Representation of Modes of Variability in Six U.S. Climate Models

Article *in* Journal of Climate · June 2020

DOI: 10.1175/JCLI-D-19-0956.1

Representation of Modes of Variability in 6 U.S. Climate Models

Clara Orbe*

NASA Goddard Institute for Space Studies, New York, NY

Luke Van Roekel

T-3 Solid Mechanics and Fluid Dynamics, Los Alamos National Laboratory, Los Alamos, NM

Ángel F. Adames

University of Michigan, Ann Arbor, MI

Amin Dezfuli

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, and Science Systems and Applications, Inc., Lanham, Maryland

John Fasullo

National Center for Atmospheric Research, Boulder, CO

Peter J. Gleckler

Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA

Jiwoo Lee

Early Online Release: This preliminary version has been accepted for publication in *Journal of Climate*, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-19-0956.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA

Wei Li

IMSG at Environmental Modeling Center, National Centers for Environmental Prediction (NCEP)/National Weather Service (NWS)/National Oceanic and Atmospheric Administration (NOAA), College Park, MD

Larissa Nazarenko

CCSR, Columbia University, New York, NY and NASA Goddard Institute for Space Studies, New York, NY

Gavin A. Schmidt

NASA Goddard Institute for Space Studies, New York, NY

Kenneth R. Sperber

Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA

Ming Zhao

Geophysical Fluid Dynamics Lab, Princeton, NJ

**Corresponding author address:* Clara Orbe, NASA Goddard Institute for Space Studies, 2880 Broadway New York, NY 10025

E-mail: clara.orbe@nasa.gov

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

ABSTRACT

We compare the performance of several modes of variability across six US climate modeling groups, with a focus on identifying robust improvements in recent models (including those participating in the Coupled Model Intercomparison Project (CMIP) Phase 6) compared to previous versions. In particular, we examine the representation of the Madden-Julian Oscillation (MJO), the El Niño/Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), the Quasi-Biennial Oscillation (QBO) in the tropical stratosphere and the dominant modes of extra-tropical variability, including the Southern Annular Mode (SAM), the Northern Annular Mode (NAM) (and the closely related North Atlantic Oscillation (NAO)), and the Pacific-North American Pattern (PNA). Where feasible, we explore the processes driving these improvements through the use of "intermediary" experiments that utilize model versions between CMIP3/5 and CMIP6 as well as targeted sensitivity experiments in which individual modeling parameters are altered. We find clear and systematic improvements in the MJO and QBO and in the teleconnection patterns associated with the PDO and ENSO. Some gains arise from better process representation, while others (e.g. the QBO) from higher resolution that allows for a greater range of interactions. Our results demonstrate that the incremental development processes in multiple climate model groups lead to more realistic simulations over time.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

1. Introduction

Around the world, and certainly in the United States, climate and weather model groups have been upgrading their codes for operational purposes and/or for contributions to new international projects (such as the Coupled Model Intercomparison Project Phase 6 - CMIP6 (Eyring et al. 2016)). Preliminary analysis of these new model versions – both published (Del Genio et al. 2015; Rind et al. 2014; Golaz et al. 2019; Danabasoglu et al. 2020) and unpublished – has shown some remarkable increases in the fidelity of representation of important modes of variability. Most notably, representations of the Madden-Julian Oscillation (MJO), the Quasi-Biennial Oscillation (QBO) and patterns associated with the El Niño/Southern Oscillation (ENSO) have greatly improved relative to model versions of only a few years ago.

This raises important scientific questions: what were the processes involved in this increase in skill? What is the balance between increases in vertical or horizontal resolution versus new physics or better tuning? Can we better predict the impact of climate change on these modes or interactions between them? The salience of these questions is increased by the upcoming IPCC 6$^{\text{th}}$ Assessment report, which will report on model evaluations and projections in 2021.

This paper reports on an in-depth comparison across all six US climate modeling centers (Table 1). Compared to broader comparisons across the CMIP archive this study has an advantage in that we are able to dig deeper into intermediate versions of models that have not been included in CMIP6 and encompasses two groups that do not contribute to CMIP since they are focused on shorter prediction windows (weather to sub-seasonal variability). Intermediate model versions are analyzed for select modes for which the simulation duration is of sufficient length to characterize the mode in question. Within these analyses, we focus on a) using consistent and robust diagnostics across all modes and models (atmospheric and coupled), b) attempting to track down

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

the reasons for model skill improvements, and c) identifying continuing and persistent systematic biases.

a. History

The inclusion of dynamic variability in the climate system has been the goal of general circulation modeling since the beginning (e.g. Phillips 1956; Manabe and Bryan 1969; Hansen et al. 1983). Some modes, which are dependent on the largest scale features of the synoptic circulation, such as the North Atlantic Oscillation (NAO) (or the closely related Northern Annular Mode (NAM)), and the Southern Annular Mode (SAM), have been represented in all models. For other modes of variability, however, it has long been recognized that they rely on wave motions or specific climate features that were not resolvable using the horizontal and vertical discretizations and/or configurations achievable in early generations of models.

Developments since then, and particularly within the CMIP process, have clearly identified necessary (though not sufficient) requirements for models to realistically exhibit specific modes of variability. An obvious example is the simulation of ENSO which, at minimum, requires sufficient resolution to resolve the equatorial Kelvin wave guide in the Pacific ocean (Kang and An 1998). Models with ocean components without sufficient resolution will exhibit tropical variability, but the magnitude and transient structure of that variability will not be realistic (Russell et al. 1995; Clement et al. 2011). Similarly, the capacity to produce a QBO relies on sufficient vertical resolution in the lower stratosphere ($\sim$ 500 m) (Geller et al. 2016).

For some modes though, resolution plays little role. For example, the inability of models to produce an MJO had long been a puzzle until the early successes of Inness et al. (2003), among others. For this mode, the key issues revolved around simulating the processes of convection and the tropical boundary layer sufficiently well to prevent excessive mixing (e.g., Kim et al. 2012).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

This is similar to representations of the Pacific Decadal Oscillation (PDO), whose status remains ambiguous - is it the decadal expression of ENSO, or something driven independently? Or is it mainly a statistical description (Newman et al. 2016)? There are no obvious resolution-based reasons to expect (or not) the presence of a realistic PDO, and yet, model representations have historically been very diverse.

The lack of any obvious barriers to simulations of these last two modes has led some to speculate that new physics or radical new approaches might be needed to improve their representation in models. Meanwhile, the 'normal' development of general circulation models (GCMs) (in the Kuhnian sense) has proceeded apace. The extent to which significant improvements have been made will be a testament (or not) to our increasing understanding of the climate system.

b. Scope

There have been far too many modes of variability identified in the literature for our analysis to be comprehensive, so our focus will be on the principal, well-recognized modes that have been robustly identified in the modern climate record. In the tropics, this includes coupled modes like ENSO, the PDO, and the MJO as well as the primarily atmospheric QBO in the tropical stratosphere. In the extra-tropics, this includes the NAO/NAM, SAM, and Pacific-North American Pattern (PNA) patterns. While we consider all seasons we focus primarily on those during which these modes are dominant (i.e. December-January-February (DJF) for the NAO/NAM, PNA, and ENSO; and June-July-August (JJA) and DJF for the SAM).

Note that, while here we distinguish between the NAM and the NAO, there have been different perspectives on their relationship (see Thompson et al. (2003) and references therein). Various studies suggest they are indistinguishable (Wallace 2000; Feldstein and Franzke 2006; Dai and Tan 2017), connected (e.g., Thompson and Wallace (1998); Gong et al. (2002); Cohen and Barlow

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

(2005); Cohen et al. (2005); Stephenson et al. (2006); Rivière and Drouard (2015); Song (2019)), independent (e.g., Ambaum et al. (2001); Deser (2000)), or mixed by season (Rogers and McHugh 2002). In this paper, however, we diagnose the NAM and NAO separately in order to provide both regional and hemispheric perspectives on northern extratropical variability. For ENSO and the PDO, in addition to the spatial patterns and teleconnections we also consider spectral behaviour.

2. Models and Analysis

We primarily use coupled atmosphere-ocean simulations together with a few historical atmosphere-only (AMIP) (Gates et al. 1999) simulations. For those models submitting to CMIP6 (e.g, CESM, ModelE, E3SM) these experiments correspond to the "Historical" experiment (Eyring et al. 2016). Other simulations analyzed were obtained directly from modeling groups (e.g. the Goddard Earth Observing System (GEOS) and Global Ensemble Forecast System (GEFS) sub-seasonal forecasts). The type and number of ensemble members per model submission considered here varies, as described in more detail below and in Table 3. We examine both current versions of the model as well as prior versions and selected development versions when available and relevant.

a. Model Descriptions

The salient details for the models used in this study (e.g., model components, resolution, parameterizations) are summarized in Table 2. Here we briefly describe the models utilized in this study, directing readers who seek full details to the references described herein.

Four versions of the Community Earth System Model (CESM) were used in this analysis: CESM1 (using the Community Atmosphere Model (CAM) version 5), CESM1 (using the Whole Atmosphere Community Climate Model (WACCM) Version 5), CESM2 (using CAM6), and CESM2 (using WACCM6), which are documented in Hurrell et al. (2013a), Mills et al. (2017),

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Danabasoglu et al. (2019), and Gettelman et al. (2019), respectively. The U.S. Department of Energy (DOE) Energy Exascale Earth System Model (E3SMv1) (Golaz et al. 2019) branched from the CESM1 model, but has evolved significantly (Rasch et al. 2019; Xie et al. 2018). Five versions of the NASA Goddard Institute for Space Studies (GISS) ModelE were included as well, two from CMIP5 (E2-R and E2-H (Schmidt et al. 2014)) and three CMIP6 versions (E2.1-G and E2.1-H (Kelley et al. 2020) and E2.2-G (Rind et al. 2020)). The -G (-R in CMIP5) and -H indicate two different ocean models. Finally, three Geophysical Fluid Dynamics Laboratory (GFDL) models were used: CM3 (Griffies et al. 2011; Donner et al. 2011), CM4 (Held et al. 2019; Zhao et al. 2018a,b) and ESM4 (Dunne et al. 2020). We also reference simulations from the suite of US models that were used in CMIP3 (Meehl et al. 2007) as a baseline for the changes seen in later CMIP iterations.

In addition to the CMIP models, we also consider an ensemble of ten free-running integrations produced by the NASA Global Modeling and Assimilation Office (GMAO) using GEOS Version-5 (GEOS-5, Molod et al. (2015)) and an ensemble of forecasts from two operational sub-seasonal forecasting modeling groups: the GMAO sub-seasonal 45-day long forecasts (Molod et al. 2020) and the Global Environmental Forecast System (GEFS) SubX forecasts from the National Centers for Environmental Prediction (NCEP) (Zhu et al. 2018). The GMAO forecasts are fully coupled, whereas the GEFS are uncoupled.

While each modeling center has different development targets, we note a few relevant developments common to models considered in this study. First, most models, particularly those participating in CMIP6, have increased the height of the model top, as well as the vertical resolution. This appears to play a critical role in the fidelity of the QBO (Geller et al. 2016, see Section 3d). Second, there have been improvements to the models' representation of gravity wave drag, which in turn have also improved simulation of the QBO. This has come by way of improved

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

parameterizations (e.g., CESM2(WACCM2), GISS E2.2, CM4) and improved tuning of current parameterizations (e.g. E3SMv1-MODGW). Third, the treatment of shallow convection has improved in E3SM, CESM2, GISS E2.1, and CM4 including new parameterizations and tunings. These improvements positively impact the simulated MJO (Section 3a).

The specific experiments submitted by each modeling center are summarized in Table 3. For most of the CMIP models, some number of ensemble members of the "Historical" experiment are analyzed. For the GMAO ensemble (hereafter M2AMIP) we have considered a ten-member ensemble of AMIP simulations initialized from meteorological fields from different dates in November 1979 using identical sea-surface temperatures and sea-ice concentrations. The 45-day NASA-GMAO sub-seasonal forecasts were initialized from the MERRA-2 and GMAO S2S-1.0 ocean analysis, respectively. An ensemble of ten forecasts were initialized at 5-day intervals during all twelve months of years 1981–2016 but only years after 1999 are considered here in order to compare fairly with the NCEP GEFS SubX forecasts, which were only available starting in 1999. For the GEFS forecasts, an 11-member ensemble of 35-day-long forecasts was used for all years spanning 1999–2016, each of which consists of one control and 10 perturbed members.

b. Intermediary Model Version Simulations

We also incorporate analysis of simulations using "intermediary" model versions that were developed between CMIP3/5 and CMIP6, as well as after an initial CMIP6 submission. While they were not originally designed as individual sensitivity experiments, we have found that these simulations contribute towards our understanding of the physical processes responsible for improved representations of different modes of variability across models.

For our analysis of the MJO, we incorporate historical coupled simulations produced using a version of GISS ModelE that represents an intermediary model tag between the CMIP5 Model

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

E2 (Schmidt et al. 2014) and the CMIP6 Model E2.1 (Kelley et al. 2020). In this model version (hereafter GISS ModelE2MJO) the main differences relative to E2 include: 1) an increase in one of the parameterization's plume entrainment rates and 2) an option to allow for re-evaporation above the cloud base. The impact of these changes on AMIP simulations of the MJO was documented in Kim et al. (2013).

In order to identify mechanisms for improved simulations of the QBO we include historical simulations produced using a version of E3SMv1 (hereafter E3SMv1-MODGWD) in which the parameterized convectively generated gravity waves were altered as described in Richter et al. (2019) (hereafter R19). In particular, two changes were made: 1) the convective fraction relating the tropospheric heating rate within convective cells to the GCM grid-box averaged heating rate was increased from 5% to 8% and 2) the efficiency with which convection generates gravity waves was decreased from a default value of 0.4 in E3SMv1 to 0.35 in E3SMv1-MODGWD. R19 showed that these two changes have significant impacts on the QBO in that model.

To further understand the influence of model tuning of clouds in simulating both tropical and extra-tropical coupled modes of variability, a sensitivity experiment conducted using CESM2 is considered, referred to here as CESM2-gamma. CESM2 utilizes the CLUBB shallow turbulence scheme. In CESM2-gamma, only the gamma coefficient, which has been identified as a critical parameter for low cloud feedback responses to climate change (Gettelman et al. 2019), is modified from the official CESM2 version. Specifically, gamma controls the width of the vertical velocity probability distribution function and exercises a strong influence over low-cloud cover.

c. Analysis Tools and Observational Products

The observational products and analysis measures we use are summarized in Table 4. In particular, the tropical and extra-tropical modes of variability examined in this study are evaluated using

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

both the Climate Variability Diagnostics Package (CVDP, Phillips et al. 2014) and the PCMDI Metrics Package (PMP, Gleckler et al. 2016). The extra-tropical modes are evaluated using both conventional Empirical Orthogonal Function (EOF) analysis in which, for example, EOF-1 in the observations is compared to EOF-1 from each of the models (Stoner et al. 2009; Phillips et al. 2014). We also utilize the Common Basis Function (CBF) approach, in which model anomalies are projected onto the observed EOF to obtain the CBF Principal Component (CBF PC; Lee et al. 2019). Using the CBF PC the model mode spatial pattern is obtained by regressing the CBF PC back onto the model anomalies. We have chosen to utilize both methods given that in the conventional approach mode swapping may preclude the relevant model mode from being compared to the observations. We find, however, that the relative performance of the models is typically consistent across the different methodologies, though as reported in Lee et al. (2019), the CBF method shows the models tend to appear more skillful, compared to the standard EOF approach.

The period of analysis for the extra-tropical modes is 1900–2005 for models and observations, for which we use both ERA 20$^{\text{th}}$ Century Reanalysis (ERA20C, Poli et al. 2016) and the NOAA 20$^{\text{th}}$ Century Reanalysis (20CR, Compo et al. 2008) for years 1900–1978 and ERA Interim (Dee et al. 2011) for 1979–present. The one exception is the SAM, which we evaluate over the period 1956–2005 since there are substantial differences during the earlier part of the 20$^{\text{th}}$ century among various observed and reanalyzed datasets (Lee et al. 2019). Furthermore, model skill for the extra-tropical modes is illustrated using Taylor Diagrams (Taylor 2001), in which the radial distance from the origin is the spatial standard deviation normalized by the observed standard deviation. The difference between the observed reference and the model statistic is the centered root mean square error (RMSE), and the azimuthal angle is the pattern correlation between the model and the reference observations. The full suite of Taylor Diagrams across all modes and seasons are too numerous to present in the main text but can be found in the online supplemental information.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

The MJO analysis is predominantly based on diagnostics performed on daily data of precipitation and 850 hPa winds over the period 1999 (for which we begin to have credible gridded precipitation observations) up to the present. The procedure follows the metrics described by Jiang et al. (2015), which are summarized here for completeness. Lag regressions of precipitation and zonal winds are obtained by projecting the daily fields to a time varying index of 20-100 day filtered precipitation along 85–95°E and 5°N/S. Time-longitude diagrams of the regression fields are obtained by averaging each lag day over latitudes spanning 15°N/S. Pattern correlations with the observations are obtained by correlating the time-longitude diagrams of models analyzed herein with precipitation from the 3b42 daily product from Tropical Rainfall Measuring Mission (TRMM) (Iguchi et al. 2000) and the 850 hPa zonal winds from ERA-5 (Hersbach and Dee 2016). A similar pattern correlation analysis is performed for the wavenumber-frequency representation of the fields, in which the signal strength (S) is defined as the ratio of the difference between the power spectrum (P) and red spectrum (R) to the power spectrum itself ($S = [P - R]/P$, where R is the red noise spectrum) (Clark et al. 2020). The calculation of the power spectrum follows that of Wheeler and Kiladis (1999), and the red noise spectrum follows the procedure of Masunaga (2007), following the guidelines outlined by Waliser et al. (2009). The East-West power ratio is calculated following the procedure outlined by Sperber and Kim (2012) as the ratio between the power spectrum of eastward- and westward-propagating zonal wavenumbers 1-5 and timescales between 20–100 days.

