Evaluation of Leading Modes of Climate Variability in the CMIP Archives

Authors: John T. Fasullo*, A. S. Phillips, C. Deser

*National Center for Atmospheric Research

*corresponding author: J. Fasullo, fasullo@ucar.edu

Submitted to Journal of Climate
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 timescales, 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 CMIP Versions 3, 5 and 6 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 the El Niño / Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO), and the Northern (NAM) and Southern (SAM) Annular Modes.

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 multi-decadal projection (Deser et al. 2012a, 2014, 2017; Kumar and Ganguly, 2018; Dai and Bloecher, 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 timescales, 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 et al. 2009), and the Southern and Northern Annular Modes (SAM, Thompson and Wallace, 2000; NAM; Calttiaux et al. 2013, Gillett and Fyfe, 2013). These modes exhibit a peak in their spectra on seasonal (Lee and Black, 2015) timescales 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 the 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 et al. 2016) and the related Interdecadal Pacific Oscillation (IPO, Power et al., 1999) represents 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).

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, Deser et al. 2019). 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 the Coupled Model Intercomparison Project, Version 6 (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 (Versions 3 and 5; 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 data sets. 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 (CESM1-LE: Kay et al. 2015) to discriminate between structural vs. sampling uncertainty in models’ representation of these modes. 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 6, 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 7.

2. Methods, Data and Mode Selection
The Climate Variability Diagnostics Package (CVDP)

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 CMIP 3, 5 and 6 archives in addition to the CESM1-LE, 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 for example: 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 Nino3.4 SST anomaly index in December (Trenberth and Hoar 1997, Deser et al. 2012), 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). 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.
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, Rhode 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 from the European Centre for Medium Range Weather Forecasts (ECMWF) Twentieth Century Reanalysis (ERA20C, Poli et al. 2016), which is combined here with estimates from ERA Interim after 1979 (Berrisford et al. 2009). Fields are also used from the coupled 20^ 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 1900-2018 is used to document modes in observations using the CVDP, though sensitivity to the precise selection of timeframe is small.

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. 2017) 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 range of pattern correlation scores estimated for modes in the CESM Large Ensemble (CESM1-LE, Kay et al. 2015) is compared with the spread of scores across the CMIP5 multi-model ensemble in Figure 1. Observational uncertainty in the modes is also indicated (Figure 1, red dots). As only a single model is used to generate the CESM1-LE, the spread in scores 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 range of noise in the CESM1-LE is 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 CESM1-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.

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), both measures of ENSO’s spatio-temporal structure will be assessed. In the Pacific, the 
S/N intrinsic to the Interdecadal Pacific Oscillation (Meehl and Hu 2007) 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.

Lastly, to differentiate the dominant spatial patterns of model bias, a principal component 
(PC) analysis is used whereby the arrays of spatial pattern biases (lon x lat) across models is 
decomposed for its empirical orthogonal functions (EOFs). Both the first two EOFs and mean 
PC values, sorted and averaged across terciles for each CMIP generation, are shown in summary 
plots, where terciles are based on the sorted weights (upper, middle, and lower thirds) of each 
PC. 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 over weighting 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 23 climate models for CMIP3, 28 models for CMIP5, and 26 models for CMIP6 (for output available at the time of this study). 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 CESM1-LE. In general, the magnitudes of the CESM1-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 non-negligible 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 modelling these interactions with global coupled models has been a long-standing challenge (Wittenberg 2009, Bellenger et al. 2015). 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 Figure 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. Composite biases in CMIP3 and CMIP6 models (Figs. 2b, c) 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 by comparing mean bias patterns in CMIP3 (Fig. 2b) versus those in CMIP6 (Fig. 2c).

The dominant patterns of model bias are summarized by the leading PCs (Fig. 2d) and EOFs (Figs. 2e and f) across terciles of the CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) generations based on their distance from observations in the PC1/2 space. 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 the pattern of overall bias in CMIP3 (Fig. 2b). Teleconnection weakness is again evident in the leading EOF of bias (Fig. 2e), which correlates negatively with the observed pattern (-0.94), 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 from those observed in some respects, such as the negative biases that extend into the northwestern Pacific Ocean in their mean biases (Fig. 2b, c) 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 the lowest model tercile. 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 the CMIP6 middle and lower tercile values deviate further from observations than corresponding CMIP3 and CMIP5 values.

