMIP archive and their mean overall, PDO, ENSO SLP, ENSO TAS, El Niño Hovmöller, La Niña Hovmöller, NAM, and SAM scores (per Fig. 9).

CMIP3 (n = 16)	CMIP5 (n = 19)	CMIP6 (n = 20)
BCCR-BCM2.0: 0.81, 0.83, 0.61, 0.54, 0.58, 0.78, 0.96, 0.95	ACCESS1.3: 0.85, 0.78, 0.77, 0.62, 0.84, 0.81, 0.92, 0.95	BCC_CSM2-MR: 0.90, 0.86, 0.77, 0.74, 0.93, 0.93, 0.95, 0.98
CCCma CGCM3.1: 0.79, 0.71, 0.66, 0.55, 0.78, 0.76, 0.86, 0.94	BCC_CSM1.1: 0.83, 0.84, 0.78, 0.63, 0.90, 0.85, 0.90, 0.93	CAMS-CSM1-0: 0.84, 0.81, 0.69, 0.57, 0.56, 0.63, 0.91, 0.92
CNRM-CM3: 0.78, 0.72, 0.71, 0.50, 0.83, 0.79, 0.85, 0.91	BNU-ESM: 0.85, 0.86, 0.81, 0.65, 0.93, 0.92, 0.90, 0.83	CanESM5: 0.85, 0.75, 0.81, 0.71, 0.88, 0.81, 0.93, 0.95
CSIRO-Mk3.5: 0.84, 0.64, 0.62, 0.56, 0.88, 0.81, 0.96, 0.96	CanESM2: 0.83, 0.74, 0.72, 0.58, 0.85, 0.81, 0.87, 0.84	CESM2: 0.85, 0.79, 0.82, 0.75, 0.78, 0.84, 0.90, 0.95
GFDL CM2.1: 0.67, 0.42, 0.15, 0.17, 0.68, 0.67, 0.86, 0.95	CESM-BGC: 0.84, 0.75, 0.80, 0.59, 0.82, 0.88, 0.91, 0.97	CNRM-CM6-1: 0.86, 0.87, 0.82, 0.56, 0.93, 0.94, 0.88, 0.91
GISS Model E-H: 0.73, 0.41, 0.30, 0.28, 0.73, 0.68, 0.88, 0.97	CMCC-CM: 0.79, 0.68, 0.74, 0.59, 0.73, 0.69, 0.79, 0.95	E3SM-1.0: 0.85, 0.84, 0.83, 0.67, 0.86, 0.84, 0.94, 0.93
IAP-FGOALS1.0-g: 0.83, 0.83, 0.69, 0.62, 0.87, 0.83, 0.90, 0.90	CNRM-CM5: 0.83, 0.71, 0.61, 0.53, 0.85, 0.76, 0.91, 0.96	EC-Earth3: 0.82, 0.82, 0.72, 0.65, 0.86, 0.82, 0.94, 0.96
INGV ECHAM4: 0.83, 0.80, 0.73, 0.63, 0.86, 0.72, 0.94, 0.95	CSIRO-Mk3-6-0: 0.81, 0.71, 0.61, 0.58, 0.87, 0.78, 0.90, 0.96	FGOALS-g3: 0.74, 0.66, 0.40, 0.32, 0.78, 0.73, 0.96, 0.88
INM-CM3.0: 0.78, 0.73, 0.58, 0.51, 0.65, 0.74, 0.90, 0.97	FGOALS-g2: 0.83, 0.71, 0.65, 0.58, 0.91, 0.82, 0.92, 0.98	GISS-E2.1-G: 0.87, 0.84, 0.85, 0.74, 0.93, 0.87, 0.90, 0.90
IPSL-CM4: 0.78, 0.75, 0.46, 0.43, 0.73, 0.78, 0.90, 0.97	FIO-ESM: 0.84, 0.71, 0.62, 0.53, 0.87, 0.82, 0.91, 0.97	GFDL-ESM4: 0.85, 0.78, 0.69, 0.61, 0.91, 0.85, 0.93, 0.96
MIROC3.2 (hires): 0.80, 0.78, 0.52, 0.53, 0.76, 0.75, 0.93, 0.97	GFDL-ESM2M: 0.86, 0.66, 0.42, 0.62, 0.92, 0.89, 0.94, 0.98	HadGEM3-GC31-LL: 0.86, 0.87, 0.82, 0.56, 0.93, 0.94, 0.88, 0.91
MIUB-ECHO-G: 0.83, 0.78, 0.71, 0.64, 0.86, 0.72, 0.86, 0.96	GISS-E2-H: 0.83, 0.72, 0.69, 0.57, 0.84, 0.81, 0.90, 0.97	IPSL-CM6A-LR: 0.88, 0.77, 0.83, 0.71, 0.85, 0.85, 0.98, 0.98
MPI-ECHAM5: 0.79, 0.67, 0.55, 0.44, 0.84, 0.73, 0.90, 0.94	HadGEM2-AO: 0.85, 0.58, 0.49, 0.31, 0.76, 0.73, 0.92, 0.94	MCM-UA-1.0: 0.77, 0.82, 0.82, 0.76, 0.93, 0.92, 0.96, 0.93
MRI-CGMC2.3.2A: 0.82, 0.75, 0.74, 0.63, 0.87, 0.72, 0.91, 0.93	INM-CM4: 0.83, 0.72, 0.63, 0.57, 0.78, 0.84, 0.96, 0.96	MIROC6: 0.88, 0.90, 0.91, 0.74, 0.93, 0.90, 0.88, 0.91
NCAR CCSM3.0: 0.81, 0.76, 0.71, 0.61, 0.84, 0.74, 0.93, 0.93	IPSL-CM5A-MR: 0.82, 0.75, 0.65, 0.61, 0.78, 0.73, 0.90, 0.97	MPI-ESM1-2-HR: 0.88, 0.90, 0.89, 0.79, 0.89, 0.83, 0.86, 0.93
UKMO HadGEM1: 0.80, 0.80, 0.29, 0.52, 0.77, 0.78, 0.90, 0.96	MIROC5: 0.80, 0.68, 0.60, 0.56, 0.84, 0.82, 0.91, 0.91	MRI-ESM2-0: 0.85, 0.77, 0.73, 0.46, 0.90, 0.85, 0.93, 0.98
	MPI-ESM-MR: 0.84, 0.77, 0.74, 0.62, 0.86, 0.79, 0.91, 0.95	NESM3: 0.85, 0.79, 0.85, 0.68, 0.90, 0.86, 0.96, 0.89
	MRI-CGCM3: 0.84, 0.76, 0.75, 0.60, 0.89, 0.86, 0.90, 0.95	NorCPM1: 0.86, 0.61, 0.21, 0.08, 0.82, 0.78, 0.87, 0.94
	NorESM1-M: 0.84, 0.76, 0.63, 0.52, 0.85, 0.82, 0.92, 0.94	SAM0-UNICON: 0.82, 0.34, 0.73, 0.74, 0.87, 0.89, 0.96, 0.94
		UKESM1.0-LL: 0.86, 0.87, 0.82, 0.56, 0.93, 0.94, 0.88, 0.91

the leading PCs (Fig. 2b), the standard deviation across all CMIP models (Fig. 2c), mean bias in CMIP3 (Fig. 2d), CMIP5 (Fig. 2e), and CMIP6 models (Fig. 2f), and the two leading bias EOFs across models (Figs. 2g,h). Differences across models are largest in the vicinity of the Aleutian low and polar regions generally (Fig. 2c). Mean biases are characterized generally by anomalies that are too weak, as biases tend to be opposite in sign to observed anomalies. Net overall improvement is however evident in the reduction in mean bias magnitudes from CMIP3 to CMIP5 and CMIP6.