The MJO forecast skill among the two sub-seasonal forecast ensembles is estimated by comparing RMM indices derived from each forecast model with an RMM index obtained from ERA5 data. The RMM index for ERA5 is obtained following Wheeler and Hendon (2004) as the combined EOF of OLR, u200 and u850. For each field, the mean and seasonal cycle is removed and the fields averaged over the 15°N/S latitude belt and normalized by its zonally-averaged variance

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

before combining them into a single vector. EOF analysis is then performed on this vector of combined fields. The EOFs from ERA5 are projected onto the GEFS and GEOS-S2S OLR, u200 and u850 anomalies to obtain their respective RMM time series. This method follows those outlined by Gottschalck et al. (2010) and Vitart (2017). Bivariate correlations are calculated for the two datasets following the method described by Gottschalck et al. (2010) but extended to include correlations corresponding to each calendar month as in Molod et al. (2020). We create subsets of the RMM indices for each calendar month, and bivariate correlations are made based on each forecast day and calendar month.

Finally, evaluations of the QBO across the models are based primarily on diagnostics derived from zonal and monthly averaged zonal wind output, available for some models only at 10, 30, 50, 70 and 100 hPa, using the metrics outlined in Schenzinger et al. (2017). (For MERRA-2, M2AMIP and the GISS ModelE simulations the native vertical resolution output was used). Comparisons of models over the period 1980–2016 are made against MERRA-2, which exhibits a realistic QBO compared to observations, both in terms of its zonal winds, mean meridional circulation, and associated ozone changes (Coy et al. 2016). The lack of stratospheric data available at higher temporal resolution in the models prevents us from doing as rigorous an evaluation as has been done in the recent Stratosphere-Troposphere Processes and their Role in Climate (SPARC) Quasi-Biennial Oscillation initiative (QBOi) (Butchart et al. 2018), for which the six-hourly output required to both calculate the Transformed Eulerian Mean circulation and compare equatorial wave spectra, was available. Nonetheless, the data analyzed here does provide some insight into the state of the QBO and its representation across the models considered in this study.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

3. Results

In this section we describe the fidelity of each climate mode separately and the improvement (or lack thereof) from CMIP3/5 to CMIP6. When improvements are found we also use intermediary model version experiments in order to understand the drivers of changes in model performance.

a. Madden-Julian Oscillation

The results of our MJO analysis for the U.S. climate models are shown in Figs. 1 and 2. As a guideline, in Fig. 2 the closer the individual points in the scatter are to the grey star, the closer the simulation is to the observations (here TRMM for precipitation, and ERA-5 for the zonal winds). Overall, the five models from CMIP6 considered here exhibit enhanced eastward propagation, compared to the CMIP5 models. The amplitudes of the MJO-related winds and precipitation are also improved in CMIP6 models relative to observations. When assessing individual models, the improvements are still clearer. For example, CESM2 exhibits stronger wind anomalies compared to CESM1 as well as more coherent eastward propagation (Fig. 1c and d). Similar results are seen for the other models both in terms of precipitation and zonal wind (see Supplementary Material).

The East-West (EW) power ratio is shown in Figs. 2c,d. When compared to the CMIP5 models, the CMIP6 models exhibit an increased EW ratio that compares more closely with observations, manifest as a rightward shift in Fig. 2 in both precipitation and wind. To showcase this change in signal, Fig. 1a-b compare the signal strength of precipitation between GFDL's CMIP5 CM3 model (Fig. 1a) and CMIP6 CM4 model (Fig. 1b). The darker shading for eastward-propagating zonal wavenumbers 1-5 and timescales ranging from 20-100 days is clearly evident in CM4.

The models considered do not only exhibit a closer agreement with the TRMM measurements and ERA5 data, but also show an improved space-time spectrum of all waves. This can be seen by considering the y-axis in Fig. 2a-b, which shows the pattern correlation of the signal strength

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

of the individual models with respect to the observations. For the spectrum in precipitation, it is clear that all CMIP6 models exhibit an increased correlation relative to their corresponding CMIP5 versions. A less distinct, but nonetheless positive, improvement is also observed for the spectrum of zonal winds.

For the two subseasonal forecast ensembles analyzed in this study we examine their forecast skills by calculating their monthly bivariate correlation coefficients with respect to RMM indices derived for ERA5 (Fig. 3). Overall, the GEOS-S2S and GEFS ensembles exhibit qualitatively similar correlations. In particular, the correlation in both models decays to 0.5 near forecast day 20, consistent with the results presented in Pegion et al. (2019) and Kim et al. (2019). When analyzing the individual months some differences are observed, however. GEOS-S2S exhibits a slower decorrelation time during late boreal summer (JAS), with correlations near 0.4 up to 40 days during August, consistent with the findings in Molod et al. (2020). On the other hand, the decorrelation time is faster during November and February, when correlations are below 0.3 at forecast day 25. In contrast to the GEOS-S2S forecasts, the GEFS ensemble exhibits similar decorrelation times for nearly every month, with the notable exception of late summer (JAS), when it decorrelates faster.

Intermediary Model Version Experiments: In order to understand improvements in the representation of the MJO, we include results from an intermediary version of GISS ModelE (denoted GISS ModelE2MJO) that represents a development version between the CMIP5 E2 submission and the CMIP6 E2.1 submission. This version (yellow squares in Fig. 2) shows significant improvements over the original GISS E2 model as a result of several modifications in the convective parameterization (see Section 2b for details) that resulted in a convection scheme that is more sensitive to environmental relative humidity and a more humid mean state, both of which have been

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

shown to be consistent with improved MJO simulation (Kim et al. 2012; Del Genio et al. 2012). This result is also consistent with Zhao et al. (2018a), who show that transient variability in the tropics increases when the rate of cumulus lateral mixing and convective rain re-evaporation are increased in an entirely different model (GFDL AM4), which suggests that the mechanism for MJO improvement demonstrated here is not specific to ModelE.

b. Extra-tropical Modes

Collective skill assessments of numerous extra-tropical modes of variability have been discussed widely in the literature (i.e. Stoner et al. 2009; Phillips et al. 2014; Lee et al. 2019). Here we analyze the SAM (Figs. 4a-b), the NAM (Fig. 4c), the PNA (Fig. 4d), the NAO (Fig. 4e), and the PDO (Fig. 4f). Our analysis of the SAM, NAM, NAO and PNA are based on seasonally averaged sea-level pressure anomalies, with a focus on the dominant (winter) season, with the exception of the SAM, for which we consider the (DJF) summer season as well since its interannual variability is nearly identical to that occurring during JJA. For the PDO we use monthly anomalies of sea surface temperature.

For the case of the SAM during JJA (Fig. 5a), in all US models the SAM appears to have been better represented in CMIP3, compared to CMIP5 and CMIP6. An evaluation of the SAM during DJF (Fig. 5b), however, shows the opposite, with most CMIP6 models outperforming earlier MIP versions, with the exception of E3SMv1. Thus, while the SAM exhibits some of the most pronounced skill improvement, compared to the other extra-tropical modes, this improvement is only realized during DJF and does not apply more generally to other seasons.

Consideration of the US modelling groups one at a time affords a somewhat clearer indication that skill has improved since CMIP3. (Note that, despite its incorporation of major changes in physics (Golaz et al. 2019), the E3SMv1 model is included in the discussion among the NCAR

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

models). In particular, for the NAM during DJF, the GFDL models shows improved skill in CMIP6 compared to previous MIPs (Fig. 5c). For the NCAR and DOE models (not shown), the Taylor Diagrams indicate that E3SMv1 and CESM1 perform best, with the remaining CMIP6 models tending to be better than the other CMIP5 and CMIP3 model versions. For the GISS models (Fig. 5d), the CMIP5 version performed best.

Of the extra-tropical modes analyzed, the PNA exhibits the largest diversity in skill across the entire ensemble of MIPs and models (not shown). CMIP3 was especially problematic, with three of the models having higher order EOFs with markedly better skill than their corresponding EOF-1. This indicates that these models were not properly simulating the hierarchy of observed EOFs due to mode swapping. Among the GFDL models (Fig. 5e), the CMIP6 simulations lie closest to the 1.0 reference line, indicating that their interannual variability (and thus pattern amplitude) is consistent with observations, whereas GFDL-ESM4 has a smaller pattern correlation and larger RMSE than GFDL-CM4. For GISS (Fig. 5f), the CMIP5 models performed best, with the CMIP3 (CMIP6) models underestimating (overestimating) the interannual variability and pattern amplitude. For the NCAR and DOE models (not shown), E3SMv1 and CESM1(CAM5) model are most skillful, with the other CMIP6 models being more skillful than the other CMIP5 or CMIP3 models.

Finally, of the modes analyzed, the NAO is best simulated overall with pattern correlations of ~ 0.95 and RMSE values less than 0.5 hPa (Fig. 5g). Collectively, CMIP6 has smaller skill dispersion than CMIP5 or CMIP3, with models tending to be located closer to the 1.00 reference line. For GFDL, CMIP6 is marginally more skillful than the other MIPs, while CMIP6 GISS E2.1-G overestimates the pattern amplitude compared to CMIP5 and CMIP3 versions of the model. For the NCAR family of models (Fig. 5h), most of the CMIP5 models (especially CESM1(CAM5) and E3SMv1), marginally outperform their CMIP6 and CMIP3 counterparts.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

To summarize: only for the SAM during DJF we do see collective improvement from CMIP3 to CMIP6 in the representation of extra-tropical coupled modes among all models. Otherwise, the inter-model scatter is large, spanning the limited and varied number of ensemble members across the MIP generations. Our conclusion is that for the extra-tropical modes the skill improvement from CMIP3 to CMIP6 is highly mode and seasonally dependent. Taylor Diagrams for additional modes and seasons are available online (See Data Availability).

Sensitivity Experiments: The relevant sensitivity experiment for the extra-tropical atmospheric modes is the CESM2-gamma historical simulation described in Section 2b. Results for this simulation are shown in the panels of Figure 5 that include the family of NCAR models (5a-b,g-h). For the most part, the differences between the CESM and CEMS2-gamma comparisons with reference data are nominal, and could possibly be owing to the limited sample (only one realization of the CESM2-gamma is included). One exception is the SAM during JJA, for which this sensitivity simulation appears to be an outlier in both pattern and amplitude. This aside, however, the performance of the extra-tropical atmospheric modes does not appear to be clearly sensitive to the gamma coefficient in the CLUBB shallow turbulence scheme as applied in CESM2.

c. Tropical Coupled Modes

1) El Niño/Southern Oscillation

Composites of ENSO events are derived from both ERA20C and the CMIP models (Fig. 6) using normalized detrended December Nino3.4 timeseries that are smoothed with a binomial filter and selected for all years when absolute anomalies exceed one standard deviation (El Niño) and all years less than -1 standard deviation (La Niña). Mean model biases (Fig. 6b) indicate a systematic

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

weakness in the ENSO pattern in models as they are negatively correlated with the observed pattern of ENSO (Fig. 6a), which is characterized by negative pressure anomalies in the eastern tropical Pacific Ocean that extend to higher latitudes over the northeastern Pacific, northern Atlantic, and Southern Ocean (Sarachik and Cane 2009) and by positive pressure anomalies over the western tropical and subtropical Pacific Ocean. There are also biases arising from a westward displacement of simulated anomalies (discussed further below), as evident in the negative biases in the western Pacific Ocean and a dipole of biases in the northern Pacific Ocean (b).

Decomposing the models' bias patterns using EOFs reveals that the leading EOF, which explains a substantial amount of inter-model variance (28%), correlates strongly with the composite pattern (Fig. 6c), particularly near the Aleutian Low (Butler et al. 2014). The second EOF, which also explains a significant amount of variance (17%), is characterized by negative values in the western Pacific Ocean and positive values in the Arctic (Fig. 6d). Negative loadings of both EOFs are found in the US CMIP6 models that most closely agree with the observations. In particular, the best US model is characterized by a pattern that strongly resembles the observations in both hemispheres and all ocean basins, although its spatial variance is somewhat stronger (Fig. 6f). In contrast, the model that least agrees with the observations is characterized by weaker than observed teleconnections, both within and outside of the tropics, and exhibits large scale teleconnections that are opposite in sign to those observed in some regions (i.e., the North Atlantic Ocean) (Fig. 6e).

The observed transient evolution of El Niño (Fig. 7), shows a gradual warming of the tropical Pacific Ocean from the dateline to the western coast of the Americas in Year 0, reaching a peak in December near 2K and transitioning on average to cooler than normal conditions of about -0.1K one year later. The mean model bias (Fig. 7b) is characterized by warm anomalies that extend too far westward (identified previously in Fig. 6b) and occur too late in the seasonal cycle, as evidenced by a broad band of positive SST biases in late spring of Year 1. The leading EOF of

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

model bias for the Niño Hovmöller plots (Fig. 7c), which explains a significant fraction of inter-model variance (45%), strongly correlates with the observed composite, indicating a systematic overestimation by many models of the time-longitude structure.

The second bias EOF, which explains 18% of the variance, is characterized by strong warm anomalies in Year 1, suggesting a failure to adequately transition to La Niña conditions over time in some models (Fig. 7d). In the best model (Fig. 7e) and consistent with observations (Fig. 7a), anomalies intensify during Year 0, are located primarily east of the dateline with cool anomalies to their west, and peak in December, after which they transition rapidly to cold anomalies in spring of Year 1. In contrast, in the poorer performing model (Fig. 7f), positive anomalies also grow gradually through Year 0 but extend well into Year 1. A large-scale transition to cool Pacific SSTs in Year 2 is simulated in the poorer scoring model, but to an extent that is weaker than observed. That said, even the poorer scoring CMIP6 US model is considerably more skillful than many other models included in the CMIP archives (not shown).

2) Pacific Decadal Oscillation

The Taylor Diagram of the PDO, derived from the leading EOF of North Pacific (20°N–70°N) SST anomalies, suggests that the RMSE has been reduced and pattern correlation increased from CMIP3 to CMIP5 to CMIP6 (Fig. 8). Among CMIP5 and CMIP6 versions of both the GFDL and GISS models, the representation of the PDO has improved and among the NCAR models, CMIP6 CESM2 has the largest pattern correlation and smallest RMSE values compared to either CMIP5 or CMIP3 versions. The overall tendency, therefore, is for CMIP6 models to have larger pattern correlations and smaller RMSE than their corresponding CMIP3 and CMIP5 versions, a result that also holds using the Common Basis Function approach (not shown).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Regressions of the principle component timeseries of the PDO against SSTs can be used to quantify the global teleconnection patterns associated with the PDO (Phillips et al. 2014), which consists of a zonal dipole of anomalies in the North Pacific centered near 40°N that resembles the structure of El Niño (Fig. 9a). As for El Niño, the mean model bias (not shown) negatively correlates with the observed pattern, which indicates an overall weakness of the PDO in the models. In addition, the leading EOF pattern of the inter-model PDO bias (Fig. 9c), which is associated with connections between the Northern and Tropical Pacific Ocean, also positively correlates with the mean bias across models suggesting that, while on average the simulated PDO connections with the Tropics are systematically underestimated, there is also considerable variation across models.

This last point is especially evident when comparing the PC weightings corresponding to the leading EOFs for the CMIP5 and CMIP6 versions of US models (Fig. 9b), where the origin may be interpreted as the CMIP mean model bias, and observations are shown in red. In particular, significant improvement from CMIP5 to CMIP6 model versions is reflected by the relative proximity of PC weightings for CMIP6 (closed circles) versus CMIP5 model versions (open circles) to the observed range. Most improved is the GISS model (green), which had amongst the most positive leading EOF1 PCs in CMIP5, yet in CMIP6 is among the closest to the observed range. Note that the reduction of bias in EOF 1 in the US climate models and inconsistent changes in EOF 2 mirror the more general changes from CMIP5 to CMIP6 seen in other climate models (Fasullo et al. 2020).