The fidelity of El Niño’s spatio-temporal structure in the equatorial Pacific is shown in Figure 3, which shows the composite evolution of equatorial (5N-5S) 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 135W, and extend westward to 170E. 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 relative to the observed structure are qualitatively similar in the CMIP3 (Fig. 3b) and CMIP6 archives (Fig. 3c). 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, Figure 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. The EOF decomposition of bias (Fig. d-f) corroborates this improvement. 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. 3B) to CMIP6 (Fig. 3C), and the proximity with which CMIP6 PC terciles lie to observations relative to CMIP3 and CMIP5 terciles.

A similar analysis of La Niña is presented in Figure 4. In nature, multi-year 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). Models have
difficulty simulating this persistence, as evident from their warm biases late in Year 1 in both CMIP3 (Fig. 4B) and CMIP6 (Fig. 4C) composite means. Cold biases in the western Pacific Ocean also characterize both CMIP3 and CMIP6 models, revealing the excessive westward extent of ENSO also seen for El Niño, though a reduction of bias on average is evident from CMIP3 to CMIP6. EOF decomposition of La Niña composites (Figs. 4d-f) 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 intra-model 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 (b,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 multi-year 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–Oyashio currents. 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 dateline. Weak
positive values are also evident in the tropical Indian and Atlantic Ocean basins.

Biases in CMIP models 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
model bias are shown Figure 5d-f, 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=-0.88) and explains a significant fraction of the variance in bias across models (32%).

The spatial structure of EOF2 highlights biases 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, 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 the highest scoring tercile, where CMIP3 lies closer
to observations than does CMIP5.
Large seasonal variability is inherent to SLP in the extratropics. In the NH, subtropical anticyclones dominate during summer and weaken and move equatorward 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 Figure 6A. Out of phase centers of action reside at approximately 40N and 65N, just east of the Azores and west of Iceland, respectively, with strong zonal coherence in the pattern. The mean model biases for CMIP3 and CMIP6 are shown in Figs. 6b,c and they suggest systematic weakness in both the Azores High and the Icelandic Low (i.e., the biases are opposite in sign to the observed anomalies). The leading EOFs differentiating simulated patterns relate mainly to the strength (EOF1, Fig. 6e) and eastward tilt of the pattern (Fig. 6f), 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. 6d), though improvement has not been monotonic across CMIP generations and significant systematic amplitude errors remain. Persisting biases are also clearly evident in the CMIP mean (Fig. 6b,c).

Observed features of the NAM (Figure 7A), include centers of action across the Arctic, in the Aleutian Low, in the Azores High that resemble the NAO (Fig. 6; see also Feldstein and Franzke, 2006). The mean bias patterns in CMIP3 (Fig. 7c) and CMIP6 (Fig. 7d) 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 (Fig. 4e), while EOF2 is dominated by a center-of-action over and downstream from the Aleutian Low (Fig. 7f). The PC weights (Fig. 7b)
demonstrate that model biases are systematically positive for both PC1 and PC2, indicative of
the weakness of the simulated NAM pattern across models (PC1/EOF1) and biases in the
weighting between the Aleutian low and Icelandic high (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 variability (Figure 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). A particularly strong center of action exists in the Amundsen Sea. As with other
extratropical modes, the mean patterns of bias and the leading patterns resulting from the PC
analysis (Figs. 8b-f) relate to weakness in simulated patterns, as the leading EOF of bias (Fig.
8C) correlates strongly with the mean pattern (r=0.88) and explains a majority of the variability
across models (61%). Much weaker is EOF2, which relates to details in the meridional extent
simulated anomalies and structure in the Tasman Sea (EOF2, Fig. 8C), explaining only 10% of
the inter-model variance. Modest reductions in these biases are evident in mean PC magnitudes
across model terciles, with CMIP6 scoring much better in the upper tercile for PC1 and across all
terciles for PC2. Nonetheless, significant systematic bias remains, particularly in PC1 where
CMIP6 values demonstrate a significant contribution of the EOF1 bias pattern.