The dominant patterns of model bias are summarized by the two leading PCs and EOFs across terciles of the CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) generations based on their distance from observations in the PC1/2 space (Fig. 2b). The magnitude of PC1/2 values for observations is indicative of the mean overall model bias. The systematic bias of CMIP PC1 terciles toward more positive values than observations indicates the contribution of EOF1 to their mean biases, consistent for example with an underestimate of ENSO teleconnection magnitude (also evident in mean CMIP biases). Teleconnection weakness is again evident in the leading EOF of bias (Fig. 2g), which correlates negatively with the observed pattern (−0.96) and explains 29% of the variance in bias across models, and for which PC1 weights are systematically greater than observations. The patterns in models are also shifted westward
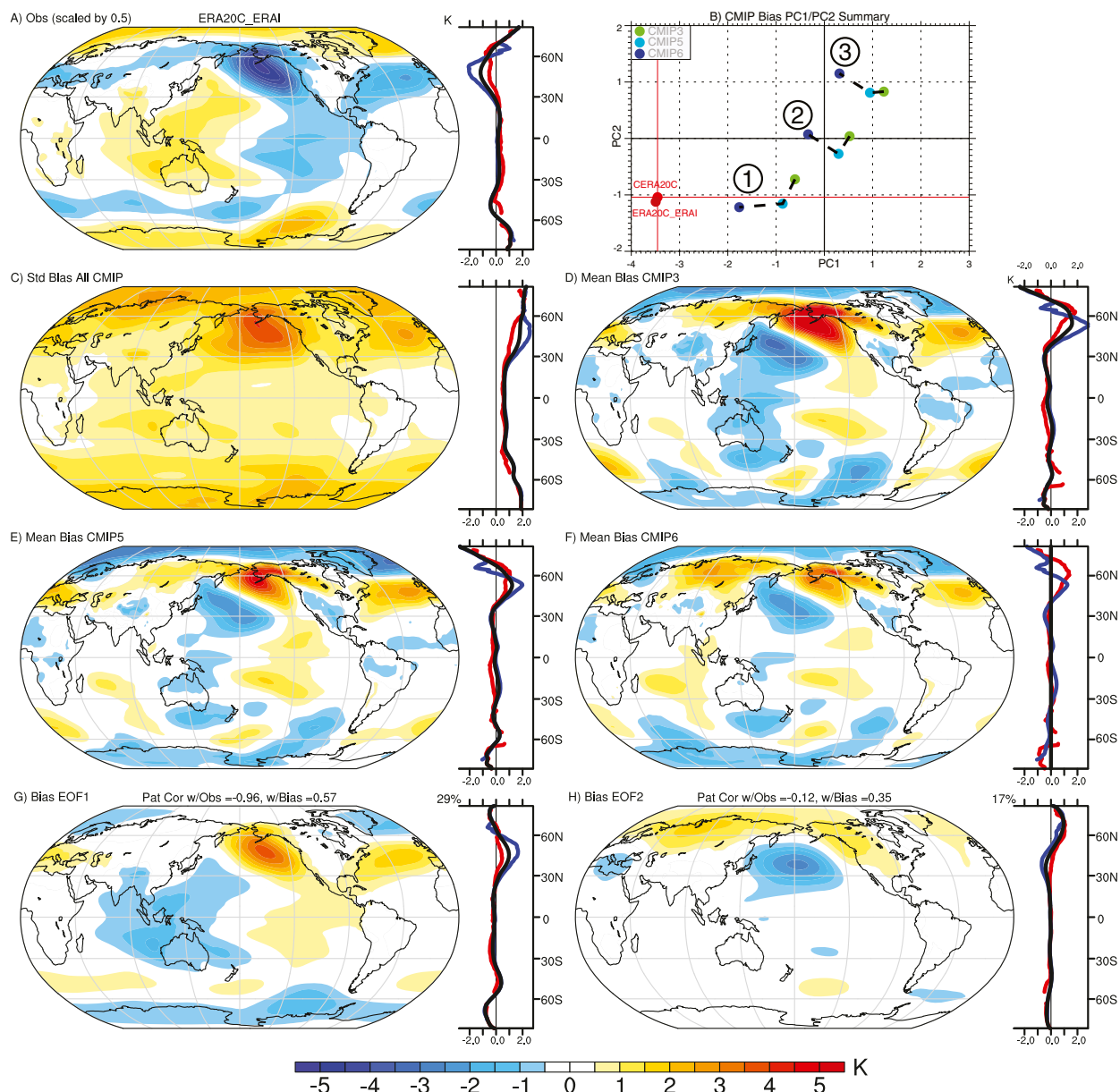
FIG. 2. (a) ENSO composites of DJF sea level pressure for 1900–2017 based on ERA-20C/ERA-I in observations. Zonal mean values are also indicated over ocean (blue), land (red), and combined (black). (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations [red; see methods (section 2)]. (c) The standard deviation of the bias of all CMIP simulations. (d) As in (a), but for mean bias in CMIP3, (e) CMIP5, and (f) CMIP6. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles.

from those observed in some respects, such as the negative biases that extend into the northwestern Pacific Ocean in their mean biases (Figs. 2d–f) and EOF2, and are too strongly positive (negative) over land (ocean) in the Arctic, and too strongly negative along the Antarctic coast. For EOF2, PC2 weights are also systematically biased positive relative to observations except for tercile 1. Improvement in PC1 across successive CMIP

generations is evident across terciles, as successive CMIP generations score closer to observations than their predecessors. Improvement in PC2 is less systematic however, evidenced by the fact that CMIP6 terciles 2 and 3 values deviate further from observations than corresponding CMIP3 and CMIP5 values.

The fidelity of El Niño's spatiotemporal structure in the equatorial Pacific is shown in Fig. 3, which shows

the composite evolution of equatorial (5°N–5°S) SST anomalies from January of year 0 through May of year 2. Observed features include warm anomalies that build early in year 0, peak at the end of year 0, are centered about 135°W, and extend westward to 170°E. On average, El Niño events transition to neutral conditions in the middle of year 1 and to cold anomalies that peak late in year 1 and then decay in year 2. The main model biases can be summarized by the leading PCs (Fig. 3b), the standard deviation across all CMIP models (Fig. 3c), mean bias in CMIP3 (Fig. 3d), CMIP5 (Fig. 3e), and CMIP6 models (Fig. 3f), and the leading bias EOFs across models (Figs. 2g,h). Differences across models are largest in the eastern tropical Pacific Ocean, particularly early in year 1 (Fig. 3c). Mean biases relative to the observed structure are qualitatively similar across the various CMIP archives. These include warm anomalies in years 0 and 1 that are too strong, extend too far west, and last too long (i.e., into the middle of year 1, Fig. 3c). Moreover, there is a tendency in some models for El Niño events to transition to conditions that are too cold on average. A reduction in bias on average from CMIP3 to CMIP6 is also clearly evident, although in instances the average bias in CMIP5 is less than in CMIP6. The EOF decomposition of bias (Figs. 3b,g,h) corroborates this improvement generally. The leading EOF is strongly correlated to the mean pattern and relates to excessive El Niño amplitude, with a tendency to simulate El Niño events, and subsequent La Niña events, that are too strong on average. The second EOF is weighted primarily toward year 1 and thus relates to the transition from El Niño to La Niña. Systematic improvements, particularly in EOF1, are evident, such as for example in the reduction of bias from CMIP3 (Fig. 3d) to CMIP6 (Fig. 3f) and the proximity with which CMIP6 PC terciles lie to observations relative to CMIP3 and CMIP5 terciles. An exception to this improvement is evident in PC2 of terciles 2 and 3.

A similar analysis of La Niña is presented in Fig. 4. In nature, multiyear events dominate the La Niña composite (Okumura and Deser 2010; DiNezio and Deser 2014), leading to weak but persistent cool anomalies on average through the end of year 2 (Fig. 4a). The main biases can again be summarized by the leading PCs (Fig. 4b), the standard deviation across all CMIP models (Fig. 4c), mean bias in CMIP3 (Fig. 4d), CMIP5 (Fig. 4e), and CMIP6 models (Fig. 4f), and the leading bias EOFs across models (Figs. 4g,h). Differences across models are again largest in the eastern Pacific Ocean near the start of year 1 (Fig. 4c). Models have difficulty simulating the persistence of cool anomalies, as evident from their warm biases late in year 1 in across CMIP3, CMIP5,

and CMIP6 composite means. Cold biases in the western Pacific Ocean also characterize CMIP models generally, revealing the excessive westward extent of ENSO also seen for El Niño, although a reduction of bias on average is evident from CMIP3 to CMIP6. EOF decomposition of La Niña composites (Figs. 4b,g,h) results in patterns that correlate moderately with both the mean pattern and bias across models. The PCs and their structure across CMIP terciles in many ways mirror those for El Niño (Fig. 3d), with some improvement being evident across generations, a large intramodel spread, and a reduction of ensemble spread from CMIP3 to CMIP6. The leading pattern of bias relates to the strength of La Niña anomalies at the end of year 1, which are too cold on average (Figs. 4b,c) while the second pattern highlights the transition into year 2 to conditions that are too warm on average, reflecting the inability of models generally to sustain multiyear events, in line with DiNezio and Deser (2014).

4. Teleconnections of the Pacific decadal oscillation

In recent years, the PDO has become recognized as resulting from a combination of distinct geographically remote processes, both in the tropics and extratropics, and particularly in the North Pacific Ocean (Newman et al. 2016). These processes operate across a range of time scales yet drive similar net responses in the North Pacific Ocean to contribute to the PDO's complex observed behavior. The main influences include changes in ocean surface heat fluxes and Ekman transport, ocean memory, and wind-driven decadal changes in the Kuroshio and Oyashio. The adequate simulation of the PDO therefore relies on the simulation of a broad range of physical processes and spatial scales, extending from the fine scales of these current systems to the planetary scale of teleconnections associated with ENSO. Figure 5a shows the pattern of the PDO in ERSSTv5 of SSTA regressed onto the PDO index derived based on the method outlined in Mantua et al. (1997). The pattern is characterized by strong negative values in the North Pacific Ocean, weak negative values in the extratropical ocean basins generally and western Pacific Ocean that extend in a horseshoe-shaped pattern poleward, and strong positive values that extend eastward and toward midlatitudes in the tropical Pacific Ocean from the date line. Weak positive values are also evident in the tropical Indian and Atlantic Ocean basins.