To illustrate this improvement directly, the simulated PDO regression patterns are shown for the CMIP5 (Fig. 9e) and CMIP6 (Fig. 9f) versions of the GISS model. In CMIP5, significant teleconnections were largely confined to the North Pacific Ocean, with minimal structure in the other ocean basins and in stark contrast to the observed pattern. In CMIP6, teleconnections to remote regions have expanded considerably, particularly in the Tropics, with a strong spatial correlation

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

with those observed, though biases in the tropical meridional structure remain. Correctly resolving these teleconnections has broad relevance to associated attribution and prediction applications and represents a major step forward, despite some persistence of biases. While the processes that contribute to the improvement remains to be understood, it is noteworthy that patterns in the North Pacific have not improved dramatically across model versions (Fig. 8), suggesting other origins for the teleconnection improvement.

3) ENSO and PDO Spectra

Another key property of both ENSO and the PDO is the spectra of their indices, which are associated with considerable socio-economic implications related to the frequency, intensity, and persistence of drought, floods, and other impacts (Dilley and Heyman 1995). In general, the power of simulated ENSO variability is too strong in models except at high frequencies ($<$ 2.5 years) (Fig. 10a), where many models underestimate the severity of El Niño-La Niña transitions (also shown in Fig. 7b). By comparison, between 2.5 and 6 years, all US climate models produce on average more power than is observed. In the 6–10 year band, the average observed power is reduced from the 2.5–6 year band but remains large, with the E3SM range on par with the observed estimates. For periods greater than 10 years, the observed power is, again, less than that for 2.5 to 6 year periods and the agreement between the models and observations is closer, with the exception of CESM2(CAM6) and CESM2(WACCM6). A systematic increase in power from CMIP5 (coincident black lines) to CMIP6 model versions is apparent in all CMIP6 US models except CESM, where the Version 1 and Version 2 ranges overlap. The general increase in ENSO power across US CMIP6 model versions for periods greater than 2.5 yrs relative to their CMIP5 counterparts is consistent with the broader increase in power from CMIP5 to CMIP6 models overall (Fasullo et al. 2020).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Model-observation agreement is generally closer for the PDO spectra (Fig. 10b), compared to that for ENSO. Models systematically underestimate the observed estimates in the less than 2.5 year band, where there is little power; by comparison, in the 2.5–6 year band the models exhibits generally good agreement with the observations, albeit with large internal variability. In the 6 to 10 year band, observational estimates again increase and fall generally within the model ranges, with WACCM6 perhaps being biased high, although more ensemble members are needed to more accurately represent the ensemble spread in that model. At low frequencies ($>$ 10 years), power in the PDO is still larger and good general agreement exists between the observed and simulated estimates.

Sensitivity Experiments: As seen in Figure 11, the gamma parameter in the CESM2 sensitivity experiments exerts an important influence on ENSO teleconnections. The spatial character of ENSO-SLP teleconnections (similar to Fig. 6, although here estimated using Nino3.4 regression) is shown for ERA-20C (Fig. 11a) and for the CESM2-gamma sensitivity experiments (Fig. 11b), along with the raw regressions for each model. In CESM2-gamma, many of the canonical biases are shown to worsen relative to CESM2 and include a weakening of the overall pattern in most locations, manifest as negative anomalies in the northeast Pacific Ocean and North Atlantic Ocean. They also include a westward shift of ENSO variance, as evidenced by negative differences in the central Pacific Ocean (Fig. 11b). The pattern correlations versus observations also decrease for CESM2-gamma (0.88) from those for CESM2 (0.93). Together the biases are analogous to PC1 and PC2 in our multi-model analysis (Fig. 6c,d) and demonstrate a basic sensitivity of ENSO teleconnections to unobserved cloud parameters that are typically tuned.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

d. Quasi-Biennial Oscillation

We evaluate the QBO only among the current (CMIP6) generation of models (Table 1). This is because, unlike for the other modes, the QBO was not consistently represented in the majority of models participating in previous CMIP intercomparisons (i.e. only 5 of 47 CMIP5 models had anything resembling a QBO (Butchart et al. 2018)). We also include in our analysis not only historical coupled runs but historical AMIP runs that were not included in the previous discussions which, by comparison, focused on modes requiring atmosphere-ocean coupling. Specifically, these include the results from the GEOS M2AMIP ensemble as well as the GISS E2.2-G AMIP historical runs (Table 3).

1) QBO PERIOD

The QBO is first depicted in terms of the evolution of the equatorial winds, averaged over 5°S to 5°N, over the course of the observational period 1980–2015 (or up to years for which model output was available, depending on the simulation) (Figure 12). MERRA-2 exhibits the characteristic oscillating propagation downward of zonal wind anomalies, also featured clearly in all of the other models, with the exception of CESM2(CAM6) and GISS E2.1. This is not surprising given that the latter models have relatively low vertical resolutions (Table 2). Hereafter, therefore, our focus will be on further quantifying various aspects of the QBO in all models exempting CESM2(CAM6) and GISS E2.1. We also exclude from our analysis the results from CESM1(WACCM5) since the QBO was imposed in that model.

As in Schenzinger et al. (2017) (hereafter SC17) we begin by calculating the Fourier transform of the equatorial zonal mean zonal wind; h_{max} is then defined as the height at which the sum of the squares of the Fourier amplitudes between 26 and 30 months maximizes. For MERRA-2 this occurs at 20 hPa which is consistent with the values of h_{max} quoted in Coy et al. (2016), as well

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

as SC17, albeit for the case of MERRA in the latter. Comparison of h_{max} among the simulations shows good consistency with MERRA-2 to within ± 15 hPa, with h_{max} occurring above 20 hPa for some models (i.e. M2AMIP, some members of GISS Model E2.2) and below 20 hPa for others (i.e. E3SMv1 and E3SMv1-MOGWD, CESM1-WACCM) (Fig. 13a). The one exception is E3SMv1, for which h_{max} spans 30–50 hPa. While this is somewhat at odds with Bushell et al. (2020), who showed that the QBO in the QBOi models is generally shifted upward compared to reanalyses, it important to note that we are considering a much smaller ensemble of models compared to the 13 models considered in that study. Furthermore, offline comparisons of h_{max} in MERRA-2 calculated used the native vertical grid of MERRA-2 (20 hPa) versus the coarser grid at which the MoV model output was available (30 hPa) demonstrates that h_{max} does exhibit some sensitivity to vertical resolution. While this implies that some caution needs to be taken when interpreting the sense of the MoV models' bias in h_{max}, as we discuss below, other measures of the QBO (e.g. amplitude, period) are less sensitive.

Time series of the equatorial zonal winds at h_{max} are used to identify the time between every other phase peak. For MERRA-2 the average period over all cycles is 28.2 months, with a minimum (maximum) period of 22 (36) months (Fig. 13b). This is in excellent agreement with equatorial radiosonde-based estimates, which are also slightly above 28 months (Baldwin et al. 2001). The mean periods in the examined models generally all agree very well with MERRA-2, particularly for the M2AMIP ensemble in which the QBO period values for all ten members range between 27 months and 29.1 months. By comparison, the QBO period is not as well captured in E3SMv1, which features a period that is almost twice as fast as in MERRA-2 for some members.

We explore this last point further by contrasting the results from E3SMv1 with those from E3SMv1-MODGWD. In response to two separate changes to the convective GWD – both with respect to the amplitude of the momentum flux phase speed spectra (which in that model is pro-

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

portional to a tunable convective fraction per GCM grid cell) and the efficiency with which convection generates gravity waves – there is a significant improvement in the QBO period exhibited in E3SMv1-MODGWD (Fig. 13b). Given that R19 tuned that simulation to obtain a credible QBO period this result is not surprising. Nonetheless, it is consistent in spirit with the development decisions that were made in tuning similar aspects of the convective component of the non-orographic gravity wave drag scheme in GISS E2.2 (Rind et al. 2020) and in previous versions of that model (Rind et al. 2014) (hereafter RJ14).

2) QBO Amplitude

Compared to the QBO period in MERRA-2 the models considered here exhibit more disagreement in terms of the amplitude of the QBO (Fig. 13c), which likely reflects the priorities governing how GWD schemes are tuned in models (i.e. to first produce a credible period and, thereafter, other aspects of the QBO). The best agreement with MERRA-2, in which the amplitude is $\sim$45 m s^{-1}, is exhibited by GISS E2.2 and M2AMIP; by comparison, in nearly all the other models the amplitude of the QBO is underestimated, consistent with the QBOi models (Bushell et al. 2020). Further decomposition of the QBO amplitude into easterly and westerly components (Figure 13, d-g) shows that this low amplitude bias among the models is most often associated with an underestimate of the easterly component of the QBO (Fig. 13, d-e). By comparison the amplitude of the westerly phase of the QBO is better represented in the models (Fig. 13, f-g).

It is interesting to ask if the differences in QBO amplitude exhibited among the models in the historical runs compare in magnitude to the differences in forecast skill among the two sub-seasonal forecast ensembles considered in this study. Figure 14 compares RMSE values in the equatorial zonal winds between the GEOS-S2S and NOAA GEFS ensemble forecasts, relative to MERRA-2. With the exception of 100 hPa, where GEFS performs slightly better, the S2S forecast

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

errors are smaller throughout the stratosphere, especially above 30 hPa. Comparisons of the vertical resolution and model top of GEFS with the underlying GEOS GCM used to produce the S2S forecasts indicates that the GEFS top is lower (0.2 hPa vs. 0.01 hPa) and has fewer vertical levels, with less distribution in the lower stratosphere/upper troposphere (Table 2). At the same time, other factors may also contribute to the differences, including the fact that the stochastic perturbation that is applied to each GEFS forecast member varies with height. In particular, in the stratosphere the weight of the stochastic perturbation decreases from 1 at 100 hPa to 0 at pressures at and above 25 hPa, which could contribute to the variable performance of the GEFS ensemble mean forecast at different stratospheric levels. This, in addition to the use of MERRA-2 as our reference against which all RMSE values have been calculated, may provide additional reasons for the relatively weaker QBO skill in the GEFS forecasts. Therefore, while the vertical resolution and model top differences between the models are consistent with the sources driving differences in model performance among the historical runs, more investigation is needed to understand how (if) skill on the climatic timescales relevant to CMIP translates to sub-seasonal timescales in any meaningful way. A more rigorous and in-depth presentation of the state of the QBO in the GEOS-S2S forecasts is currently in preparation and soon to be submitted for publication (personal communication with Dr. Lawrence Coy (NASA, Global Modeling Assimilation Office)).

Intermediary Model Version Experiments: Comparisons between pairs of models show that the QBO period depends sensitively on which aspects of the non-orographic gravity wave drag are altered. In particular, comparisons of GISS E2.1 versus E2.2 and E3SMv1 versus E3SMv1-MODGWD, in which changes in the efficiency with which convection generates gravity waves were made in both cases, resulted in significant improvements in the QBO period, relative to MERRA-2.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

By comparison, our analysis does not reveal any systematic changes that clearly improve the representation of the amplitude of the QBO. In particular, the QBO amplitude in E3SMv1-MODGWD is approximately the same as in E3SMv1 (Fig. 13c), albeit with some differences, depending on QBO phase (Fig. 13d). This can also be seen in the perhaps surprising result that, despite their good representation of the overall mean QBO amplitude, all M2AMIP members consistently underestimate (overestimate) the easterly (westerly) QBO amplitude (Fig. 13d,e(f,g)). This is interesting because the M2AMIP ensemble was generated using the exact same version of GEOS that was used to produce MERRA-2. This demonstrates that, while changes in the non-orographic GWD parameterizations may suffice in terms of improving aspects like QBO period, they may not be sufficient for constraining other aspects like QBO amplitude. For that, assimilation of observed fields (as in MERRA-2) can counteract underlying free-running biases in the models (Geller et al. 2016).

4. Discussion

We have presented a comprehensive assessment of the performance of US climate models with respect to multiple modes of variability. Overall, we show that for many modes (though not all), improvements in model skill over time are impressive and a testament to the improvements in the representation of key processes. In addition to improved representations of the MJO and QBO (which have been reported in previous studies (Kim et al. 2013; Rind et al. 2014; Danabasoglu et al. 2020)), the overall improvement in ENSO and the PDO in recent CMIP6 models is remarkable (Figs. 15a,b). At the same time, however, there is no clear improvement in the representation of the NAM and possible degradation of skill in the SAM (Figs. 15c,d), although the correlations were very high already.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

We can distinguish between two kinds of improvement exhibited among most of the modes considered in this study: those that rely on a threshold of model representation that is crossed at a distinct moment in model development, and improvements that rely on more gradual, collective improvements in processes. As an example of the first, the ability of GISS E2.2, CESM2(WACCM6) and E3SMv1-MODGWD to produce a realistic QBO signal is predominantly a function of increased vertical resolution in the lower stratosphere and sufficiently complex spectra of parameterized gravity waves that are tied to the underlying physics in the models (e.g. convection, shear). Models without either do not have a QBO worth discussing, while those with have at least the possibility of being able to tune for a realistic amplitude and period. In this latter group, the QBO period is much easier to tune than the amplitude (Geller et al. 2016), which is consistently underestimated. While data assimilation can remedy this bias (evident in comparing M2AMIP with MERRA-2), it remains a challenge for future development.

Improvement in the simulation of coupled and extra-tropical modes falls into the second category of model improvement, likely being attributed to gradual improvement of the base climate and a range of relevant processes. Evidence of improved fidelity across generations is apparent in some cases (e.g., the amplitude of the SAM in DJF), but less clear in others (NAM, NAO, PNA, the SAM during JJA). While the limited and varied number of samples hamper definitive statements that can be generalized across models, our analysis nevertheless suggests that progress has been made in some areas, most notably for ENSO and the PDO. Improvements seen in the MJO are also an example of this latter approach, although the improvements are much clearer, compared to the extra-tropical modes. The drivers of these improvements also appear to be much better understood and related to consistent approaches to treating rain re-evaporation within convective parameterizations.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Finally, while the results from our analysis suggests a clear progression in model fidelity in a climate context, it is not clear how (if) this improved performance translates to skill in subseasonal forecasting. Our limited analysis comparing two subseasonal forecast groups suggests that the factors contributing to improved QBO performance in the climate context may also improve skill on subseasonal timescales. Owing, however, to the limited number of subseasonal models considered in this study, our analysis is not conclusive. As more forecast systems become available in parallel with new CMIP6 models, however, it will become easier to address this question.

5. Data Availability

All CMIP simulations are available through the Earth System Grid Federation (ESGF). In addition, all intermediary and sensitivity experiments as well as supplementary figures are publicly available. Specifically, the MJO and QBO data can be found at `https://portal.nccs.nasa.gov/datashare/GISS_MOV/`. The summary data for the CESM simulations is available at `ftp://ftp.cgd.ucar.edu/archive/andrew/cesm2ecs` while raw model output can be found at: `https://doi.org/10.26024/zrad-5z41`. CESM1 and CESM2 code bases are available via links from `http://www.cesm.ucar.edu/models/`. The analysis codes used for the extra-tropical modes are available via PMPv1.2 `https://github.com/PCMDI/pcmdi_metrics`. Taylor Diagrams for additional modes and seasons are available online at `https://pcmdi.llnl.gov/pmp-preliminary-results/variability_modes/US_models_taylor_diagrams`. The M2AMIP GMAO ensembles can be accessed at `https://portal.nccs.nasa.gov/datashare/gmao_m2amip/`, while a subset of the GEOS S2S data analyzed in this manuscript is available at `https://gmao.gsfc.nasa.gov/gmaoftp/gmaofcst/subx/GEOS_S2S_V2.1/`. The GEFS SubX output is available via the IRI Data Library `http:`

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

`//iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/` (doi: `https://dx.doi.org/10.7916/D8PG249H`).

Acknowledgments. This work was funded from NASA MAP, DOE and NOAA, and arises from the 2019 US Climate Modeling Summit in Washington D.C. co-chaired by Steven Pawson and Gavin Schmidt. Climate modeling at GISS and the GMAO is supported by the NASA Modeling, Analysis and Prediction program, and resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center. This research was also supported as part of the Energy Exascale Earth System Model (E3SM) project and Regional and Global Climate Modeling Program, both funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. Work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. All presented material relevant to the CESM project, which is supported primarily by the National Science Foundation (NSF), is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement 1852977. Portions of this study were 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 and Environmental Research (BER) Cooperative Agreement DE-FC02-97ER62402. Computing and data storage resources, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. The GEFS SubX forecasts were partially supported through NWS OSTI and NOAA's Climate Program Office (CPO) Modeling, Analysis, Predictions, and Projections (MAPP) program. We also thank the GFDL model development team, the leadership of

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

NOAA/GFDL for their efforts and support in developing CM4/ESM4 as well as the many GFDL scientists and engineers who conducted the CM4/ESM4 runs and made the data available at ESGF. We acknowledge the modeling groups, PCMDI and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the CMIP multi-model datasets. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. We would like to thank the three reviewers whose comments helped greatly increase the clarity of the manuscript.

References

Alexander, M. J., and T. J. Dunkerton, 1999: A spectral parameterization of mean-flow forcing due to breaking gravity waves. *J. Atmos. Sci.*, **56**, 4167 4182.