7. 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 Figure 9. An
average of the full set of scored metrics (see Phillips et al. 2014) is used to compute the overall score (Fig. 9a). To provide a more complete sampling of histogram distributions, all ensemble members are included in this analysis, though the main results are unchanged by considering only one member per modeling center. Also shown are the corresponding ranges of scores from CESM-LE, 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 generally non-gaussian, with a tail of low scoring models being evident, particularly for CMIP3 where overall scores fell below 0.7 for some models. 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). Differences within and across CMIP generations are considerably larger than internal variability in the CESM1-LE. 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 CESM1-LE. Corresponding scores for ENSO DJF TAS composites (Fig. 9d) are 0.53, 0.59, and 0.67. Spatio-temporal patterns for El Niño (Fig. 9e) and La Niña (Fig. 9f) hovmöölers 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 hovmöölers 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 CESM1-LE. 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). 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.

As a complement to pattern correlation scores, the spectra of ENSO and the PDO summarized across various frequency bands are shown in Figure 10, where the full distribution of each CMIP generation is summarized by its minimum and maximum values and interquartile range. Also shown is the corresponding spread in power from the CESM-LE to provide context for apparent model bias. For ENSO, the power on timescales less than 2.5 years is less than observed for all CMIP generations and simulations and the estimated influence of internal variability based on the CESM-LE range is small. In the 2.5 to 6-year band, the 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 that many models may overestimate power in this band. This possibility is further supported by the relatively narrow range of the CESM-LE. Though 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 range, 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 yrs., >10 yrs.), a progressive increase in 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 yrs., agreement between observations and CMIP is greatest for CMIP6 as observations tend to show more power than the interquartile range of CMIP3. Spectral agreement of the PDO with observations (Fig. 10b) in many ways mirrors that of ENSO, with deficient power for short periods across CMIP generations, reasonable agreement in the 2.5-6 and 6-10 year bands. 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 is 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., 2019) 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 year band and a reduction at periods greater than 10 years.

8. **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. Modes have been selected for which the signal of structural uncertainty across models is relatively large compared to intrinsic noise due to sampling fluctuations, and consideration has been given to both internal variability and observational uncertainty in evaluation of model bias. 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 intra-model 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 inter-model 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 major internal modes of variability, albeit with reduced but persisting biases, making climate models increasingly suitable for attributing past changes, predicting future climate, and estimating associated uncertainties.
Data Availability

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).

Acknowledgements

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. The efforts of Dr. Fasullo in this work were partially supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological & Environmental Research (BER) via National Science Foundation IA 1844590. The efforts of Dr. Fasullo in this work were also supported in part by NSF Award #AGS-1419571.
References:


10.1007/s00382-010-0977-x.

Deser, C., A.S. Phillips, R. A. Tomas, Y. Okumura, M. A. Alexander, A. Capotondi, J. D. Scott,
Y. -O. Kwon, and M. Ohba (2012b) ENSO and Pacific Decadal Variability in Community
Climate System Model Version 4. *J. Climate*, **25**, 2622-2651, doi: 10.1175/JCLI-D-11-
00301.1.

climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**(6),

Deser, C., 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

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, doi: 10.1175/JCLI-D-16-
0844.1.

Deser, C., F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C.
Frankignoul, J. C. Fyfe, D. E. Horton, J. E. Kay, R. Knutti, N. S. Lovenduski, J. Marotzke,
http://www.cgd.ucar.edu/staff/cdeser/


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

Gettelman, A., C. Hannay, J.T. Bacmeister, R.B. Neale, A.G. Pendergrass, G. Danabasoglu, G.,
... M.J. Mills, (2019) High climate sensitivity in the Community Earth System Model


Gleckler, P. C. Doutriaux, P.J. Durack, K.E. Taylor, Y. Zhang, D.N. Williams, ... and J.
https://doi.org/10.1029/2016eo051663.