Biases in CMIP models can be summarized by the leading PCs (Fig. 5b), the standard deviation across all CMIP models (Fig. 5c), mean bias in CMIP3 (Fig. 5d), CMIP5 (Fig. 5e), and CMIP6 models (Fig. 5f),
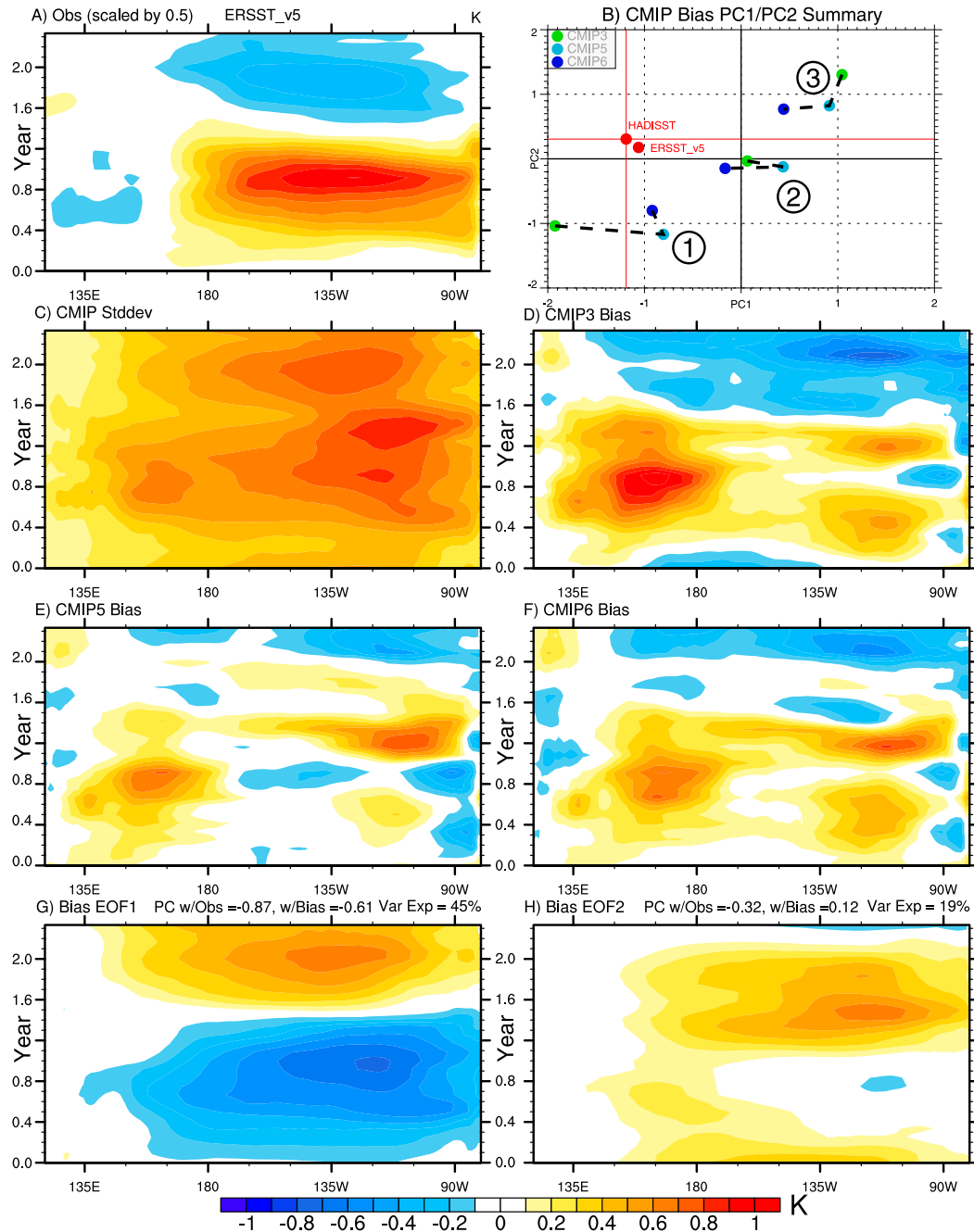
FIG. 3. Hovmöller diagrams (longitude vs time) of (a) composite equatorial SST anomalies during El Niño events based on ERSSTv5 observational estimates. (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red; see methods). (c) The standard deviation of the bias of all CMIP simulations. (d) As in (a), but for mean bias in CMIP3, (e) CMIP5, and (f) CMIP6. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles. Vertical axes extend from January of year 0 to May of year 2.

and the leading bias EOFs across models (Figs. 5g,h). Differences across models are largest in the North Pacific Ocean and tropical Pacific Ocean (Fig. 5c). Mean biases are characterized by patterns that are too weak,

with negative biases in the eastern Pacific Ocean and positive biases in the western Pacific Ocean. A reduction in bias from CMIP3 to CMIP6 in most regions is also clearly evident. The leading PCs and EOFs of
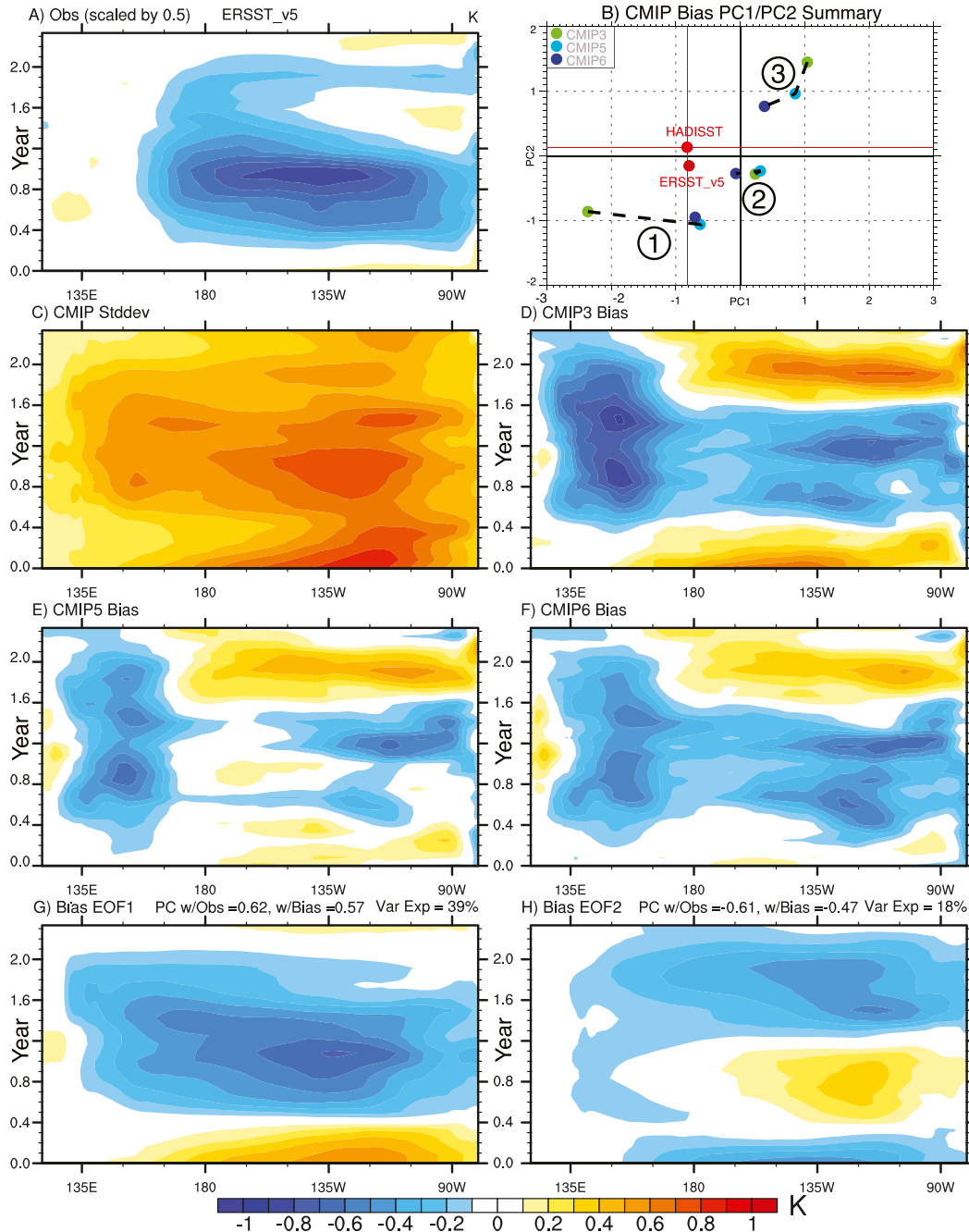
FIG. 4. As in Fig. 3, but for La Niña.

model bias are shown Figs. 5b, 5g, and 5h and highlight connections with the tropics as being significantly underestimated in some models and a general weakness in simulated patterns. These aspects are evident for example in EOF1, which is negatively correlated with the mean bias ($r = -0.89$) and explains a significant fraction of the variance in bias across models (32%). The spatial structure of EOF2 resembles the

North Pacific Gyre Oscillation (DiLorenzo et al. 2008). The mode highlights contrasts in the zonal structure of anomalies in the North Pacific Ocean, a structure that is not strongly correlated to the mean model bias ($r = -0.23$) and only accounts for about half of the variance (16%) explained by EOF1. As is the case for the ENSO composite in Fig. 2, systematic model biases are evident in PC1 terciles,
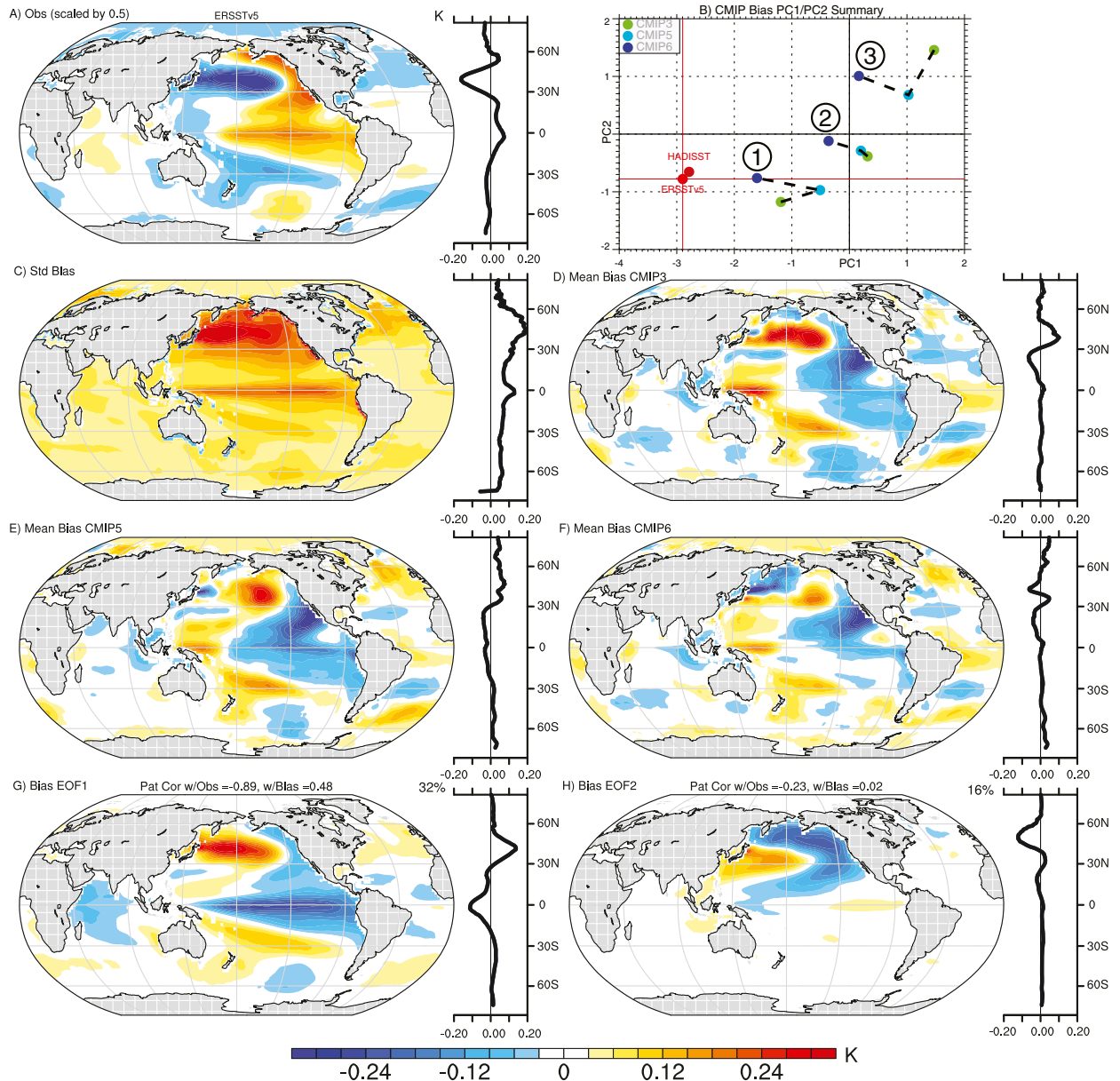
FIG. 5. (a) Spatial pattern of the observed Pacific decadal oscillation (based on NOAA's ERSSTv5 for 1900–2018). Zonal mean values are also indicated over ocean (black). (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations [red; see methods (section 2)]. (c) The standard deviation of bias across all CMIP simulations. (d) As in (a), but for mean PDO pattern bias in CMIP3, (e) CMIP5, and (f) CMIP6. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and the pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles.