Ambaum, M. H., B. J. Hoskins, and D. B. Stephenson, 2001: Arctic oscillation or North Atlantic oscillation? *Journal of Climate*, **14 (16)**, 3495–3507.

Baldwin, M. P., and Coauthors, 2001: The quasi-biennial oscillation. *Reviews of Geophysics*, **39 (2)**, 179–229.

Borovikov, A., R. Cullather, R. Kovach, J. Marshak, G. Vernieres, Y. Vikhliaev, B. Zhao, and Z. Li, 2017: GEOS-5 seasonal forecast system. *Climate Dynamics*, 1–27.

Bretherton, C. S., J. R. McCaa, and H. Grenier, 2004: A new parameterization for shallow cumulus convection and its application to marine subtropical cloud-topped boundary layers. Part I: Description and 1-D results. *Mon. Wea. Rev.*, **132**, 864–882.

Bushell, A. C., and Coauthors, 2020: Evaluation of the Quasi-Biennial Oscillation in global climate models for the SPARC QBO-initiative. *Quarterly Journal of the Royal Meteorological Society*, 1–31, doi:10.1002/qj.3765.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Butchart, N., and Coauthors, 2018: Overview of experiment design and comparison of models participating in phase 1 of the SPARC Quasi-Biennial Oscillation initiative (QBOi). *Geoscientific Model Development*, **11 (3)**, 1009–1032.

Butler, A. H., L. M. Polvani, and C. Deser, 2014: Separating the stratospheric and tropospheric pathways of El Niño – Southern Oscillation teleconnections. *Env. Res. Lett.*, **9**, 024 014, doi: 10.1088/1748-9326/9/2/024014.

Chun, H.-Y., and J.-J. Baik, 1998: Momentum flux by thermally induced internal gravity waves and its approximation for large-scale models. *Journal of the atmospheric sciences*, **55 (21)**, 3299–3310.

Clark, S. K., Y. Ming, and A. F. Adames, 2020: Monsoon low pressure system-like variability in an idealized moist model. *Journal of Climate*, **33 (6)**, 2051–2074, doi:10.1175/JCLI-D-19-0289.1.

Clement, A., P. DiNezio, and C. Deser, 2011: Rethinking the ocean's role in the Southern Oscillation. *Journal of Climate*, **24 (15)**, 4056–4072, doi:10.1175/2011jcli3973.1.

Cohen, J., and M. Barlow, 2005: The NAO, the AO, and global warming: How closely related? *Journal of Climate*, **18 (21)**, 4498–4513.

Cohen, J., A. Frei, and R. D. Rosen, 2005: The role of boundary conditions in AMIP-2 simulations of the NAO. *Journal of climate*, **18 (7)**, 973–981.

Compo, G., J. Whitaker, and P. Sardeshmukh, 2008: The 20th Century Reanalysis. *Proceedings of the Third WCRP International Conference on Reanalysis, The University of Tokyo, Japan. 28 Jan?1 Feb 2008.*, URL http://wcrp.ipsl.jussieu.fr/Workshops/Reanalysis2008/Documents/V5?511_ea.pdf.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Coy, L., K. Wargan, A. M. Molod, W. R. McCarty, and S. Pawson, 2016: Structure and dynamics of the quasi-biennial oscillation in MERRA-2. *Journal of Climate*, **29 (14)**, 5339–5354.

Dai, P., and B. Tan, 2017: The nature of the Arctic Oscillation and diversity of the extreme surface weather anomalies it generates. *Journal of Climate*, **30 (14)**, 5563–5584.

Danabasoglu, G., J. Lamarque, J. Bacmeister, D. Bailey, A. DuVivier, J. Edwards, and Coauthors, 2019: The Community Earth System Model version 2 (CESM2). *J. Advances in Modeling Earth Systems*, Under Review.

Danabasoglu, G., and Coauthors, 2020: The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, (submitted).

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597.

Del Genio, A. D., Y. Chen, D. Kim, and M.-S. Yao, 2012: The MJO Transition from Shallow to Deep Convection in CloudSat/CALIPSO Data and GISS GCM Simulations. *Journal of Climate*, **25 (11)**, 3755–3770, doi:10.1175/JCLI-D-11-00384.1, URL https://doi.org/10.1175/JCLI-D-11-00384.1.

Del Genio, A. D., J. Wu, A. B. Wolf, Y. Chen, M.-S. Yao, and D. Kim, 2015: Constraints on cumulus parameterization from simulations of observed MJO events. *Journal of Climate*, **28 (16)**, 6419–6442, doi:10.1175/jcli-d-14-00832.1.

Del Genio, A. D., M.-S. Yao, and J. Jonas, 2007: Will moist convection be stronger in a warmer climate? *Geophysical Research Letters*, **34 (16)**.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Deser, C., 2000: On the teleconnectivity of the “Arctic Oscillation”. *Geophysical Research Letters*, **27 (6)**, 779–782.

Dilley, M., and B. N. Heyman, 1995: ENSO and disaster: Droughts, floods and El Niño/Southern Oscillation warm events. *Disasters*, **19 (3)**, 181–193, doi:10.1111/j.1467-7717.1995.tb00338.x.

Donner, L., C. Seman, R. Hemler, and S.-M. Fan, 2001: A cumulus parameterization including mass fluxes, convective vertical velocities, and mesoscale effects. *J. Climate*, **14**, 3444–3463.

Donner, L. J., B. L. Wyman, R. Hemler, L. W. Horowitz, Y. Ming, M. Zhao, and Coauthors, 2011: The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *J. Climate*, **24**, 3484–3519.

Dunne, J. P., L. W. Horowitz, A. J. Adcroft, P. Ginoux, I. M. Held, and Coauthors, 2020: The GFDL Earth System Model version 4.1 (GFDL-ESM4.1): Model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, in review.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, **9 (5)**, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Fasullo, J. T., A. S. Phillips, and C. Deser, 2020: Evaluation of leading modes of climate variability in the CMIP archives. *Journal of Climate*, accepted.

Feldstein, S. B., and C. Franzke, 2006: Are the North Atlantic Oscillation and the northern annular mode distinguishable? *Journal of the Atmospheric Sciences*, **63 (11)**, 2915–2930.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Garcia, R. R., and B. A. Boville, 1994: "Downward control" of the mean meridional circulation and temperature distribution of the polar winter stratosphere. *Journal of the atmospheric sciences*, **51 (15)**, 2238–2245.

Garner, S. T., 2005: A topographic drag closure built on an analytical base flux. *J. Atmos. Sci.*, **62**, 2302–2315.

Gates, W. L., and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP1). *Bull. Am. Meteor. Soc.*, **80**, 29–55.

Geller, M. A., and Coauthors, 2016: Modeling the QBO - Improvements resulting from higher-model vertical resolution. *Journal of Advances in Modeling Earth Systems*, **8 (3)**, 1092–1105.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *Journal of Climate*, **24 (19)**, 4973–4991, doi:10.1175/2011jcli4083.1.

Gettelman, A., M. Mills, D. Kinnison, R. Garcia, A. Smith, D. Marsh, and Coauthors, 2019: The Whole Atmosphere Community Climate Model version 6 (WACCM6). *J. Geophysical Research Atmospheres*, Under Review.

Gleckler, P., C. Doutriaux, P. J. Durack, K. E. Taylor, Y. Zhang, D. N. Williams, E. Mason, and J. Servonnat, 2016: A more powerful reality test for climate models. *Eos*, **97**, doi:10.1029/2016EO051663.

Golaz, J.-C., V. E. Larson, and W. R. Cotton, 2002: A pdf-based model for boundary layer clouds. Part I: Method and model description. *Journal of the Atmospheric Sciences*, **59 (24)**, 3540–3551.

Golaz, J.-C., and Coauthors, 2019: The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, **11 (7)**, 2089–2129.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Gong, G., D. Entekhabi, and J. Cohen, 2002: A large-ensemble model study of the wintertime AO–NAO and the role of interannual snow perturbations. *Journal of Climate*, **15 (23)**, 3488–3499, doi:10.1175/1520-0442(2002)015⟨3488:ALEMSO⟩2.0.CO;2.

Gottschalck, J., and Coauthors, 2010: A framework for assessing operational madden?julian oscillation forecasts. *Bulletin of the American Meteorological Society*, **91 (9)**, 1247–1258, doi:10.1175/2010BAMS2816.1.

Griffies, S. M., and Coauthors, 2011: GFDL's CM3 coupled climate model: Characteristics of the ocean and sea ice simulations. *J. Climate*, **24**, 3520–3544, doi:10.1175/2011JCLI3964.1.

Hansen, J. E., G. L. Russell, D. Rind, P. Stone, A. Lacis, R. Ruedy, and L. Travis, 1983: Efficient three-dimensional models for climatic studies. *Mon. Wea. Rev.*, **111**, 609–662, doi:10.1175/1520-0493(1983)111⟨0609:ETDGMF⟩2.0.CO;2.

Held, I. M., and Coauthors, 2019: Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, doi:10.1029/2019MS001829.

Hersbach, H., and D. Dee, 2016: ERA-5 reanalysis is in production. *ECMWF newsletter*, **147**, 7.

Hurrell, J., M. Holland, P. Gent, S. Ghan, , J. Kay, P. Kushner, and Coauthors, 2013a: The Community Earth System Model. *Bull. Amer. Met. Soc.*, doi:10.1175/BAMS-D-12-00121.1.

Hurrell, J. W., and Coauthors, 2013b: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94 (9)**, 1339–1360, doi:10.1175/bams-d-12-00121.1.

Iguchi, T., T. Kozu, R. Meneghini, J. Awaka, and K. Okamoto, 2000: Rain-profiling algorithm for the TRMM precipitation radar. *Journal of Applied Meteorology*, **39 (12)**, 2038–2052.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Inness, P. M., J. M. Slingo, E. Guilyardi, and J. Cole, 2003: Simulation of the Madden–Julian Oscillation in a coupled general circulation model. Part II: The role of the basic state. *Journal of Climate*, **16 (3)**, 365–382, doi:10.1175/1520-0442(2003)016⟨0365:sotmjo⟩2.0.co;2.

Jiang, X., and Coauthors, 2015: Vertical structure and physical processes of the Madden-Julian Oscillation: Exploring key model physics in climate simulations. *Journal of Geophysical Research: Atmospheres*, **120 (10)**, 4718–4748, doi:10.1002/2014JD022375.

Kang, I.-S., and S.-I. An, 1998: Kelvin and Rossby wave contributions to the SST oscillation of ENSO. *Journal of Climate*, **11 (9)**, 2461–2469.

Kelley, M., and Coauthors, 2020: GISS-E2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems (JAMES)*.

Kim, D., A. D. Del Genio, and M.-S. Yao, 2013: Moist convection scheme in Model E2. *arXiv preprint arXiv:1312.7496*.

Kim, D., A. H. Sobel, A. D. Del Genio, Y. Chen, S. J. Camargo, M.-S. Yao, M. Kelley, and L. Nazarenko, 2012: The tropical subseasonal variability simulated in the NASA GISS general circulation model. *Journal of Climate*, **25 (13)**, 4641–4659, doi:10.1175/JCLI-D-11-00447.1.

Kim, H., M. A. Janiga, and K. Pegion, 2019: Mjo propagation processes and mean biases in the subx and s2s reforecasts. *Journal of Geophysical Research: Atmospheres*, **124 (16)**, 9314–9331, doi:10.1029/2019JD031139.

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 Dynamics*, **52 (7-8)**, 4057–4089.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Manabe, S., and K. Bryan, 1969: Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci.*, **26 (4)**, 786–789, doi:10.1175/1520-0469(1969)026⟨0786:ccwaco⟩2.0.co;2.

Masunaga, H., 2007: Seasonality and regionality of the Madden-Julian oscillation, Kelvin wave and equatorial Rossby wave. *J. Atmos. Sci.*, **64**, 4400–4416, doi:10.1175/2007JAS2179.1.

McFarlane, N., 1987: The effect of orographically excited gravity wave drag on the general circulation of the lower stratosphere and troposphere. *Journal of the atmospheric sciences*, **44 (14)**, 1775–1800.

Meehl, G. A., 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. *Bulletin of the American Meteorological Society*, **88 (9)**, 1383–1394, doi:10.1175/bams-88-9-1383, URL https://doi.org/10.1175/bams-88-9-1383.

Mills, M., J. Richter, S. Tilmes, B. Kravitz, D. MacMartin, A. Glanville, and Coauthors, 2017: Radiative and chemical response to interactive stratospheric sulfate aerosols in fully coupled CESM1(WACCM). *J. Geophysical Research Atmospheres*, doi:10.1002/2017JD027006.

Molod, A., L. Takacs, M. Suarez, and J. Bacmeister, 2015: Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2. *Geoscientific Model Development*, **8 (5)**, 1339–1356, doi:10.5194/gmd-8-1339-2015.

Molod, A., and Coauthors, 2020: GEOS-S2S Version 2: The GMAO high resolution coupled model and assimilation system for seasonal prediction. *Submitted to Journal of Geophysical Research*.

Moorthi, S., and M. J. Suarez, 1992: Relaxed Arakawa-Schubert. A parameterization of moist convection for general circulation models. *Monthly Weather Review*, **120 (6)**, 978–1002.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

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

Park, S., and C. S. Bretherton, 2009: The University of Washington shallow convection and moist turbulence schemes and their impact on climate simulations with the Community Atmosphere Model. *Journal of Climate*, **22 (12)**, 3449–3469.

Pegion, K., and Coauthors, 2019: The subseasonal experiment (subx): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society*, **100 (10)**, 2043–2060, doi:10.1175/BAMS-D-18-0270.1.

Phillips, A. S., C. Deser, and J. T. Fasullo, 2014: Evaluating modes of variability in climate models. *Eos, Transactions American Geophysical Union*, **95**, 453–455, doi:10.1002/2014EO490002.

Phillips, N. A., 1956: The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, **82 (352)**, 123–164, doi:10.1002/qj.49708235202.

Poli, P., and Coauthors, 2016: ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, **29 (11)**, 4083–4097.

Rasch, P. J., and Coauthors, 2019: An overview of the atmospheric component of the Energy Exascale Earth System Model. *Journal of Advances in Modeling Earth Systems*.

Richter, J. H., C.-C. Chen, Q. Tang, S. Xie, and P. J. Rasch, 2019: Improved simulation of the QBO in E3SMv1. *Journal of Advances in Modeling Earth Systems*, **11 (11)**, 3403–3418, doi:10.1029/2019ms001763.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Richter, J. H., F. Sassi, and R. R. Garcia, 2010: Toward a physically based gravity wave source parameterization in a general circulation model. *Journal of the Atmospheric Sciences*, **67 (1)**, 136–156.

Rind, D., J. Jonas, N. K. Balachandran, G. A. Schmidt, and J. Lean, 2014: The QBO in two GISS global climate models: 1. Generation of the QBO. *Journal of Geophysical Research: Atmospheres*, **119 (14)**, 8798–8824.

Rind, D., and Coauthors, 2020: GISS Model E2.2: A climate model optimized for the middle atmosphere. *Journal of Geophysical Research*, submitted.

Rivière, G., and M. Drouard, 2015: Dynamics of the northern annular mode at weekly time scales. *Journal of the Atmospheric Sciences*, **72 (12)**, 4569–4590.

Rogers, J., and M. McHugh, 2002: On the separability of the North Atlantic oscillation and Arctic oscillation. *Climate Dynamics*, **19 (7)**, 599–608.

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

Russell, G. L., J. R. Miller, and D. Rind, 1995: A coupled atmosphere-ocean model for transient climate change. *Atmosphere-Ocean*, **33 (4)**, 683–730.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *Journal of Climate*, **27 (6)**, 2185–2208, doi:10.1175/jcli-d-12-00823.1.

Sarachik, E. S., and M. A. Cane, 2009: *The El Niño–Southern Oscillation Phenomenon*. Cambridge University Press, doi:10.1017/cbo9780511817496.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Schenzinger, V., S. Osprey, L. Gray, and N. Butchart, 2017: Defining metrics of the Quasi-Biennial Oscillation in global climate models. *Geoscientific Model Development*, **10 (6)**.

Schmidt, G. A., and Coauthors, 2014: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *J. Adv. Model. Earth Syst.*, **6**, 141–184, doi:10.1002/2013MS000265.

Scinocca, J., and N. McFarlane, 2000: The parametrization of drag induced by stratified flow over anisotropic orography. *Quarterly Journal of the Royal Meteorological Society*, **126 (568)**, 2353–2393.

Song, J., 2019: The long-and short-lived North Atlantic Oscillation events in a simplified atmospheric model. *Journal of the Atmospheric Sciences*, **76 (9)**, 2673–2700.

Sperber, K. R., and D. Kim, 2012: Simplified metrics for the identification of the Madden-Julian Oscillation in models. *Atmospheric Science Letters*, **13 (3)**, 187–193, doi:10.1002/asl.378.

Stephenson, D., V. Pavan, M. Collins, M. Junge, R. Quadrelli, and Coauthors, 2006: North Atlantic Oscillation response to transient greenhouse gas forcing and the impact on european winter climate: a CMIP2 multi-model assessment. *Climate Dynamics*, **27 (4)**, 401–420.