Guo, R., C. Deser, L. Terray and F. Lehner, 2019: Human influence on winter precipitation

Huang, B., Peter W. Thorne, et. al. (2017) Extended Reconstructed Sea Surface Temperature
version 5 (ERSSTv5), Upgrades, validations, and intercomparisons. *J. Climate*, doi:
10.1175/JCLI-D-16-0836.1.


detection and attribution to the characterization of internal climate variability. *J. Climate*,
27(10), 3477-3491, doi: 10.1175/JCLI-D-12-00622.1.

Kay, J. E., C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, ... and M. Holland (2015). The
Community Earth System Model (CESM) large ensemble project: A community resource for
studying climate change in the presence of internal climate variability. *Bulletin of the

Kumar, D., and A.R. Ganguly (2018) Intercomparison of model response and internal variability
3914-4.

Laloyaux, P., E. de Boisseson, M. Balmaseda, J.R. Bidlot, S. Broennimann, R. Buizza, ... and
Kosaka (2018) CERA-20C: A coupled reanalysis of the Twentieth Century. *J. Advances in

annular mode in CMIP5 simulations. *J. Climate*, 28(1), 86-107, doi: 10.1175/JCLI-D-13-
00570.1.

agreement between observed and simulated extratropical modes of interannual variability.

Lehner, F., C. Deser, I. R. Simpson and L Terray, 2018: Attributing the US Southwest’s recent

(2019) Improvements in the uncertainty model in the Goddard Institute for Space Studies


Table 1: Simulations considered in this study by CMIP archive.