which suggest systematic positive contributions from EOF1 (i.e., weakness) in model patterns generally. Improvement is also apparent across CMIP generations with the reduction of distances in tercile PC1/2 values from observational estimates, with some variability across the terciles of each generation, such as for example in PC1 for tercile 1, where CMIP3 lies closer to observations than does CMIP5, and in PC2

for tercile 3, where CMIP5 lies closer to observations than does CMIP6.

5. Extratropical modes of variability

Large seasonal variability is inherent to SLP in the extratropics. In the NH, subtropical anticyclones dominate during summer and weaken and move equatorward
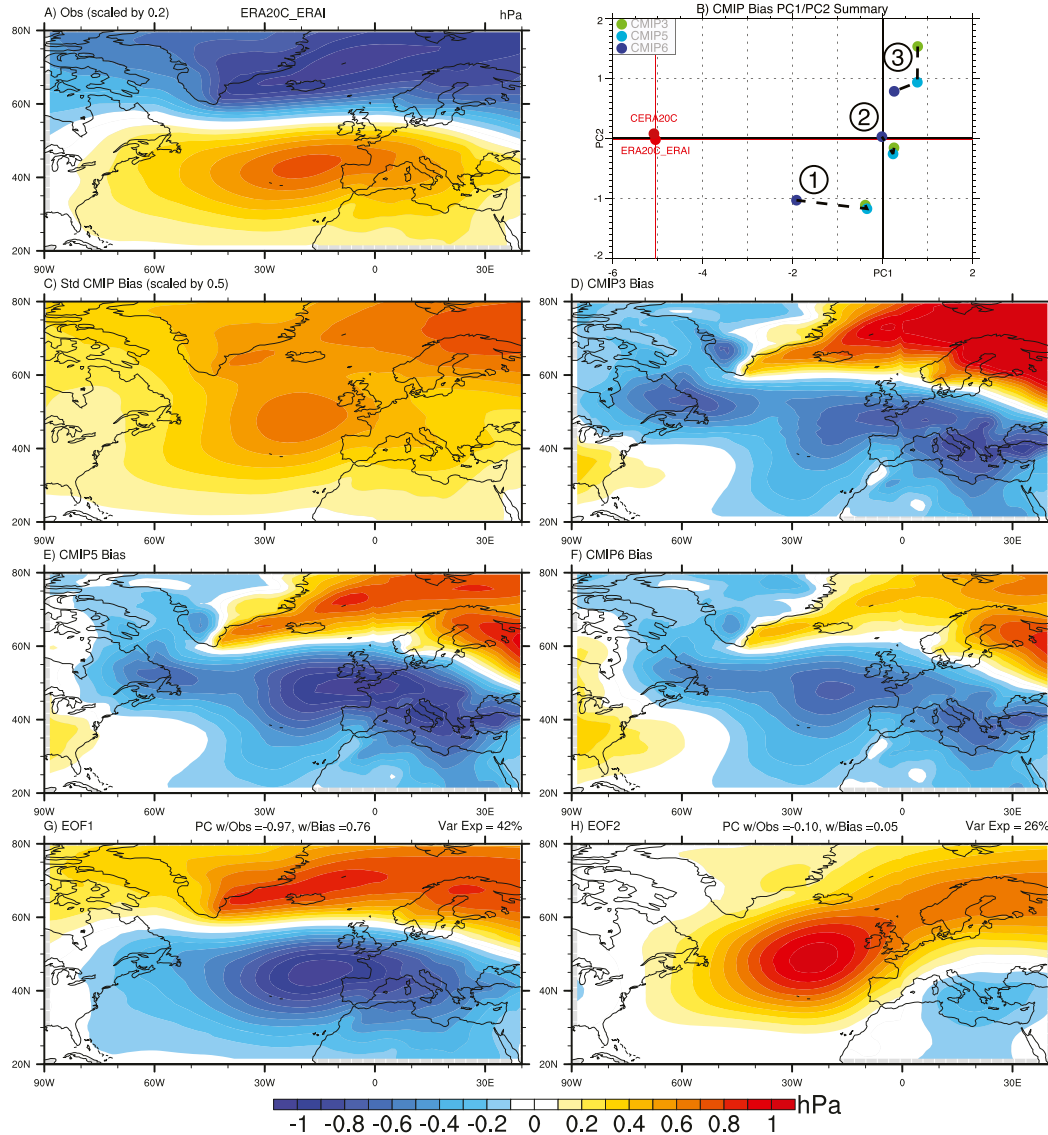
FIG. 6. (a) Observed pattern of the North Atlantic Oscillation in DJF (based on ERA20C/ERA-Interim from 1950 to 2018). (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations [red; see methods (section 2)]. (c) The standard deviation of bias across all CMIP simulations. (d) As in (a), but for mean bias in CMIP3, (e) CMIP5, and (f) CMIP6. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and the pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles.

by winter, interacting with the high-latitude Aleutian and Icelandic low pressure centers. Variations in these features largely comprise the centers of action for NH variability and its dominant modes, the NAM and NAO. Features of the NAO pattern are shown in Fig. 6a. Out-of-phase centers of action reside at approximately 40° and 65°N, just east of the Azores and west of Iceland, respectively, with strong zonal coherence in the pattern. The mean model biases can be summarized by the leading PCs (Fig. 6b), the standard deviation across all

CMIP models (Fig. 6c), mean bias in CMIP3 (Fig. 6d), CMIP5 (Fig. 6e), and CMIP6 models (Fig. 6f), and the leading bias EOFs across models (Figs. 6g,h). Differences across models are largest in the vicinity north of the Azores high and Nordic seas (Fig. 6c). Mean biases for CMIP3, CMIP5, and CMIP6 suggest systematic weakness in both the Azores high and the Icelandic low (i.e., the biases are opposite in sign to the observed anomalies), particularly in CMIP5 models in the Azores region. The leading EOFs differentiating simulated
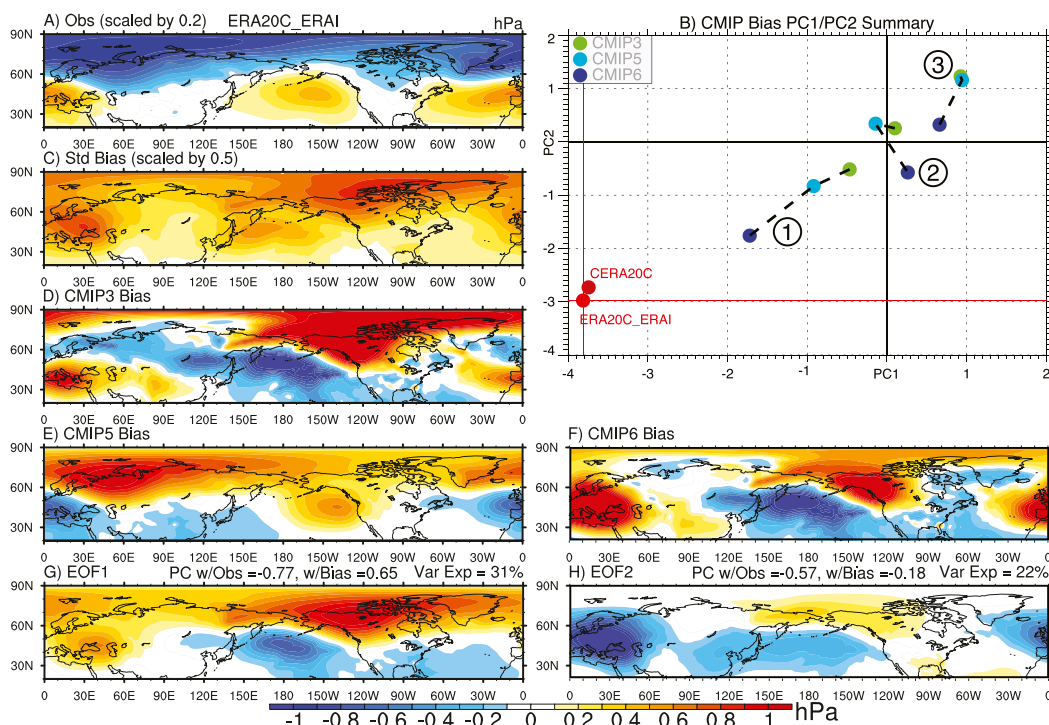
FIG. 7. Observed (a) pattern of the northern annular mode in DJF (based on ERA20C/ERA-Interim from 1950 to 2018). (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations [red; see methods (section 2)]. (c) The standard deviation of bias across all CMIP simulations. (d)–(f) As in (a), but for mean bias in CMIP3, CMIP5, and CMIP6, respectively. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and the pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles.

patterns relate mainly to the strength (EOF1; Fig. 6g) and eastward tilt of the pattern (Fig. 6h), with some models simulating a pattern that is too strong with a northward center of action that extends well to the east and north of Iceland. Some reduction of these biases is evident across CMIP generations, with CMIP6 simulations exhibiting PC weightings closer to observations than their predecessors for all terciles (Fig. 6b), although improvement has not been monotonic across CMIP generations as significant systematic amplitude errors are evident in CMIP5.