Stern, W. F., and R. T. Pierrehumbert, 1988: The impact of an orographic gravity wave drag parameterization on extended range predictions with a GCM. *Eighth Conf. on Numerical Weather Prediction, Baltimore, MD, Amer. Meteor. Soc.*, 745–750.

Stoner, A. M. K., K. Hayhoe, and D. J. Wuebbles, 2009: Assessing general circulation model simulations of atmospheric teleconnection patterns. *Journal of Climate*, **22 (16)**, 4348–4372.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Thompson, D. W., S. Lee, and M. P. Baldwin, 2003: Atmospheric processes governing the Northern Hemisphere annular mode/North Atlantic oscillation. *Geophysical Monograph-American Geophysical Union*, **134**, 81–112.

Thompson, D. W., and J. M. Wallace, 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical research letters*, **25 (9)**, 1297–1300.

Vitart, F., 2017: Madden?julian oscillation prediction and teleconnections in the s2s database. *Quarterly Journal of the Royal Meteorological Society*, **143 (706)**, 2210–2220, doi:10.1002/qj.3079.

Waliser, D., and Coauthors, 2009: MJO Simulation Diagnostics. *J. Climate*, **22**, 3006–3030, doi:10.1175/2008JCLI2731.1.

Wallace, J. M., 2000: North Atlantic oscillation/annular mode: two paradigms – one phenomenon. *Quarterly Journal of the Royal Meteorological Society*, **126 (564)**, 791–805.

Wheeler, M., and G. N. Kiladis, 1999: Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber-frequency domain. *J. Atmos. Sci.*, **56**, 374–399, doi:10.1175/1520-0469(1999)056⟨0374:CCEWAO⟩2.0.CO;2.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate mjo index: Development of an index for monitoring and prediction. *Monthly Weather Review*, **132 (8)**, 1917–1932, doi:10.1175/1520-0493(2004)132⟨1917:AARMMI⟩2.0.CO;2.

Xie, S., and Coauthors, 2018: Understanding cloud and convective characteristics in version 1 of the E3SM atmosphere model. *Journal of Advances in Modeling Earth Systems*, **10 (10)**, 2618–2644.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Zhang, G. J., and N. A. McFarlane, 1995: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model. *Atmosphere-ocean*, **33 (3)**, 407–446.

Zhao, M., J.-C. Golaz, I. M. Held, H. Guo, and . co authors, 2018a: The GFDL global atmosphere and land model AM4.0/LM4.0 – Part II: Model description, sensitivity studies, and tuning strategies. *J. of Adv. Model. Earth Syst.*, **10**, doi:10.1002/2017MS001209.

Zhao, M., J.-C. Golaz, I. M. Held, H. Guo, and co authors, 2018b: The GFDL global atmosphere and land model AM4.0/LM4.0 – Part I: Simulation characteristics with prescribed SSTs. *J. of Adv. Model. Earth Syst.*, **10**, doi:10.1002/2017MS001208.

Zhu, Y., and Coauthors, 2018: Toward the improvement of subseasonal prediction in the national centers for environmental prediction global ensemble forecast system. *Journal of Geophysical Research: Atmospheres*, **123 (13)**, 6732–6745.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

TABLE 1: Climate models analyzed in this study, summarized in terms of corresponding modeling center (Col. 1), version name (Col. 2) and reference (Col. 3).

Modeling Group	Model	Reference
Department of Energy (DOE)	E3SMv1	Golaz et al. (2019); Xie et al. (2018)
NOAA Geophysical Fluid Dynamics Laboratory (GFDL)	CM3, CM4, ESM4	Griffies et al. (2011); Held et al. (2019); Dunne et al. (2020)
NASA Goddard Institute for Space Studies (GISS)	GISS E2-R/H, E2.1-G/H, E2.2-G	Schmidt et al. (2014); Kelley et al. (2020); Rind et al. (2020)
NASA Global Modeling and Assimilation Office (GMAO)	GEOS-5	Molod et al. (2015); Borovikov et al. (2017)
National Center for Atmospheric Research (NCAR)	CESM(1,2)(CAM/WACCM(5,6))	Gent et al. (2011); Hurrell et al. (2013b); Mills et al. (2017); Danabasoglu et al. (2020)
NOAA National Center for Environmental Prediction (NCEP)	CFS v2	Saha et al. (2014)

Accepted for publication in *Journal of Climate*. DOI10.1175/JCLI-D-19-0956.1.

TABLE 2: Details of atmospheric model components considered in this study. Listed is model name (Col. 1), number of vertical levels and distribution within the troposphere/stratosphere/mesosphere (Col. 2), model top (Col. 3), horizontal resolution (Col. 4), references describing the convection schemes (Col. 5) and references describing the gravity wave drag scheme (Col. 6).

Model	Vertical Layers (Total/Trop/Strat+Mes)	Model Top (hPa)	Horizontal Resolution	Convection Scheme	Gravity Wave Drag
NCAR-CESM1 (CAM5)	32/24/8	3.6	1 degree	Zhang and McFarlane (1995) Park and Bretherton (2009)	McFarlane (1987) Richter et al. (2010)
NCAR-CESM1 (WACCM5)	70/24/28	6×10^{-6}	1 degree	Zhang and McFarlane (1995) Park and Bretherton (2009)	McFarlane (1987) Richter et al. (2010)
NCAR-CESM2 (CAM6)	32/22/10	3.6	1 degree	Updated ZM95 Golaz et al. (2002)	Scinocca and McFarlane (2000) Richter et al. (2010)
NCAR - CESM2 (WACCM6)	70/24/28	6×10^{-6}	1 degree	Updated ZM95 Golaz et al. (2002)	Scinocca and McFarlane (2000) Richter et al. (2010)
DOE-E3SM1	72/47/25	0.01	1 degree	Xie et al. (2018) Golaz et al. (2002)	McFarlane (1987) Richter et al. (2010)
GFDL-CM3	48/23/25	0.01	2 degree	Bretherton et al. (2004) Donner et al. (2001)	Stern and Pierrehumbert (1988) Alexander and Dunkerton (1999)
GFDL-CM4	33/24/9	1	1 degree	Zhao et al. (2018a)	Garner (2005) Alexander and Dunkerton (1999)
GFDL-ESM4	49/24/25	0.01	1 degree	Zhao et al. (2018a)	Garner (2005) Alexander and Dunkerton (1999)
GISS - E2	40/25/15	0.1	2.5 degrees	Del Genio et al. (2007)	Schmidt et al. (2014)
GISS-E2.1	40/25/15	0.1	2.5 degrees	Kim et al. (2013) Del Genio et al. (2015)	Schmidt et al. (2014)
GISS - E2.2	102/58/44	0.002	2.5 degrees	Kim et al. (2013) Del Genio et al. (2015)	Rind et al. (2014) Rind et al. (2020)
GEOS-M2AMIP	72/35/37	0.01	50 km	Moorthi and Suarez (1992)	McFarlane (1987) Garcia and Boville (1994)
GEOS-S2S	72/35/37	0.01	0.5 degrees	Moorthi and Suarez (1992)	McFarlane (1987) Garcia and Boville (1994)
NCEP GEFS	64/43/21	0.2	T574/T384	Saha et al. (2014)	Chun and Baik (1998)

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

TABLE 3: Main simulation experiments analyzed in this study described in terms of submitting modeling center (Col.1), model version (Col. 2), simulation type (Col.3), number of ensemble members (Col. 4), coupling to the ocean (Col. 5), and DOI where relevant.

Modeling Center	Version	Type	Ensemble Size	AMIP/Coupled	DOI
NCAR	CCSM4	Historical	6	Coupled	10.1594/WDCC/CMIP5.NRS4hi
	CESM1 (CAM5)	Historical	3	Coupled	10.1594/WDCC/CMIP5.NFCChi
	CESM1 (BGC)	Historical	1	Coupled	10.1594/WDCC/CMIP5.NFCBhi
	CESM1 (WACCM5)	Historical	7	Coupled	10.1594/WDCC/CMIP5.NFCWhi
	CESM2 (CAM6)	Historical	6	Coupled	10.22033/ESGF/CMIP6.7627
		Intermediary	2	Coupled	N/A
	CESM2 (WACCM6)	Historical	6	Coupled	10.22033/ESGF/CMIP6.11298
GISS	E2-R	Historical	18	Coupled	10.1594/WDCC/CMIP5.GIGRhi
	E2-R-CC	Historical	1	Coupled	10.1594/WDCC/CMIP5..GIHChi
		Intermediary	1	Coupled	N/A
	E2-H	Historical	18	Coupled	10.1594/WDCC/CMIP5.GIGHhi
	E2-H-CC	Historical	1	Coupled	10.1594/WDCC/CMIP5..GIRChi
	E2.1-G	Historical	20	Coupled	10.22033/ESGF/CMIP6.1400
	E2.1-H	Historical	20	Coupled	10.22033/ESGF/CMIP6.1421
	E2.2-G	AMIP	5	Atm.	10.22033/ESGF/CMIP6.6986
		Historical	3	Coupled	10.22033/ESGF/CMIP6.2081
GEOS	M2AMIP	Historical	10	Atm.	N/A
	S2S-v2	45-day Forecasts	4	Coupled	N/A
DOE	E3SMv1	Historical	5	Coupled	10.22033/ESGF/CMIP6.4497
		AMIP	1	Atm.	10.22033/ESGF/CMIP6.4492
	E3SMv1-MODGWD	Intermediary	1	Atm.	N/A
			1	Coupled	N/A
GFDL	CM2.1	Historical	10	Coupled	10.1594/WDCC/CMIP5.NGG2hi
	CM3	Historical	5	Coupled	10.1594/WDCC/CMIP5.NGG3hi
	ESM2G	Historical	1	Coupled	10.1594/WDCC/CMIP5.NGEGhi
	ESM2M	Historical	1	Coupled	10.1594/WDCC/CMIP5.NGEMhi
	CM4	Historical	3	Coupled	10.22033/ESGF/CMIP6.8594
	ESM4	Historical	3	Coupled	10.22033/ESGF/CMIP6.8597
NOAA	GEFS	35-day Forecasts	11	Atm.	10.7916/D8PG249H

Accepted for publication in *Journal of Climate*. DOI10.1175/JCLI-D-19-0956.1.

TABLE 4: Analysis approach used for evaluating modes of variability, described in terms of mode (Col. 1), observational product (Col. 2), time period of analysis (Col. 3) and output used for analyses (Col. 4). *ERA20C (ERA 20th Century Reanalysis) used for the evaluation of ENSO teleconnections. **20CR (NOAA 20th Century Reanalysis) used for evaluating the SAM, NAM, PNA, NAO and PDO. *** Years 1956-2005 were used for the SAM analysis.

Mode	Observation Product	Years	Output for Analysis
MJO	TRMM, ERA5	1998-2014	daily precipitation, daily zonal winds (U) at 850 hPa
QBO	MERRA-2	1980-2016	monthly zonal winds (U) (10-100 hPa)
ENSO and PDO	ERSSTv5, HadISST,	1920-present	monthly sea level pressure (slp)
	ERA20C*/ERAI, BEST, 20CR**		and surface temperature (ts)
SAM, NAM, NAO	NOAA 20CR**	1900-2005***	monthly sea level pressure (slp)

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

LIST OF FIGURES

Accepted for publication in *Journal of Climate*. DOI10.1175/JCLI-D-19-0956.1.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

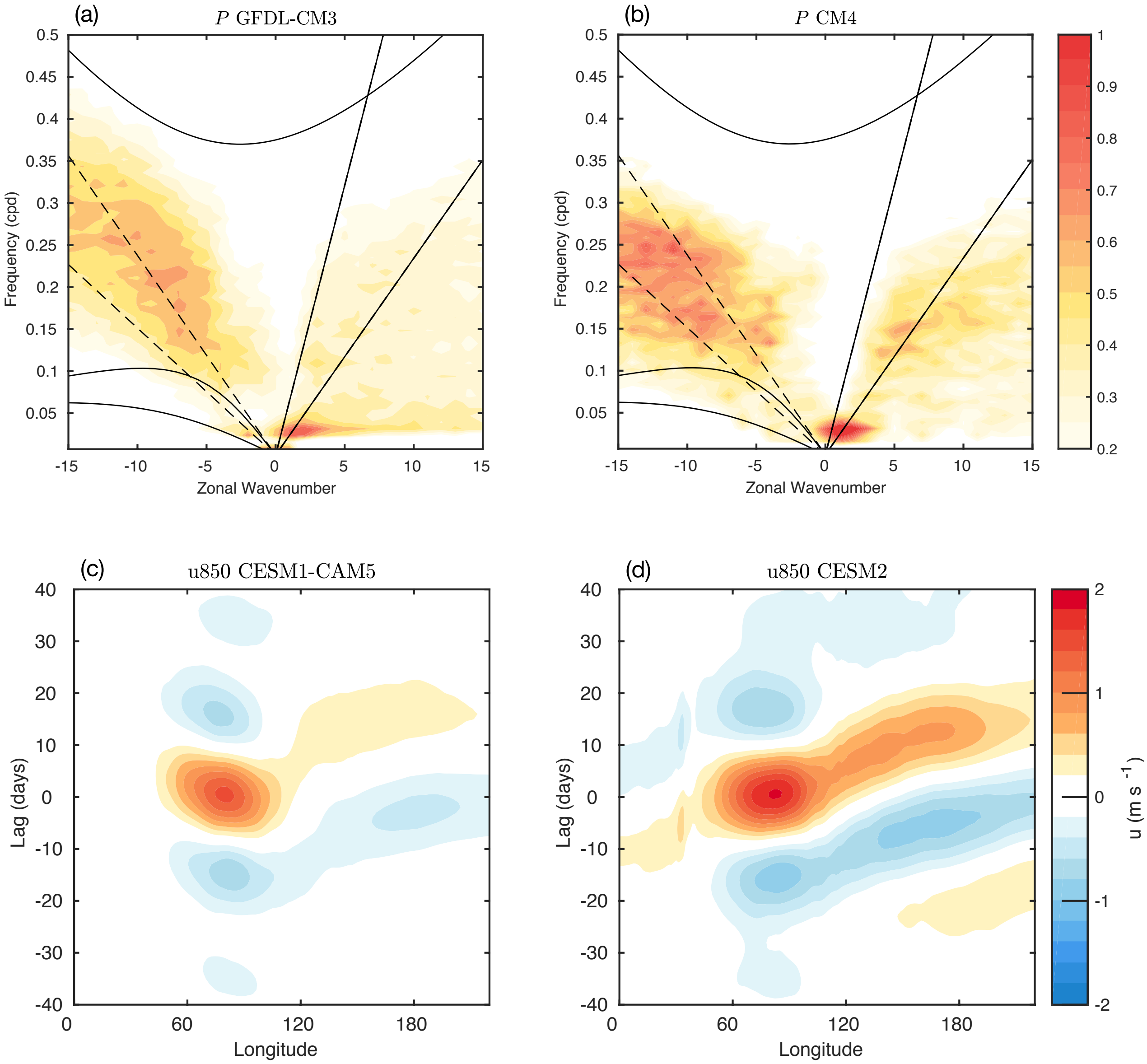

FIG. 1: Top: Signal strength of precipitation for (a) CM3 and (b) CM4. The solid lines in (a) and (b) correspond to the dispersion curves of the equatorial wave solutions for equivalent depths of 12 m and 90 m, respectively. The dashed lines correspond to constant phase speeds of 7 m s^{-1} and 11 m s^{-1}. Bottom: Time-longitude diagrams of zonal winds at 850 hPa (u850) averaged over 15°N/S for CESM1(CAM5) and CESM2(CAM6).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