<table>
<thead>
<tr>
<th>CMIP3</th>
<th>CMIP5</th>
<th>CMIP6</th>
</tr>
</thead>
<tbody>
<tr>
<td>bccr_bcm2_0</td>
<td>ACCESS1-3</td>
<td>BCC-CSM2-MR</td>
</tr>
<tr>
<td>cccma_cgcm3_1</td>
<td>bcc-csm1-1</td>
<td>CAMS-CSM1-0</td>
</tr>
<tr>
<td>cnrm_cm3</td>
<td>BNU-ESM</td>
<td>CanESM5</td>
</tr>
<tr>
<td>csiro_mk3_5</td>
<td>CanESM2</td>
<td>CNRM-CM5-6-1</td>
</tr>
<tr>
<td>IDL&gt; for i=0,15 do print,mods3(i)</td>
<td>CESM1-BGC</td>
<td>E3SM-1-0</td>
</tr>
<tr>
<td>bccr_bcm2_0</td>
<td>CMCC-CM</td>
<td>EC-Earth3</td>
</tr>
<tr>
<td>cccma_cgcm3_1</td>
<td>CNRM-CM5</td>
<td>IPSL-CM6A-LR</td>
</tr>
<tr>
<td>cnrm_cm3</td>
<td>CNRM-CM5-2</td>
<td>MIROC6</td>
</tr>
<tr>
<td>csiro_mk3_5</td>
<td>CSIRO-Mk3-6-0</td>
<td>HadGEM3-GC31-LL</td>
</tr>
<tr>
<td>gfdl_cm2_1</td>
<td>FGOALS-g2</td>
<td>MRI-ESM2-0</td>
</tr>
<tr>
<td>giss_model_e_h</td>
<td>FIO-ESM</td>
<td>GISS-E2-1-G</td>
</tr>
<tr>
<td>iap_fgoals1_0_g</td>
<td>GFDL-CM3</td>
<td>GFDL-ESM4</td>
</tr>
<tr>
<td>ingv_echam4</td>
<td>GFDL-ESM2G</td>
<td>CESM2</td>
</tr>
<tr>
<td>mpi_echam5</td>
<td>GFDL-ESM2M</td>
<td>CESM2-WACCM</td>
</tr>
<tr>
<td>immcm3_0</td>
<td>GISS-E2-H</td>
<td>NorCPM1</td>
</tr>
<tr>
<td>ipsl_cm4</td>
<td>GISS-E2-H-CC</td>
<td>GFDL-ESM4</td>
</tr>
<tr>
<td>miroc3_2_hires</td>
<td>GISS-E2-R</td>
<td>NEM3</td>
</tr>
<tr>
<td>miub_echo_g</td>
<td>GISS-E2-R-CC</td>
<td>FGOALS-g3</td>
</tr>
<tr>
<td>mri_egcm2_3_2a</td>
<td>HadCM3</td>
<td></td>
</tr>
<tr>
<td>ncar_cesm3_0</td>
<td>HadGEM2-AO</td>
<td></td>
</tr>
<tr>
<td>ukmo_hadgem1</td>
<td>immcm4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPSL-CM5A-LR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPSL-CM5A-MR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPSL-CM5B-LR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MIROC5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPI-ESM-MR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MPI-ESM-P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MRI-CGCM3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NorESM1-M</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Mean pattern correlations (vertical bars) between simulations (1920-2005) and observations (1900-2018) for members of the CESM1-LE along with their 2-standard deviation (2σ) range (black whiskers) and the corresponding 2σ range for the 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 CESM1-LE (i.e. noise, N) is on par with the CMIP5 range, which is comprised of both model structural error (S) and internal variability (N), suggests a weak S/N for mode evaluation.
Figure 2: ENSO composites of DJF sea level pressure for 1900 through 2017 based on (a) ERA-20C/ERA-I in observations. Zonal mean values are also indicated over ocean (blue), land (red), and combined (black). (b) As in (a) but for mean bias in CMIP3 and (c) CMIP6. (d) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles.
Figure 3: Hovmoller diagrams (longitude vs. time) of (a) composite equatorial SST anomalies during El Niño events based on ERSSTv5 observational estimates. (b) As in (a) but for mean bias in CMIP3 and (c) CMIP6. (d) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles. Vertical axes extend from January of Year 0 to May of Year 2.
Figure 4: As in Fig. 3 except for La Niña.
Figure 5 a) Spatial pattern of the observed Pacific Decadal Oscillation (based on NOAA’s ERSSTv5 for 1900 to 2018). Zonal mean values are also indicated over ocean (black). (b) As in (a) but for mean PDO pattern bias in CMIP3 and (c) CMIP6. (d) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles.
Figure 6 Observed (a) pattern of the Northern Atlantic Oscillation (based on ERA20C/ERAI from 1950 to 2018). (b) As in (a) but for mean bias in CMIP3 and (c) CMIP6. (d) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles.
Figure 7: Observed (a) pattern of the Northern Annular Mode in DJF (based on ERA20C/ERA1 from 1950 to 2018). (b) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (c) As in (a) but for mean bias in CMIP3 and (d) CMIP6. (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles.
Figure 8 a) Observed pattern of the Southern Annular Mode in DJF (based on CERA20C/ERAI from 1950 to 2018). (b) The 1st and 2nd PC weights of model bias for terciles of CMIP3 (green), CMIP5 (aqua), and CMIP6 (blue) and observations (red, see methods). (c) As in (a) but for mean bias in CMIP3 and (d) CMIP6. (e) The first and (f) second EOFs of model bias (including all CMIP generations); pattern correlations of each against observations and CMIP mean bias are indicated in the panel titles.
Figure 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 hovmöëllers, and DJF (g) NAM and (h) SAM. Horizontal whisker lines on each plot correspond to the 10th to 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 CESM-LE are shown (red).
Figure 10 Whisker plots of the band-averaged power spectra in CMIP3 (green), CMIP5 (aqua), and CMIP5 (blue) for a) Niño3.4 SST anomalies and b) the PDO index. Thin colored lines span the full ensemble range while thick lines span the 25 to 75 percentile range. The equivalent spans from CESM1-LE are also shown (red). Two CMIP3 simulations greatly exceed the range for the 2.5-6 yr band (CNRM-CM3, >200; IAP-FGOALS, >600). Units are °C cycles mo⁻¹.