Observed features of the NAM (Fig. 7a) include centers of action across the Arctic, and in the vicinity of the Aleutian low and Azores high, with resemblance to the NAO in the Atlantic sector (Fig. 6; see also Feldstein and Franzke 2006). The mean bias patterns can again be summarized by the leading PCs (Fig. 7b), the standard deviation across all CMIP models (Fig. 7c), mean bias in CMIP3 (Fig. 7d), CMIP5 (Fig. 7e), and CMIP6 models (Fig. 7f), and the leading bias EOFs across models (Figs. 7g,h). Differences across models are largest in Europe and north of Canada (Fig. 7c). Mean biases vary across CMIP models although generally they exhibit

patterns that are negatively correlated to the observed pattern and thus relate mainly to its amplitude. EOF1 of the model bias captures this aspect in the Pacific Ocean (Fig. 7g), while EOF2 is dominated by a center of action over and downstream from the Aleutian low (Fig. 7h). The PC weights (Fig. 7b) demonstrate that model biases are systematically positive for both PC1 and PC2, indicative of the regional weaknesses of the simulated NAM pattern across models (PC1/EOF1) and biases in the weighting between the Aleutian low and anomalies over Europe (PC2/EOF2). Some improvement across CMIP archives is evident in the PC weights, particularly for the tercile of models that lies closest to observations where bias is reduced considerably in CMIP6.

The SAM is a quasi-zonally symmetric mode of variability in the SH, typically identified through SLP or lower tropospheric geopotential height variability (Fig. 8a), with opposing centers of action over Antarctica and a weaker opposing zonal band centered near 45°S (Hartmann and Lo 1998; Thompson and Wallace 2000). Here we use only data after 1950 due to data scarcity in the early twentieth century (Schneider and Fogt 2018). A particularly strong center of action exists in the Amundsen
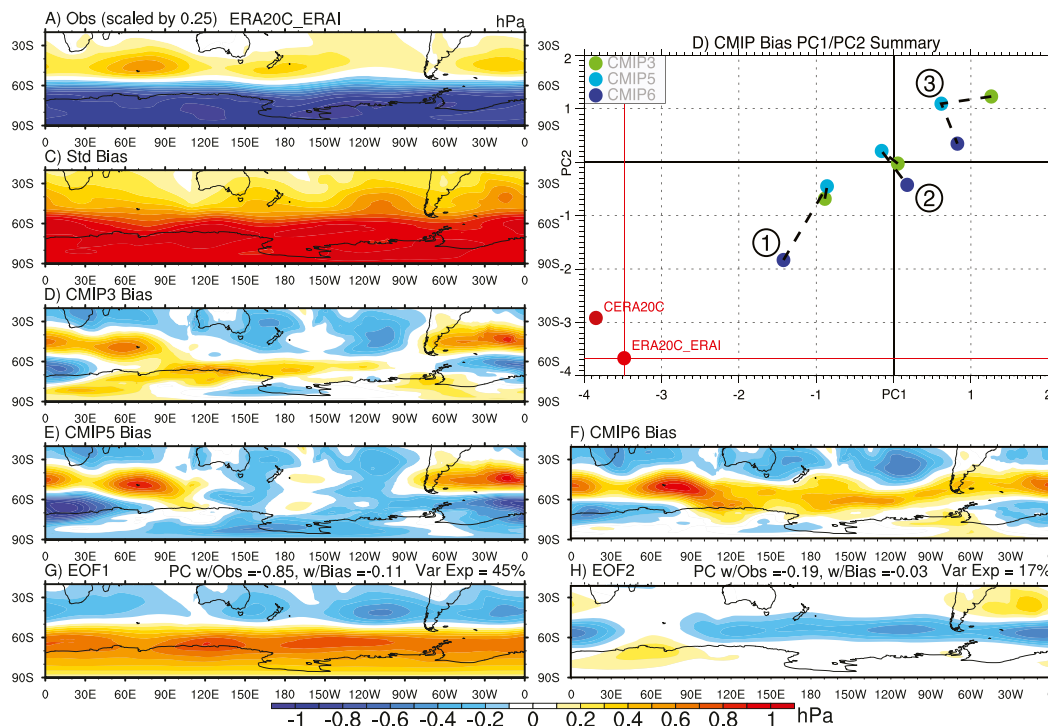
FIG. 8. (a) Observed pattern of the southern annular mode in DJF (based on ERA20C/ERA-Interim from 1950 to 2018). (b) The first and second PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations [red; see methods (section 2)]. (c) The standard deviation of bias across all CMIP simulations. (d) As in (a), but for mean bias in CMIP3, (e) CMIP5, and (f) CMIP6. (g) The first and (h) the second EOFs of model bias (including all CMIP generations); variance explained and the pattern correlations of each against observations [in (a)] and CMIP mean bias are indicated in the panel titles.

Sea. Biases in the SAM can be summarized by the leading PCs (Fig. 8b), the standard deviation across all CMIP models (Fig. 8c), mean bias in CMIP3 (Fig. 8d), CMIP5 (Fig. 8e), and CMIP6 models (Fig. 8f), and the leading bias EOFs (Figs. 8g,h). Differences across models are largest in the vicinity of the Antarctica generally, and particularly along east Antarctica and the Weddell Sea (Fig. 8c). As with other extratropical modes, the mean patterns of bias and the leading patterns resulting from the PC analysis (Figs. 8b,g,h) relate to weakness in simulated patterns, as the leading EOF of bias (Fig. 8c) correlates strongly with the mean pattern ($r = -0.85$) and explains a majority of the variability across models (45%). Much weaker is EOF2, which relates to details in the meridional extent of simulated anomalies across the Southern Ocean (EOF2; Fig. 8h), explaining only 17% of the intermodel variance. Modest reductions in these biases are evident in mean PC magnitudes across model terciles, with CMIP6 scoring much better in tercile 1 for PC1 and across all terciles for PC2. Bias increases are notable as well as CMIP5 positive biases in the Southern Ocean and negative biases along Antarctica are larger than in CMIP3. Nonetheless,

significant systematic bias remains, particularly in PC1 where CMIP6 values indicate a significant contribution of the EOF1 bias pattern.

6. Performance across CMIP generations

A summary of the broader distributions of model scores in reproducing the major indices of variability is provided in the CVDP metrics tables and displayed graphically in Fig. 9. Metric scores for individual models are listed in Table 1. An average of the full set of scored metrics (see Phillips et al. 2014) is used to compute the overall score. To provide a more complete sampling of histogram distributions, all ensemble members are included in this analysis, although the main results are unchanged by considering only one member per modeling center. Also shown are the corresponding ranges of scores from the CESM-LE and MPI-GE, where the spread is solely driven by internal variability, providing context for interpreting CMIP ensemble spread and evolution across generations. Most notably, the scoring distributions are skewed and generally non-Gaussian, with a tail of low scoring models being evident,
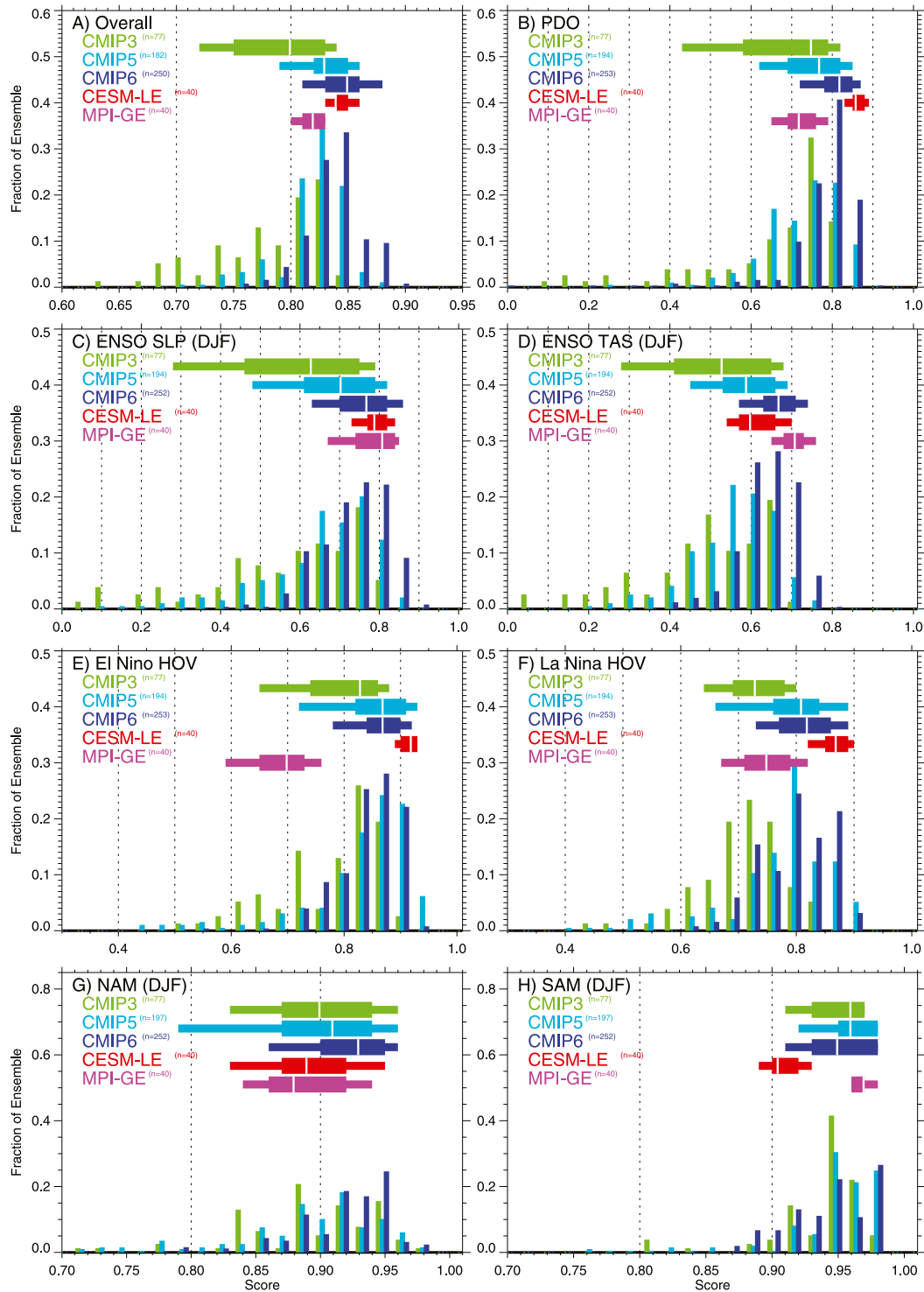
FIG. 9. Histogram plots of the distributions of pattern correlations for CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) for (a) the overall CVDP score (Phillips et al. 2014), and scores for (b) the PDO, (c),(d) El Niño DJF TAS and SLP composites, (e),(f) El Niño and La Niña TS Hovmöllers, and DJF (g) NAM and (h) SAM. Horizontal whisker lines on each plot correspond to the 10th–90th percentile range (thin lines), interquartile range (thick lines), and median value (white). The number of simulations considered *n* is also indicated. Corresponding results from the CESM-LE (red) and MPI-LE (cyan) are also shown.