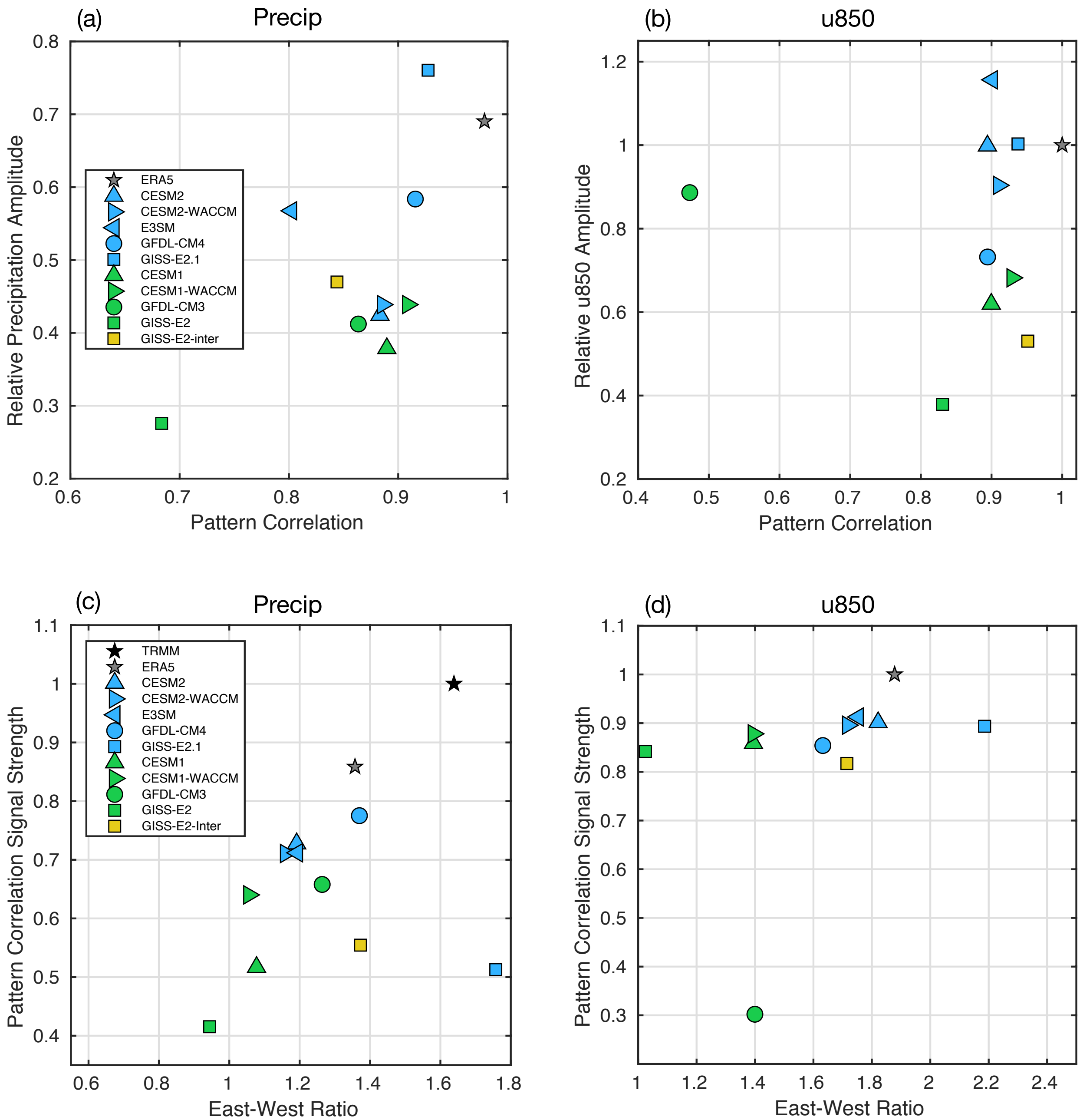

FIG. 2: Top: Scatter plots of pattern correlation (x-axis) versus relative precipitation amplitude (y-axis) (top panels) and east-west ratio (x-axis) versus the pattern correlation of the signal strength (y-axis) (bottom panels) for CMIP6 versions of the models considered in this study (blue), CMIP5 models (green), and the intermediate GISS model (yellow). The different shapes correspond to the model center: the GFDL models are shown as circles, GISS is shown as squares, NCAR's CESM-CAM is shown as an upward-pointed triangle while CESM-WACCM is shown as a rightward-pointed triangle. E3SM is denoted by the leftward-pointed (blue) triangle.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

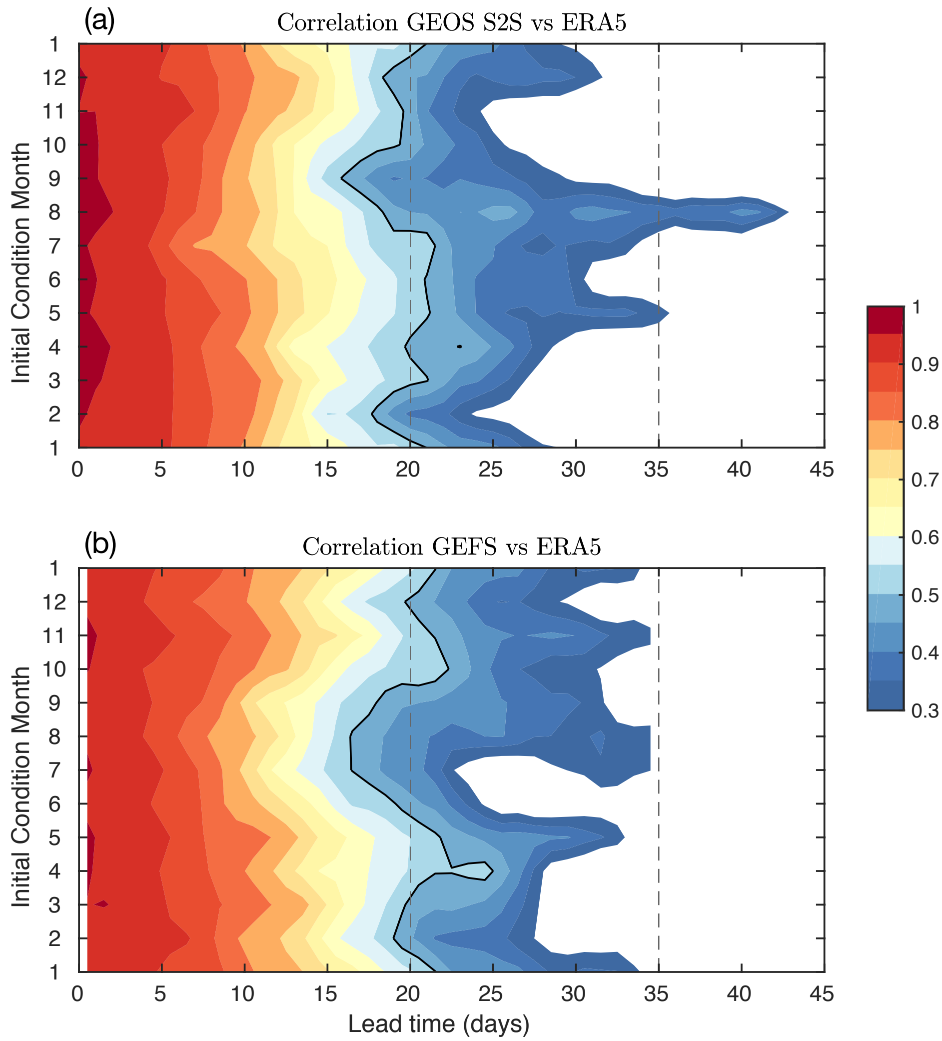

FIG. 3: Bivariate anomaly correlation of the RMM indices from ERA5 and ensemble means from (a) NOAA GEFS and (b) GEOS-S2S as a function of forecast lead day and the month of initialization of the forecast. The dashed line at day 35corresponds to the longest lead time in the GEFS dataset, while the line at day 20 is shown for reference. Note that the January is shown in the top and bottom to elucidate the wintertime variability in each model more clearly.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

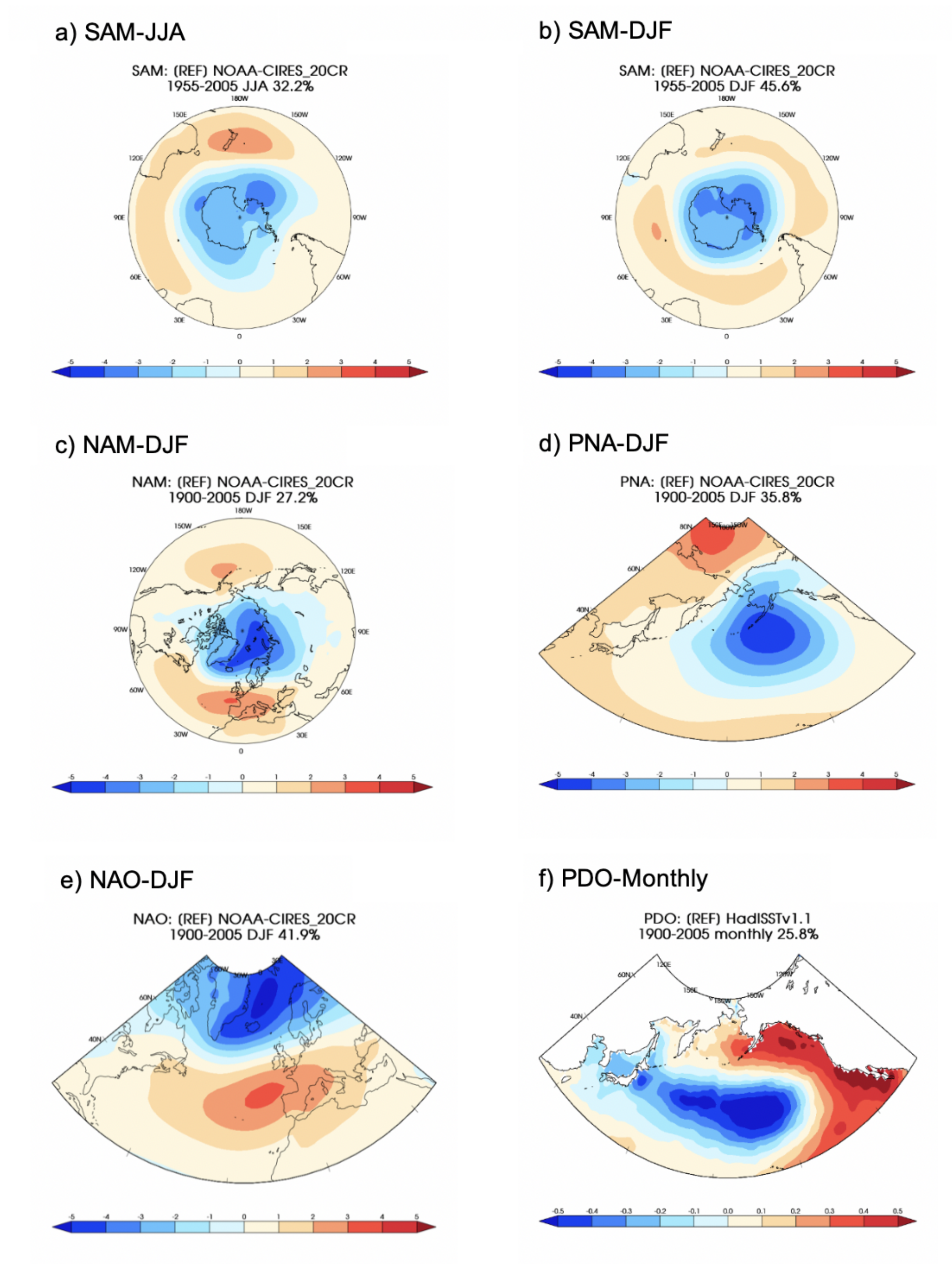

FIG. 4: Observed modes (EOF-1) of sea-level pressure variability for the (a) Southern Annular Mode (SAM) during JJA, (b) the SAM during DJF, (c) the Northern Annular Mode (NAM) during DJF, (d) the Pacific-North American Pattern (PNA) during DJF and (e) the North Atlantic Oscillation (NAO) during DJF, based on anomalies from the 20$^{\text{th}}$ Century Reanalysis (20CR). Monthly sea surface temperature anomalies from HadISSTv1.1 are used for the Pacific Decadal Oscillation (PDO) (f). Maps show the positive phases of the individual modes and the percentage of total variance (%) explained by the EOF is noted at the top of each plot.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

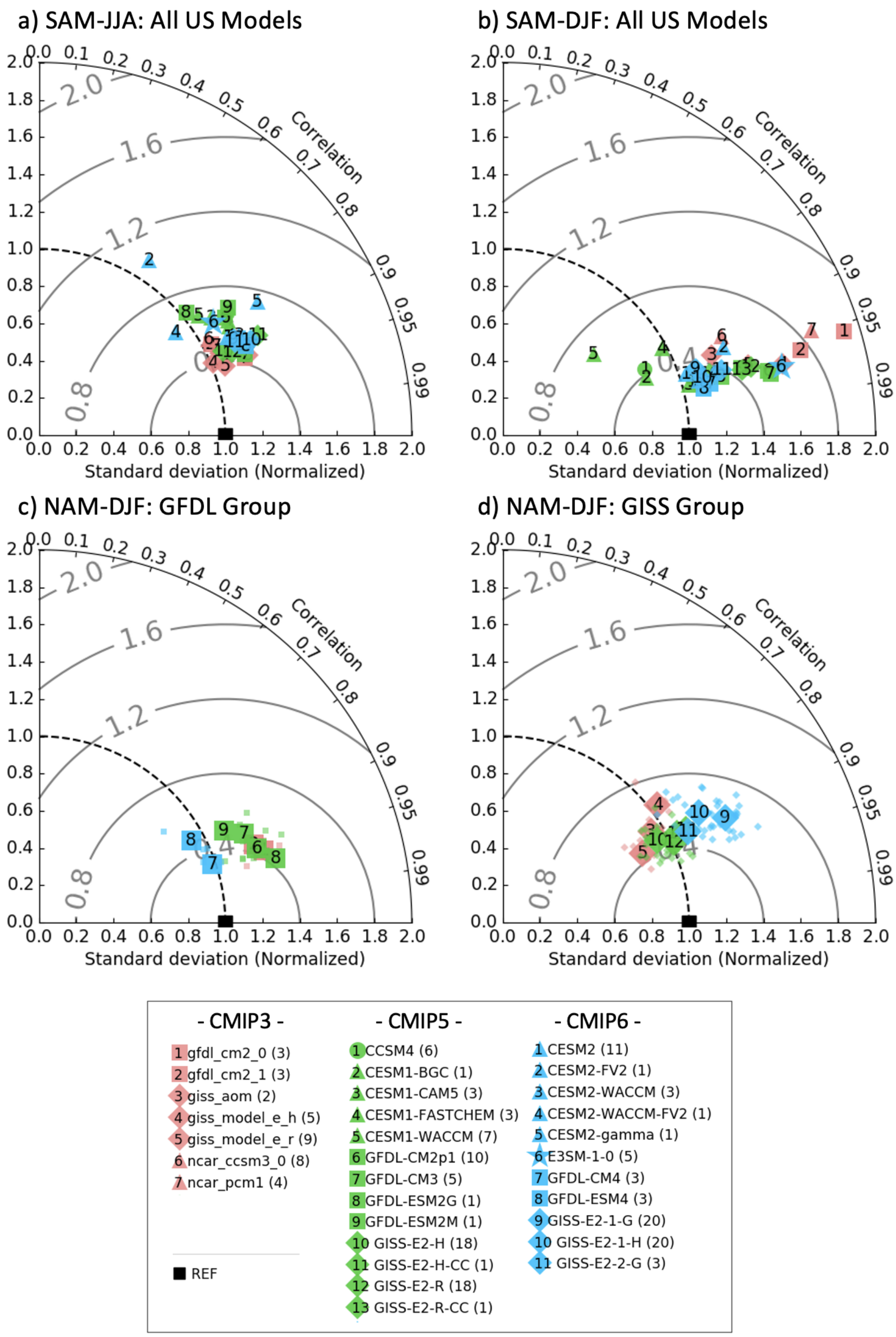

FIG. 5: Taylor Diagrams illustrating model skill for the SAM for all models during (a) JJA and (b) DJF; skill for the NAM during DJF is shown in (c) and (d) for GFDL and GISS models, respectively. Red, green and blue represent CMIP3, CMIP5 and CMIP6 generations of the US models, respectively. The black squares on all abscissa represent the observationally-based references used for evaluating skill. Larger symbols represent statistics averaged across multiple realizations for a given model, with the number of realizations shown in parenthesis after model labels in the legend. Smaller symbols in the panels indicate results from individual realizations.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