TABLE 2. Correlations between metrics in Table 1 across all CMIP models. HOV stands for Hovmöller.

	Overall	PDO	ENSO-SLP	ENSO-TAS	El Niño-HOV	La Niña-HOV	NAM	SAM
Overall	—	0.51	0.64	0.72	0.61	0.57	0.17	−0.03
PDO	0.51	—	0.63	0.73	0.29	0.49	0.08	−0.22
ENSO-SLP	0.64	0.63	—	0.83	0.59	0.50	0.02	−0.22
ENSO-TAS	0.72	0.73	0.83	—	0.61	0.56	0.06	−0.13
El Niño-HOV	0.61	0.29	0.59	0.61	—	0.64	0.14	−0.27
La Niña-HOV	0.57	0.49	0.50	0.56	0.64	—	0.22	−0.32
NAM	0.17	0.08	0.02	0.06	0.14	0.22	—	0.0
SAM	−0.03	−0.22	−0.22	−0.13	−0.27	−0.32	0.0	—

particularly for CMIP3 where overall scores fell below 0.7 for some models (Fig. 9a). The median overall score for CMIP3 (0.80) improves somewhat in CMIP5 to an average of 0.83 and continues to improve with CMIP6 (0.85), due in part to fewer low scoring simulations (Fig. 9a). Differences within and across CMIP generations are considerably larger than internal variability in the CESM-LE and MPI-GE. Median scores for PDO patterns also increase, from 0.75 in CMIP3 to 0.77 in CMIP5 and 0.82 in CMIP6. ENSO SLP median scores (Fig. 9c) increase considerably across generations, from 0.63 to 0.71 and 0.77 in CMIP3, CMIP5, and CMIP6, respectively. Differences within and across CMIP generations for the PDO and ENSO SLP teleconnections are again considerably larger than internal variability in the CESM-LE though only somewhat larger than in the MPI-GE. Corresponding scores for ENSO DJF TAS composites (Fig. 9d) are 0.53, 0.59, and 0.67. Spatiotemporal patterns for El Niño (Fig. 9e) and La Niña (Fig. 9f) Hovmöllers are generally greater than for the overall score and they exhibit less spread. Scores across model generations improve for El Niño from 0.83 in CMIP3 to 0.87 in both CMIP5 and CMIP6 while scores for La Niña Hovmöllers increase from 0.73 in CMIP3 to 0.81 in CMIP5 and 0.82 in CMIP6. Differences within and across CMIP generations are also larger than internal variability in the CESM-LE although spread in the MPI-GE is substantially larger for the El Niño and La Niña Hovmöllers and is not significantly less than intermodel differences. Mean correlations are higher for CESM generally than for the MPI-GE, except for SAM scores. Scores for the NAM and SAM are systematically higher across models than scores for ENSO or the PDO, and changes across model generations are less significant, with NAM (Fig. 9e) and SAM (Fig. 9f) CMIP3 median scores of 0.90 and 0.96, respectively, and CMIP6 scores of 0.93 and 0.95, respectively. Changes across interquartile ranges are inconsistent, unlike changes for the overall score (Fig. 9a) and many other modes where progressive improvements are generally evident in both the mean, interquartile ranges, and 10th and 90th percentile ranges (whisker

lines). Internal variability contributes significantly to the intermodel spread, although spread in the MPI-GE scores for SAM are small relatively. Beyond improvements in median scores across modes, perhaps the most notable aspect of the evolution of the CMIP archives is the substantial improvement in the lowest scoring models across generations.

Correlations across models between the various scores in Fig. 9 are summarized in Table 2. They show that the overall score as computed by the CVDP is strongly influenced by ENSO and the PDO, which is expected as it incorporates multiple scores related to ENSO. A strong relationship also exists between the PDO and various ENSO scores, indicating that models with greater realism in their PDO pattern also exhibit greater fidelity in their ENSO patterns. There is no connection between NAM and SAM scores across models, and only a weak relationship in the scores between the extratropical atmospheric modes and ENSO or the PDO. This lack of strong correlations in fidelity is consistent with the finding of Deser et al. (2017a) in which connections between other extratropical modes (NAO/PNA) were found to be largely absent.

As a complement to pattern correlation scores, the spectra of ENSO and the PDO summarized across various frequency bands are shown in Figs. 10a and 10b, respectively, where the full distribution of each CMIP generation is summarized by its minimum and maximum values and interquartile range. Also shown are the corresponding ranges in spectral power from the CESM-LE and MPI-GE to provide context for apparent model bias. For ENSO, the power on time scales less than 2.5 years is less than observed for all CMIP generations and simulations and the estimated influence of internal variability based on either the CESM-LE or MPI-GE ranges is small (Fig. 10a). In the 2.5–6-yr band, ENSO power across models varies widely, particularly for CMIP3 models where two models exhibit variance that is an order of magnitude too large (see caption). Observational estimates tend to fall within the interquartile range of the CMIP3 and CMIP5 ensemble range and within the full CMIP6 range, with a suggestion
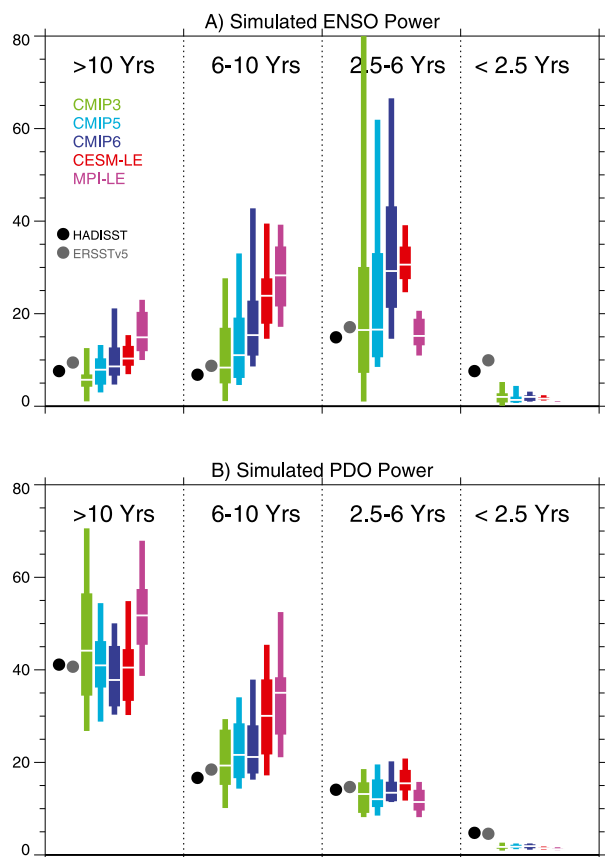
FIG. 10. Whisker plots of the band-averaged power spectra in CMIP3 (green), CMIP5 (aqua), and CMIP5 (blue) for (a) Niño-3.4 SST anomalies and (b) the PDO index. Thin colored lines span the full ensemble range while thick lines span the 25th–75th percentile range. The equivalent spans from CESM-LE (red) and MPI-LE (cyan) are also shown. Two CMIP3 simulations greatly exceed the range for the 2.5–6-yr band (CNRM-CM3, >200; IAP-FGOALS, >600). Units are °C^2 (cycles per month)$^{-1}$.

that many models may overestimate power in this band. This possibility is further supported by the relatively narrow ranges of the CESM-LE and MPI-GE in the 2.5–6-yr band, as also identified in Maher et al. (2018). Although biases in the spectra of individual models can be identified in some instances (e.g., Bellenger et al. 2014), such as for example as suggested by those model–observational differences lying well beyond the magnitude of the CESM-LE and MPI-GE ranges, definitive statements regarding model bias across the archives are limited by internal variability, which comprises a significant fraction of the ensemble spread, and the limited duration of the observational record (see also Deser et al. 2012b). Over longer periods (6–10 years, >10 years), a progressive increase in ENSO power across CMIP generations is again evident and observations are generally consistent with the full ensemble range for each CMIP generation, although for periods >10 years,

agreement between observations and CMIP is greatest for CMIP6 as observations tend to show more power than the interquartile range of CMIP3 (Fig. 10a). Power in the 6–10-yr band in the MPI-GE is larger than in the CESM-LE, although the ranges of both lie above observed estimates.

Spectral agreement of the PDO with observations in many ways mirrors that of ENSO, with deficient power for short periods across CMIP generations, and reasonable agreement in the 2.5–6- and 6–10-yr bands (Fig. 10b). At lower frequencies, agreement between observed power and the interquartile ranges of all CMIP generations is good. One of the more interesting aspects of the comparison is the finding that the estimates of internal variability from the CESM-LE and MPI-GE are in many instances greater than the spread of the CMIP archives, which is unexpected given that the CMIP distributions include both structural uncertainty combined with internal variability. This finding highlights the model dependency of estimated contributions from internal variability (see also Deser et al. 2020) and the associated challenges in evaluating model spectra. Unlike for ENSO, there is no systematic change in spectral power across CMIP generations for the PDO index, with a slight increase for the 6–10-yr band and a reduction at periods greater than 10 years (Fig. 10b).

7. Summary, discussion, and conclusions

Indices for the major modes of climate variability and their associated patterns have been computed and used to evaluate coupled climate models across CMIP generations. The CESM-LE and MPI-GE have been used estimate the internal variability of these modes (N) as it compares to the broader CMIP model spread (S). Modes have been selected for which S/N is relatively large and observational uncertainty is small. The evaluation has demonstrated systematic improvement of the representation of modes of variability across the CMIP archives, particularly for ENSO and the PDO, while extratropical atmospheric modes, though exhibiting general increases in scores across generations, tend to score high across all with changes that are small relative to intramodel spread.

A focus here has been on composite patterns of the modes, both in space and time, and strong correlations between leading modes of bias with these patterns suggests that leading order errors in both simulated amplitude and structure are important contributors to model bias. The reduction of these bias components is evidenced by their PCs, which tend to migrate toward observed values across their upper, middle, and lower terciles, with some notable exceptions that are identified. The distributions of pattern correlations also shift

toward higher values systematically across the CMIP generations, though for extratropical modes correlations have been high in all CMIP archives and improvements have been small relative to the intermodel spread, making improvements harder to detect. The bandpass spectra of ENSO and PDO indices have also been assessed, although in these cases a clear progression toward higher fidelity across model generations is not apparent, perhaps in part due to the presence of large intrinsic noise.

While the aspects of variability assessed here are strong indicators of model fidelity, other characteristics remain to be assessed. These include impact relevant teleconnections, such as rainfall and temperature (for modes identified through SLP), and connections to clouds, their feedbacks, and contributors to atmospheric diabatic heating. Such an assessment is likely to be particularly useful for evaluating the drivers of biased simulation of variability and may be useful for constraining longer-term feedbacks and projections (Lutsko and Takahashi 2018), a topic of particular interest given early indications of higher climate sensitivity from some CMIP6 models (Gettelman et al. 2019). It will also be central to efforts to evaluate and interpret model estimation of associated impacts and their changes in a changing climate. In total, the results presented here suggest a progressive increase in model fidelity in simulating many, but not all, major internal modes of variability considered, albeit with reduced but persisting biases, making climate models increasingly suitable for attributing past changes, predicting future climate, and estimating associated uncertainties.

Data availability statement: Simulations used for this study are available on the Earth System Grid (https://

www.earthsystemgrid.org) while CVDP output is available from the CVDP Repository (http://www.cesm.ucar.edu/working_groups/CVC/cvdp/data-repository.html).

REFERENCES

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115**, 1083–1126, https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dyn.*, **42**, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z.

Berrisford, P., D. Dee, K. Fielding, M. Fuentes, P. Kallberg, S. Kobayashi, and S. Uppala, 2009: The ERA-Interim archive. ERA Rep. Series 1, 16 pp., https://www.ecmwf.int/en/elibrary/8173-era-interim-archive.

Bojariu, R., and F. Giorgi, 2005: The North Atlantic Oscillation signal in a regional climate simulation for the European region. *Tellus*, **57A**, 641–653, https://doi.org/10.3402/tellusa.v57i4.14709.

Booth, B. B. B., N. J. Dunstone, P. R. Halloran, T. Andrews, and N. Bellouin, 2012: Aerosols implicated as a prime driver of twentieth-century North Atlantic climate variability. *Nature*, **484**, 228–232, https://doi.org/10.1038/nature10946.

Capotondi, A., Y. G. Ham, A. Wittenberg, and J. S. Kug, 2015: Climate model biases and El Niño Southern Oscillation (ENSO) simulation. *US CLIVAR Variations*, No. 13, U.S. International CLIVAR Project Office, Southampton, United Kingdom, 21–25.

Cattiaux, J., and C. Cassou, 2013: Opposite CMIP3/CMIP5 trends in the wintertime northern annular mode explained by combined local sea ice and remote tropical influences. *Geophys. Res. Lett.*, **40**, 3682–3687, https://doi.org/10.1002/grl.50643.

Chen, Z., B. Gan, L. Wu, and F. Jia, 2018: Pacific–North American teleconnection and North Pacific Oscillation: Historical simulation and future projection in CMIP5 models. *Climate Dyn.*, **50**, 4379–4403, https://doi.org/10.1007/s00382-017-3881-9.

Compo, G. P., and Coauthors, 2011: The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, https://doi.org/10.1002/qj.776.

Dai, A., and C. E. Bloecker, 2019: Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models. *Climate Dyn.*, **52**, 289–306, https://doi.org/10.1007/s00382-018-4132-4.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012a: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, https://doi.org/10.1007/s00382-010-0977-x.

——, and Coauthors, 2012b: ENSO and Pacific decadal variability in Community Climate System Model version 4. *J. Climate*, **25**, 2622–2651, https://doi.org/10.1175/JCLI-D-11-00301.1.

——, A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**, 2271–2296, https://doi.org/10.1175/JCLI-D-13-00451.1.

——, L. Terray, and A. S. Phillips, 2016: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *J. Climate*, **29**, 2237–2258, https://doi.org/10.1175/JCLI-D-15-0304.1.

——, I. R. Simpson, K. A. McKinnon, and A. S. Phillips, 2017a: The Northern Hemisphere extra-tropical atmospheric circulation

response to ENSO: How well do we know it and how do we evaluate models accordingly? *J. Climate*, **30**, 5059–5082, https://doi.org/10.1175/JCLI-D-16-0844.1.

——, J. W. Hurrell, and A. S. Phillips, 2017b: The role of the North Atlantic Oscillation in European climate projections. *Climate Dyn.*, **49**, 3141–3157, https://doi.org/10.1007/s00382-016-3502-z.

——, and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Di Lorenzo, E., and Coauthors, 2008: North Pacific Gyre Oscillation links ocean climate and ecosystem change. *Geophys. Res. Lett.*, **35**, L08607, https://doi.org/10.1029/2007GL032838.

DiNezio, P. N., and C. Deser, 2014: Nonlinear controls on the persistence of La Niña. *J. Climate*, **27**, 7335–7355, https://doi.org/10.1175/JCLI-D-14-00033.1.

Eyring, V., and Coauthors, 2016a: ESMValTool (v1.0)—A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geosci. Model Dev.*, **9**, 1747–1802, https://doi.org/10.5194/gmd-9-1747-2016.

——, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016b: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

——, and Coauthors, 2016c: Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dyn.*, **7**, 813–830, https://doi.org/10.5194/esd-7-813-2016.

Feldstein, S. B., and C. Franzke, 2006: Are the North Atlantic Oscillation and the northern annular mode distinguishable? *J. Atmos. Sci.*, **63**, 2915–2930, https://doi.org/10.1175/JAS3798.1.

Furtado, J. C., E. Di Lorenzo, N. Schneider, and N. A. Bond, 2011: North Pacific decadal variability and climate change in the IPCC AR4 models. *J. Climate*, **24**, 3049–3067, https://doi.org/10.1175/2010JCLI3584.1.

Gettelman, A., and Coauthors, 2019: High climate sensitivity in the Community Earth System Model version 2 (CESM2). *Geophys. Res. Lett.*, **46**, 8329–8337, https://doi.org/10.1029/2019GL083978.

Gillett, N. P., and J. C. Fyfe, 2013: Annular mode changes in the CMIP5 simulations. *Geophys. Res. Lett.*, **40**, 1189–1193, https://doi.org/10.1002/GRL.50249.

Gleckler, P. C., C. Doutriaux, P. Durack, K. Taylor, Y. Zhang, D. Williams, E. Mason, and J. Servonnat, 2016: A more powerful reality test for climate models. *Eos, Trans. Amer. Geophys. Union*, **97**, https://eos.org/science-updates/a-more-powerful-reality-test-for-climate-models.

Guilyardi, E., H. Bellenger, M. Collins, S. Ferrett, W. Cai, and A. Wittenberg, 2012: A first look at ENSO in CMIP5. *CLIVAR Exchanges*, No. 58, International CLIVAR Project Office, Southampton, United Kingdom, 29–32.

Guo, R., C. Deser, L. Terray, and F. Lehner, 2019: Human influence on winter precipitation trends (1921–2015) over North America and Eurasia revealed by dynamical adjustment. *Geophys. Res. Lett.*, **46**, 3426–3434, https://doi.org/10.1029/2018GL081316.

Hartmann, D. L., and F. Lo, 1998: Wave-driven zonal flow vacillation in the Southern Hemisphere. *J. Atmos. Sci.*, **55**, 1303–1315, https://doi.org/10.1175/1520-0469(1998)055<1303:WDZFVI>2.0.CO;2.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, **269**, 676–679, https://doi.org/10.1126/science.269.5224.676.