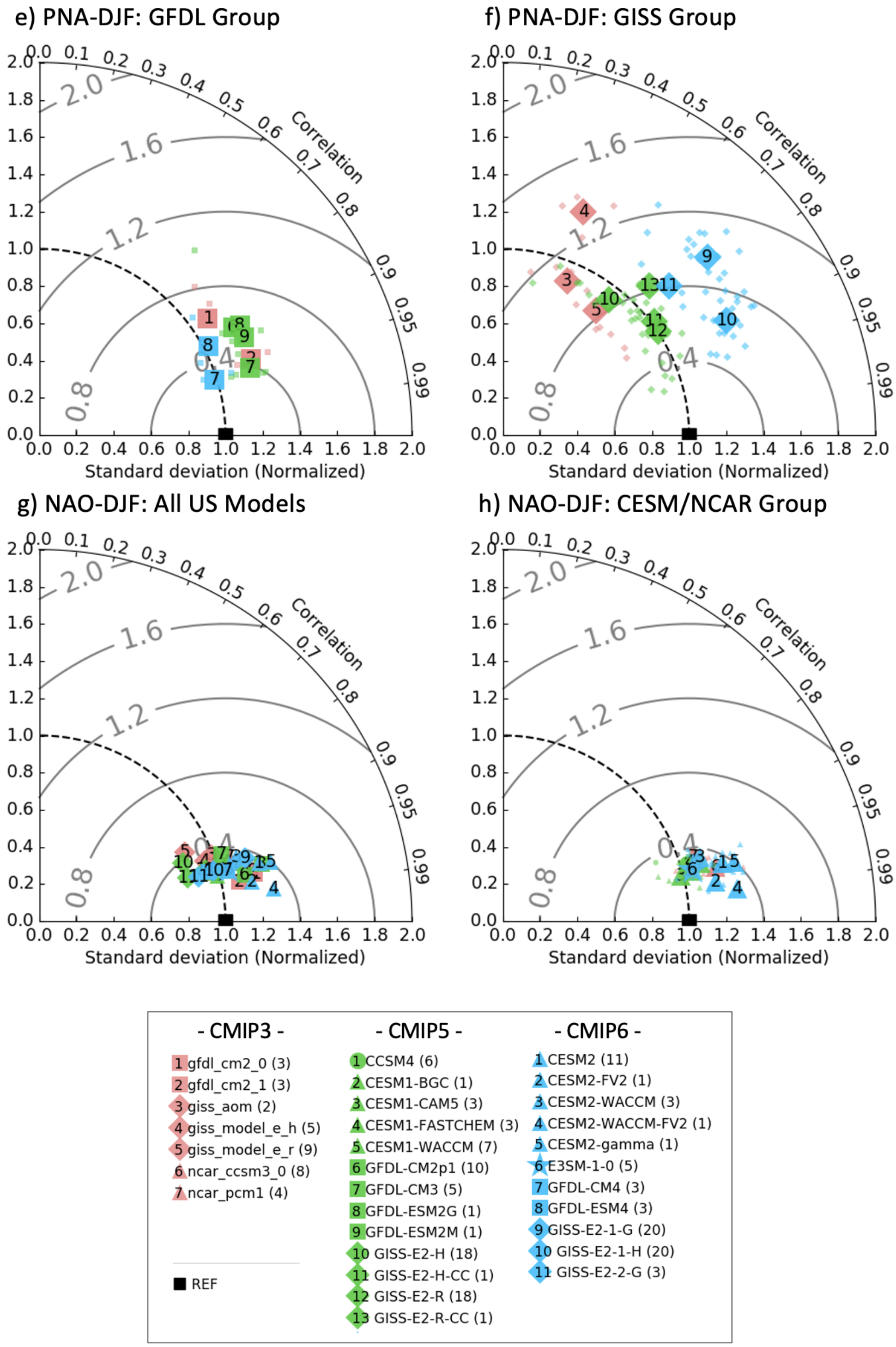

FIG. 5: (cont) As above, but for the PNA for (e) GFDL and (f) GISS model versions. The NAO for boreal winter is shown for all models in (g) and the NCAR subset of models in (h).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

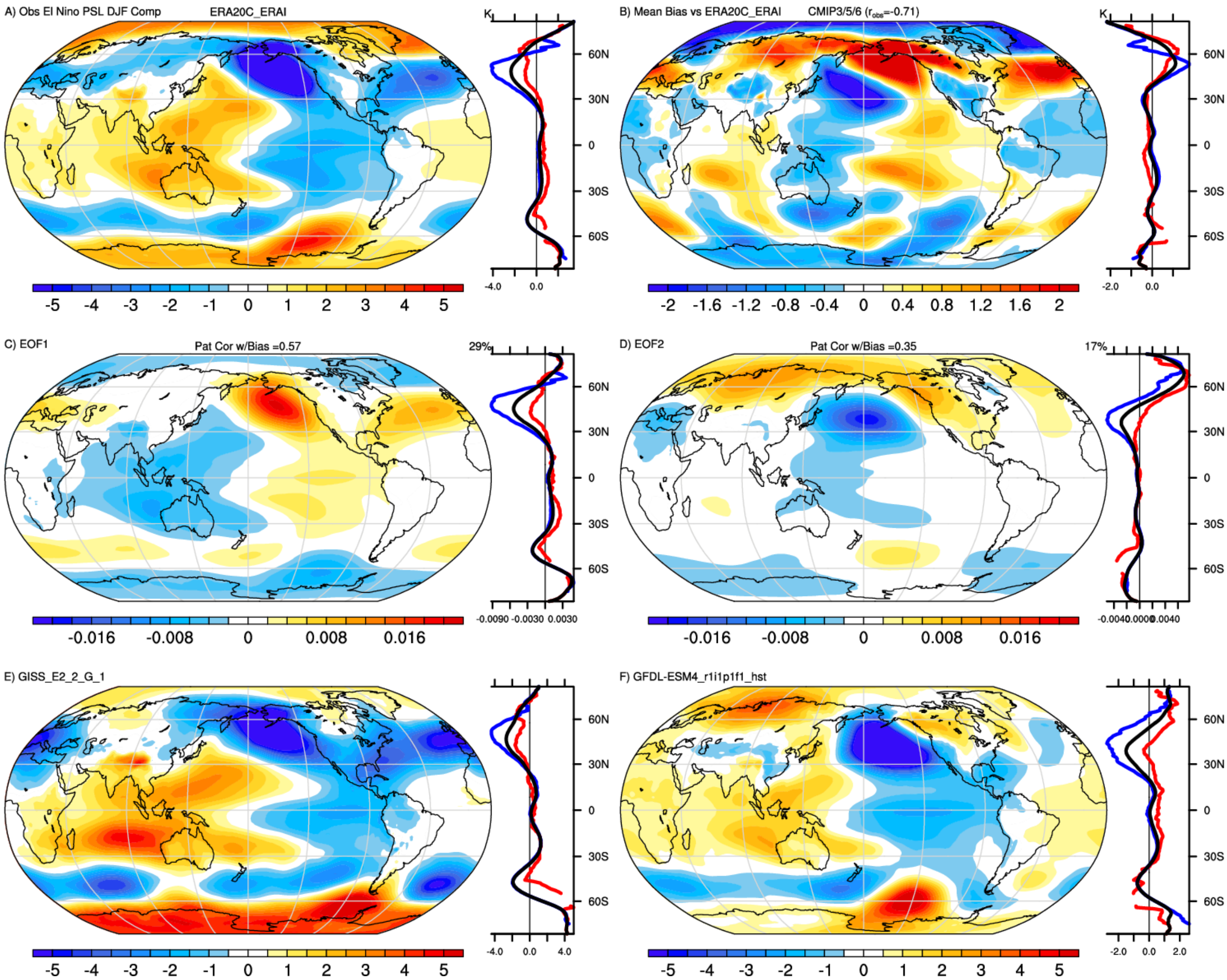

FIG. 6: Composites of DJF sea level pressure for El Niño minus La Niña in observations (A, 1920–2017, see methods) and the mean bias for climate model simulations (from 1900) in the CMIP archives (B, units of hPa). Zonal mean values are also indicated over ocean (blue), land (red), and combined (black). The first (C) and second (D) EOFs of model bias are also shown to illustrate the leading patterns that distinguish simulated modes. The US CMIP6 simulations with the (E) least and (F) greatest difference in EOFs 1 and 2 from observations are also shown.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

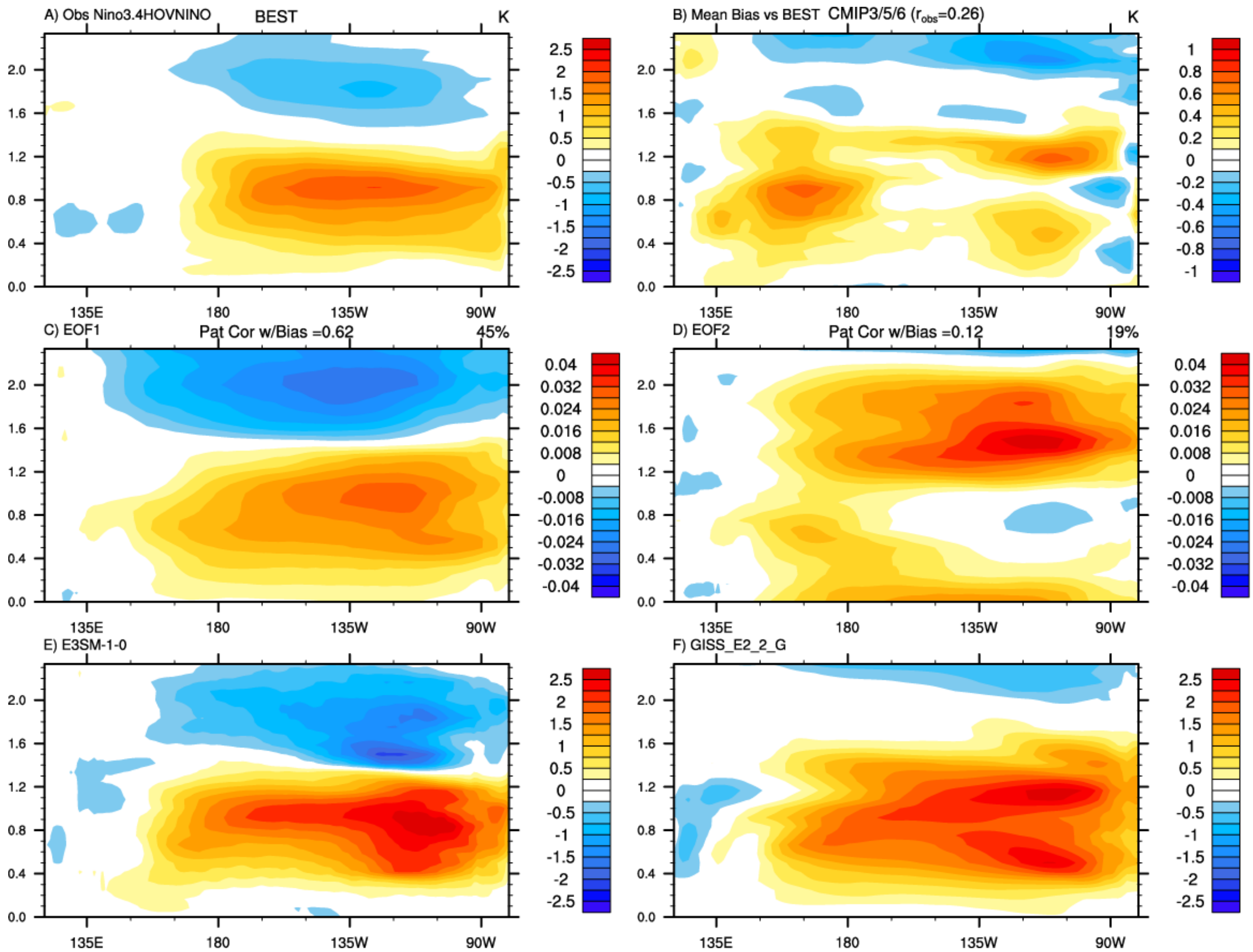

FIG. 7: Hovmöller diagram of surface temperature anomalies during El Niño events (as defined in main text) based on an observational estimate (A, Berkeley Earth, 1920–2017; Rohde et al. (2013)) and the mean CMIP model bias (B, since 1900). The leading patterns differentiating models are shown in C) EOF1 and D) EOF2 and the US CMIP6 simulations with the (E) least and (F) greatest difference in EOFs 1 and 2 from observations are also shown.

Accepted for publication in *Journal of Climate*. DOI10.1175/JCLI-D-19-0956.1.

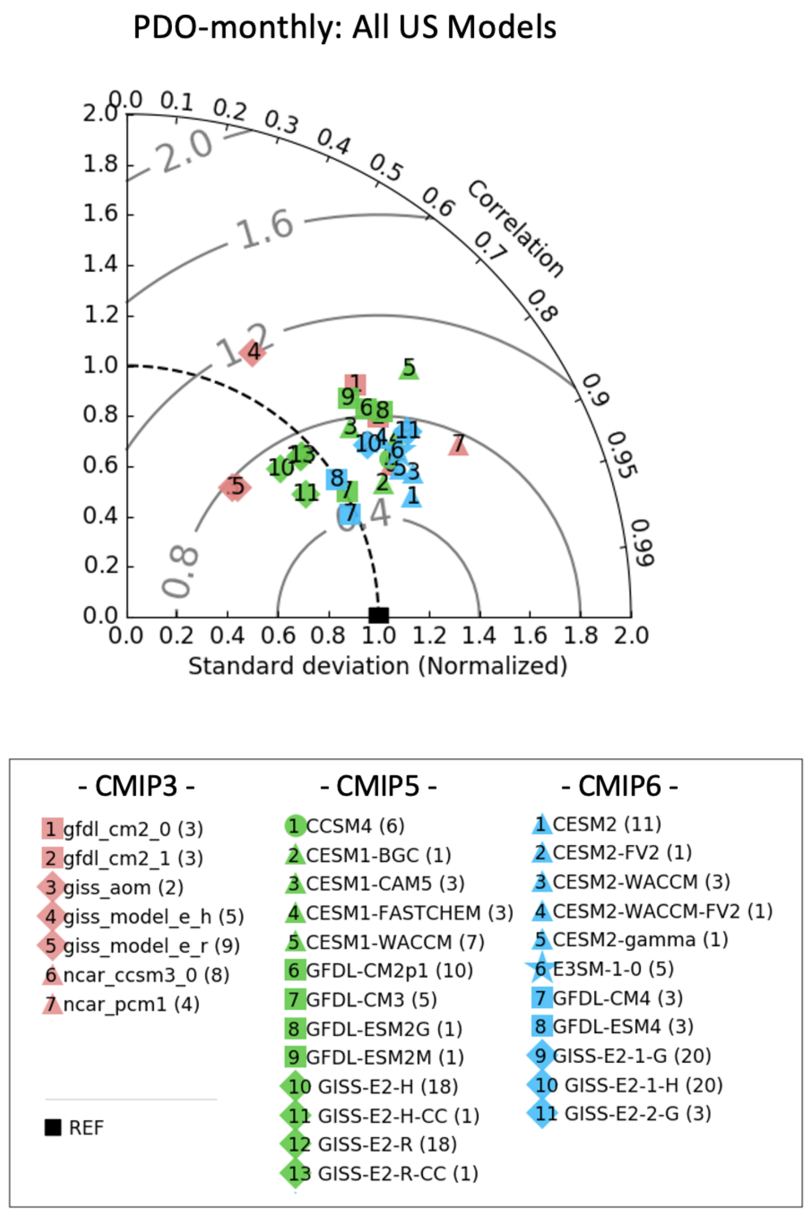

FIG. 8: As in Fig. 5 but for the Pacific Decadal Oscillation for all models.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

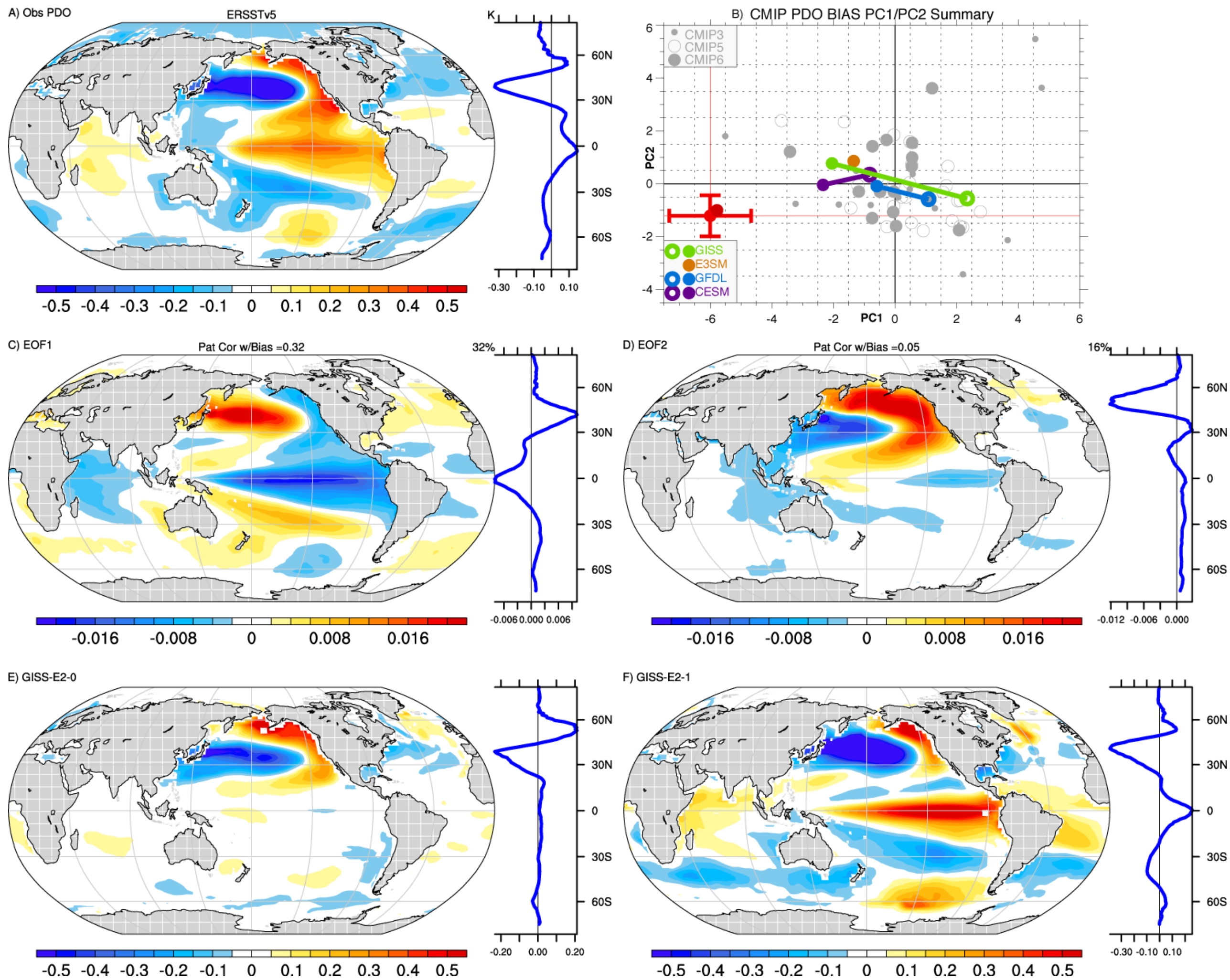

FIG. 9: A) Spatial pattern of the observed Pacific Decadal Oscillation based on NOAAs ERSSTv5 (1920–2017) and its zonal mean structure (blue line). Evolving values of the principle components of mode bias across US climate models (colors, from 1920) relative to other CMIP simulations (grey) and observations (red) (B). Note that open circles denote CMIP5 versions and large (small) closed circles denote CMIP6 (CMIP3) versions. The structure of the leading bias EOF (C) and second EOF (D) are also shown along with the PDO patterns in CMIP5 and CMIP6 versions of the GISS model. Red symbols in (B) denote the PC values for observations (ERSSTv5 and HadISST-dark red) and estimated 2σ range of internal variability based on the CESM Large Ensemble.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