——, and C. Deser, 2009: North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.*, **78**, 28–41, https://doi.org/10.1016/j.jmarsys.2008.11.026.

Imbers, J., A. Lopez, C. Huntingford, and M. Allen, 2014: Sensitivity of climate change detection and attribution to the characterization of internal climate variability. *J. Climate*, **27**, 3477–3491, https://doi.org/10.1175/JCLI-D-12-00622.1.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kumar, D., and A. R. Ganguly, 2018: Intercomparison of model response and internal variability across climate model ensembles. *Climate Dyn.*, **51**, 207–219, https://doi.org/10.1007/s00382-017-3914-4.

Laloyaux, P., and Coauthors, 2018: CERA-20C: A coupled reanalysis of the twentieth century. *J. Adv. Model. Earth Syst.*, **10**, 1172–1195, https://doi.org/10.1029/2018MS001273.

Lee, J., K. R. Sperber, P. J. Gleckler, C. J. Bonfils, and K. E. Taylor, 2019: Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dyn.*, **52**, 4057–4089, https://doi.org/10.1007/s00382-018-4355-4.

Lee, Y. Y., and R. X. Black, 2015: The structure and dynamics of the stratospheric northern annular mode in CMIP5 simulations. *J. Climate*, **28**, 86–107, https://doi.org/10.1175/JCLI-D-13-00570.1.

Lehner, F., C. Deser, I. R. Simpson, and L. Terray, 2018: Attributing the US Southwest's recent shift into drier conditions. *Geophys. Res. Lett.*, **45**, 6251–6261, https://doi.org/10.1029/2018GL078312.

Lenssen, N. J., G. A. Schmidt, J. E. Hansen, M. J. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, 6307–6326, https://doi.org/10.1029/2018JD029522.

Li, G., and S.-P. Xie, 2012: Origins of tropical-wide SST biases in CMIP multi-model ensembles. *Geophys Res. Lett.*, **39**, L22703, https://doi.org/10.1029/2012GL053777.

Lienert, F., J. C. Fyfe, and W. J. Merryfield, 2011: Do climate models capture the tropical influences on North Pacific sea surface temperature variability? *J. Climate*, **24**, 6203–6209, https://doi.org/10.1175/JCLI-D-11-00205.1.

Lutsko, N. J., and K. Takahashi, 2018: What can the internal variability of CMIP5 models tell us about their climate sensitivity? *J. Climate*, **31**, 5051–5069, https://doi.org/10.1175/JCLI-D-17-0736.1.

Maher, N., D. Matei, S. Milinski, and J. Marotzke, 2018: ENSO change in climate projections: Forced response or internal variability? *Geophys. Res. Lett.*, **45**, 11–390, https://doi.org/10.1029/2018GL079764.

——, and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.*, **11**, 2050–2069, https://doi.org/10.1029/2019MS001639.

Mantua, N. J., and S. R. Hare, 2002: The Pacific decadal oscillation. *J. Oceanogr.*, **58**, 35–44, https://doi.org/10.1023/A:1015820616384.

——, ——, Y. Zhang, J. M. Wallace, and R. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079, https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2.

McKinnon, K. A., and C. Deser, 2018: Internal variability and regional climate trends in an observational large ensemble. *J. Climate*, **31**, 6783–6802, https://doi.org/10.1175/JCLI-D-17-0901.1.

McPhaden, M. J., S. E. Zebiak, and M. H. Glantz, 2006: ENSO as an integrating concept in Earth science. *Science*, **314**, 1740–1745, https://doi.org/10.1126/science.1132588.

Meehl, G. A., and A. Hu, 2006: Megadroughts in the Indian monsoon region and Southwest North America and a mechanism for associated multidecadal Pacific sea surface temperature anomalies. *J. Climate*, **19**, 1605–1623, https://doi.org/10.1175/JCLI3675.1.

——, C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.*, **88**, 1383–1394, https://doi.org/10.1175/BAMS-88-9-1383.

——, A. Hu, and H. Teng, 2016: Initialized decadal prediction for transition to positive phase of the interdecadal Pacific oscillation. *Nat. Commun.*, **7**, 11718, https://doi.org/10.1038/ncomms11718.

Newman, M., and Coauthors, 2016: The Pacific decadal oscillation, revisited. *J. Climate*, **29**, 4399–4427, https://doi.org/10.1175/JCLI-D-15-0508.1.

Okumura, Y. M., and C. Deser, 2010: Asymmetry in the duration of El Niño and La Niña. *J. Climate*, **23**, 5826–5843, https://doi.org/10.1175/2010JCLI3592.1.

Oshima, K., and Y. Tanimoto, 2009: An evaluation of reproducibility of the Pacific decadal oscillation in the CMIP3 simulations. *J. Meteor. Soc. Japan*, **87**, 755–770, https://doi.org/10.2151/jmsj.87.755.

Otterå, O. H., M. Bentsen, H. Drange, and L. Suo, 2010: External forcing as a metronome for Atlantic multidecadal variability. *Nat. Geosci.*, **3**, 688–694, https://doi.org/10.1038/ngeo955.

Phillips, A. S., C. Deser, and J. Fasullo, 2014: Evaluating modes of variability in climate models. *Eos, Trans. Amer. Geophys. Union*, **95**, 453–455, https://doi.org/10.1002/2014EO490002.

Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *J. Climate*, **29**, 4083–4097, https://doi.org/10.1175/JCLI-D-15-0556.1.

Power, S. B., T. Casey, C. Folland, A. Colman, and V. Metha, 1999: Interdecadal modulation of the impact of ENSO on Australia. *Climate Dyn.*, **15**, 319–324, https://doi.org/10.1007/s003820050284.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Robertson, A. W., A. Kumar, M. Peña, and F. Vitart, 2015: Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, **96** (3), ES49–ES53, https://doi.org/10.1175/BAMS-D-14-00139.1.

Rohde, R., and Coauthors, 2013: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinfor. Geostat. Overview*, **1** (1), https://doi.org/10.4172/2327-4581.1000101.

Santer, B. D., and Coauthors, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. USA*, **106**, 14 778–14 783, https://doi.org/10.1073/pnas.0901736106.

Schneider, D. P., and R. L. Fogt, 2018: Artifacts in century-length atmospheric and coupled reanalyses over Antarctica due to historical data availability. *Geophys. Res. Lett.*, **45**, 964–973, https://doi.org/10.1002/2017GL076226.

Schurer, A. P., G. C. Hegerl, M. E. Mann, S. F. Tett, and S. J. Phipps, 2013: Separating forced from chaotic climate variability over the past millennium. *J. Climate*, **26**, 6954–6973, https://doi.org/10.1175/JCLI-D-12-00826.1.

Simpson, I. R., S. G. Yeager, K. A. McKinnon, and C. Deser, 2019: Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nat. Geosci.*, **12**, 613–619, https://doi.org/10.1038/s41561-019-0391-x.

Stoner, K., K. Hayhoe, and D. J. Wuebbles, 2009: Assessing general circulation model simulations of atmospheric teleconnection patterns. *J. Climate*, **22**, 4348–4372, https://doi.org/10.1175/2009JCLI2577.1.

Stott, P. A., N. P. Gillett, G. C. Hegerl, D. J. Karoly, D. A. Stone, X. Zhang, and F. Zwiers, 2010: Detection and attribution of climate change: A regional perspective. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 192–211, https://doi.org/10.1002/WCC.34.

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

Thoma, M., R. J. Greatbatch, C. Kadow, and R. Gerdes, 2015: Decadal hindcasts initialized using observed surface wind stress: Evaluation and prediction out to 2024. *Geophys. Res. Lett.*, **42**, 6454–6461, https://doi.org/10.1002/2015GL064833.

Thompson, D. W. J., and J. M. Wallace, 2000: Annular modes in the extratropical circulation. Part I: Month-to-month variability. *J. Climate*, **13**, 1000–1016, https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771–2777, https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2.

——, and T. J. Hoar, 1997: El Niño and climate change. *Geophys. Res. Lett.*, **24**, 3057–3060, https://doi.org/10.1029/97GL03092.

——, and D. J. Shea, 2006: Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.*, **33**, L12704, https://doi.org/10.1029/2006GL026894.

——, and J. T. Fasullo, 2013: An apparent hiatus in global warming? *Earth's Future*, **1**, 19–32, https://doi.org/10.1002/2013EF000165.

Vannière, B., E. Guilyardi, G. Madec, F. J. Doblas-Reyes, and S. Woolnough, 2013: Using seasonal hindcasts to understand the origin of the equatorial cold tongue bias in CGCMs and its impact on ENSO. *Climate Dyn.*, **40**, 963–981, https://doi.org/10.1007/s00382-012-1429-6.

Vimont, D. J., 2005: The contribution of the interannual ENSO cycle to the spatial pattern of decadal ENSO-like variability. *J. Climate*, **18**, 2080–2092, https://doi.org/10.1175/JCLI3365.1.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wallace, J. M., C. Deser, B. V. Smoliak, and A. S. Phillips, 2016: Attribution of climate change in the presence of internal variability. *Climate Change: Multidecadal and Beyond*, C.-P. Chang, M. Ghil, and J. M. Wallace, Eds., World Scientific Series on Asia-Pacific Weather and Climate, Vol. 6, World Scientific, 1–29, https://doi.org/10.1142/9789814579933_0001.

Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations? *Geophys. Res. Lett.*, **36**, L12702, doi:10.1029/2009GL038710.

Yang, Y., S.-P. Xie, L. Wu, Y. Kosaka, and J. Li, 2018: ENSO forced and local variability of north tropical Atlantic SST: Model simulations and biases. *Climate Dyn.*, **51**, 4511–4524, https://doi.org/10.1007/s00382-017-3679-9.

Zhang, T., and D.-Z. Sun, 2014: ENSO asymmetry in CMIP5 models. *J. Climate*, **27**, 4070–4093, https://doi.org/10.1175/JCLI-D-13-00454.1.