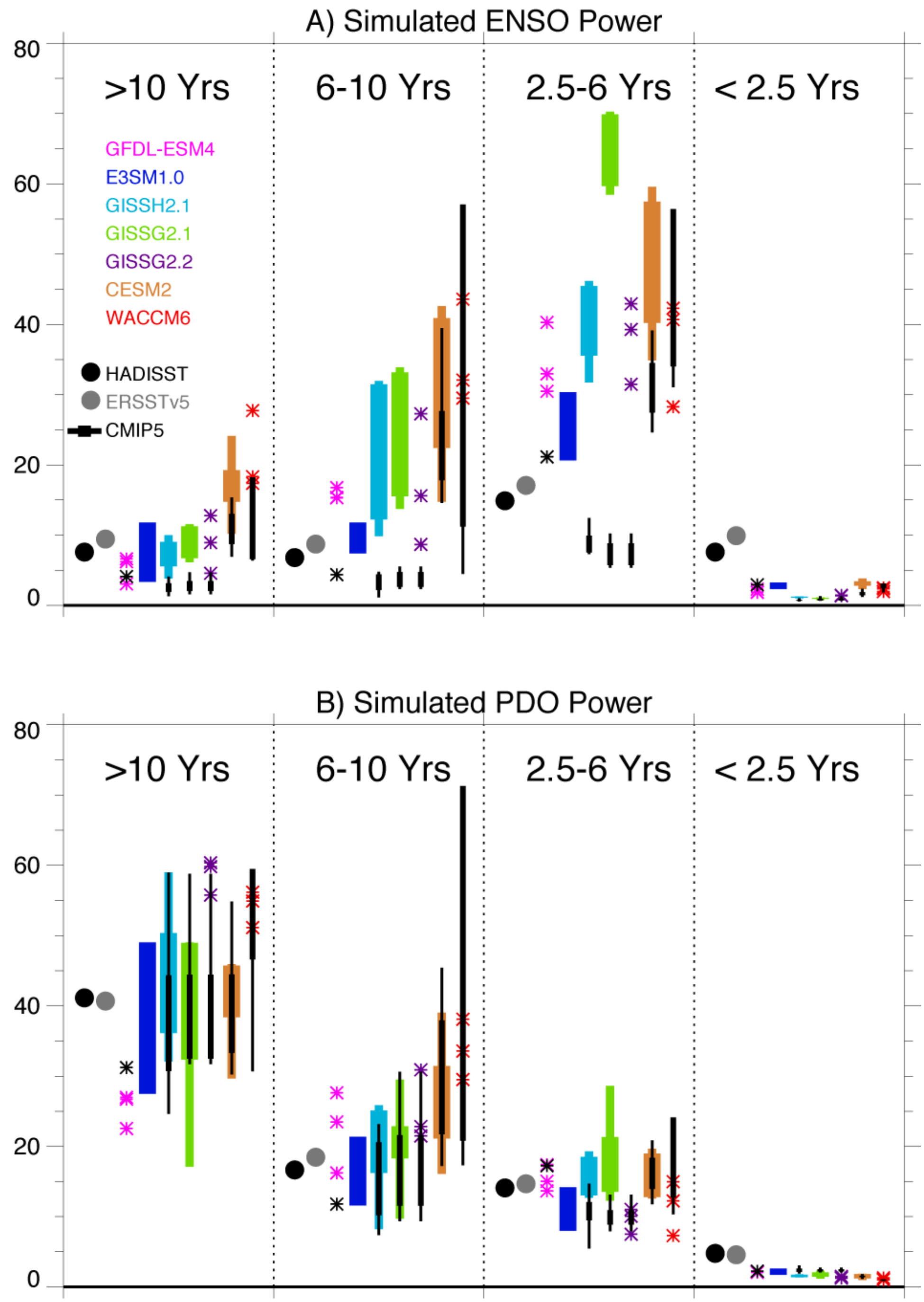

FIG. 10: Power of major coupled modes of variability in US climate models including A) Nino3.4 SSTa and B) the PDO timeseries across various bands. Thick lines indicate the interquartile range and thin lines indicate the full ensemble range for each model where at least 5 simulations are available while asterisks denote values for individual members of other models. Also shown are observed estimates from the Hadley Centre (black circle) and NOAA ERSSTv5 SSTa products. Analogous ranges for the corresponding CMIP5 model versions (i.e. from the same center) are shown in thinner black lines.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

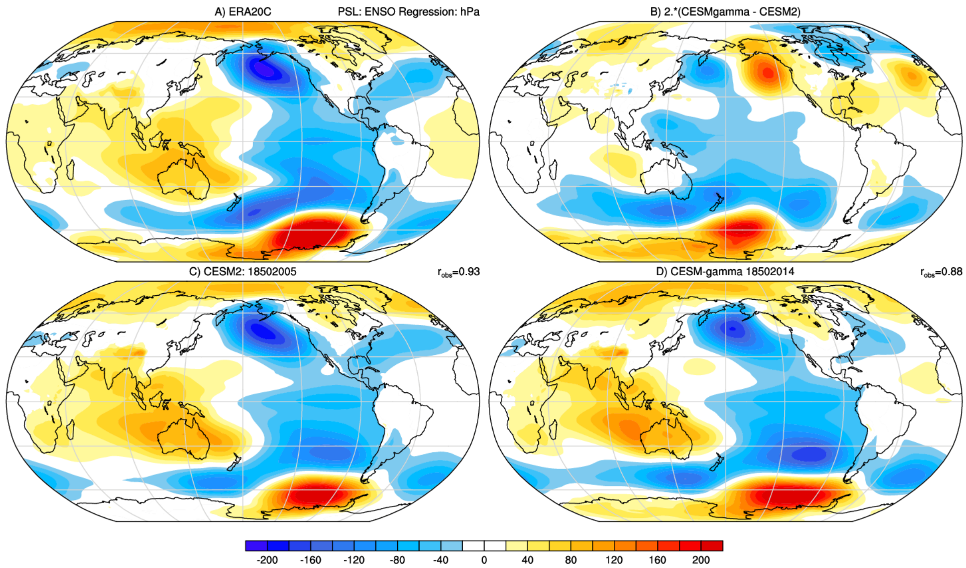

FIG. 11: A) Regression between SLP and Nino3.4 SSTa for observations (ERA20C, 1920–2017) and B) the difference between the same regressions for CESM2 (regression is shown in C) and CESM2-gamma (regression shown in D, 1900-2005). The difference field (B) has been multiplied by two in order to use one common color bar.

Accepted for publication in *Journal of Climate*. DOI10.1175/JCLI-D-19-0956.1.

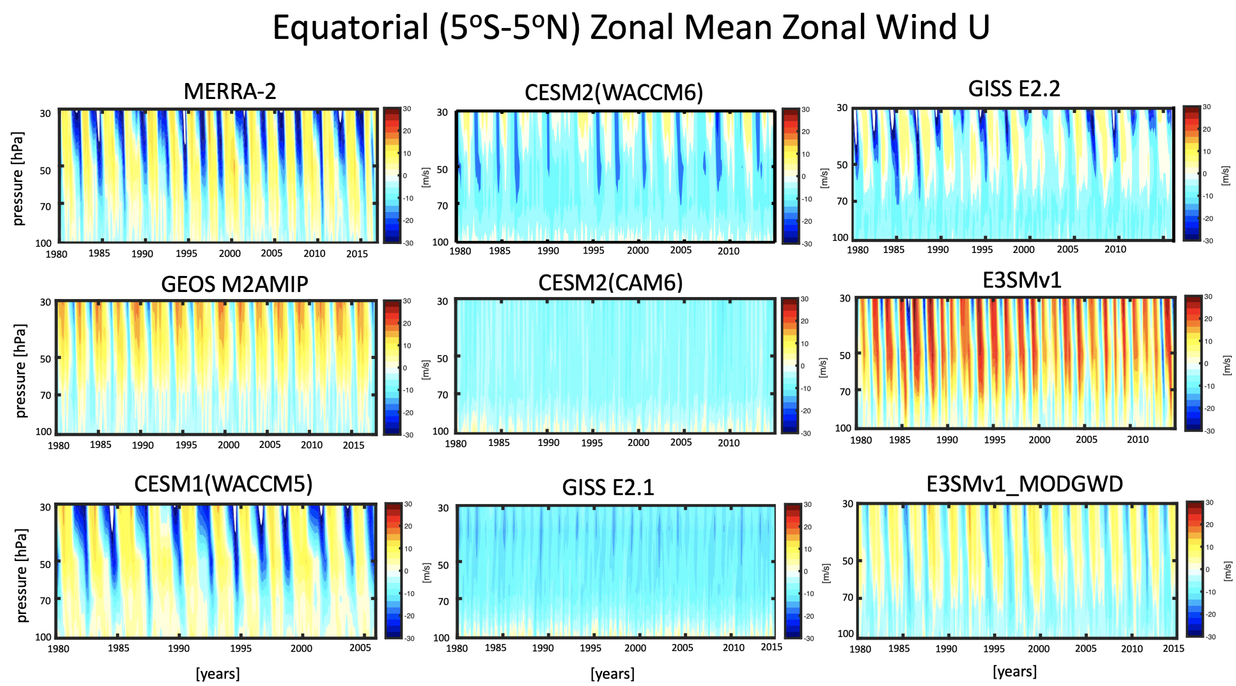

FIG. 12: Evolution of the equatorial (5°S-5°N averaged) zonal mean zonal winds for the various models considered for QBO evaluation. MERRA-2 (a) is treated as the reference against which the GEOS M2AMIP, CESM1(WACCM5), CESM2(WACCM6), CESM2(CAM6), GISS E2.1, GISS E2.2, E3SMv1 and E3SMv1-MODGWD are compared. For models providing ensembles only one member is shown in order to avoid averaging over (phase-lagged) oscillations among different members.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

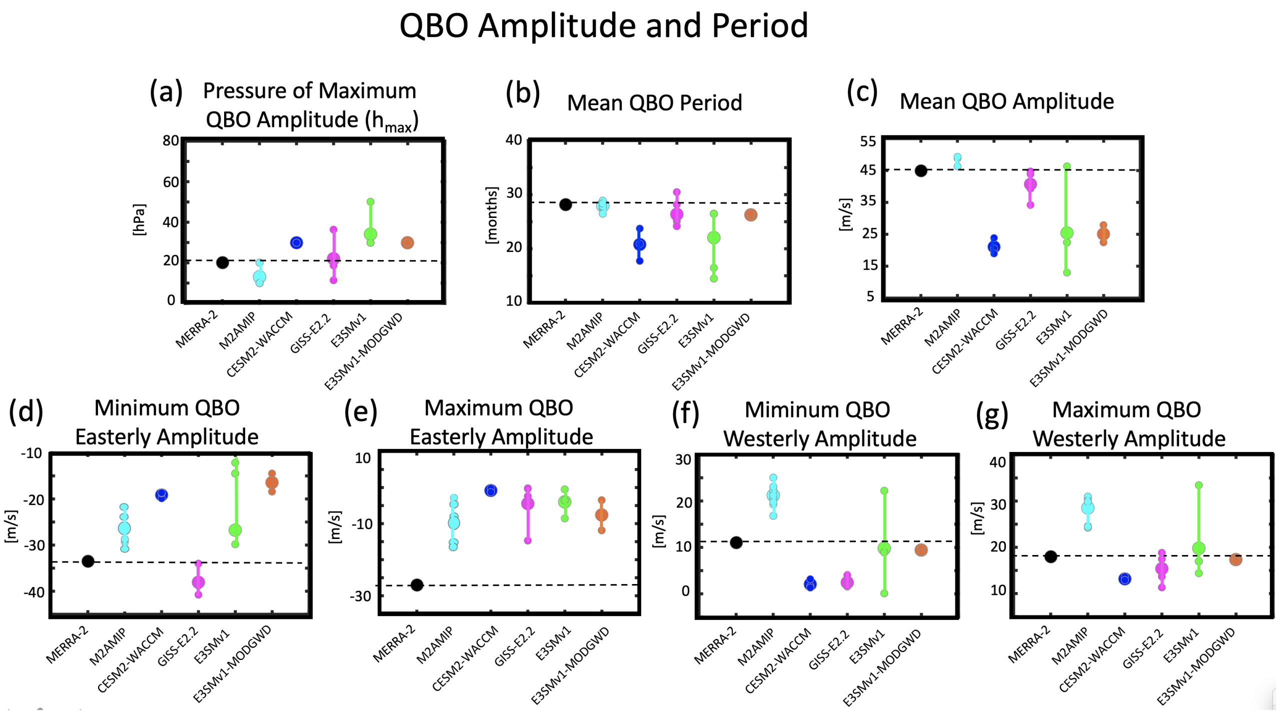

FIG. 13: Comparison of different measures of the QBO ranging from (a) $h_{\max}$, the pressure at which the squared Fourier amplitudes ranging from 26–30 months of the equatorial zonal mean winds maximizes, (b) the mean QBO period, (c) the mean QBO amplitude, (d,e) the maximum (minimum) QBO amplitude occurring during the easterly phase of the QBO and (f,g) the maximum (minimum) QBO amplitude occurring during the westerly phase. Small (large) circles denote individual ensemble members (ensemble means) while lines span the ensemble range. Note that the results for CESM2(CAM6), GISS E2.1 and CESM1(WACCM5) are not shown since the first two models do not simulate a QBO and the QBO is prescribed in the latter

.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

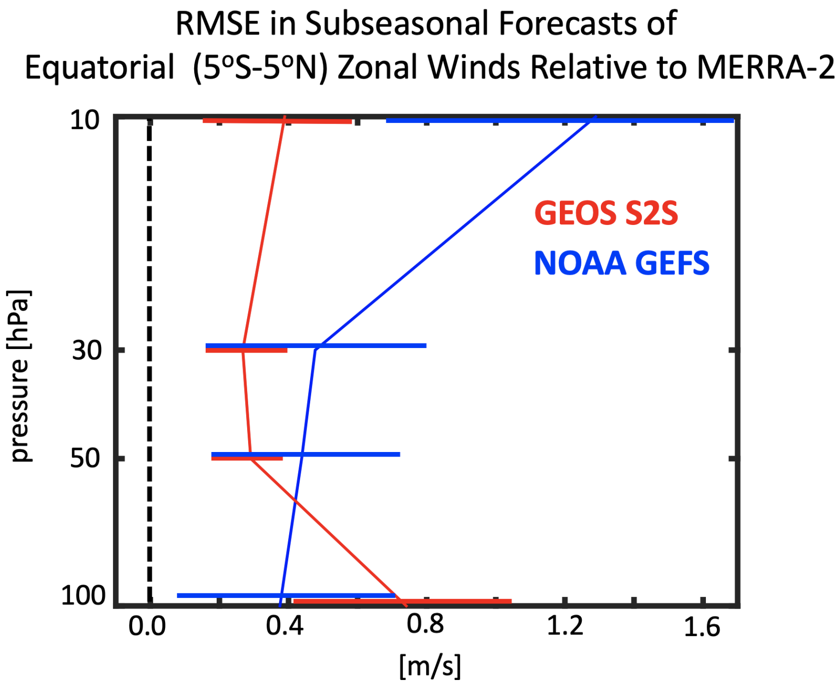

FIG. 14: Root mean square error (RMSE) of the equatorial (5°S-5°N) zonally averaged zonal winds, compared between the 45-day-long GEOS-S2S and 35-day-long NOAA GEFS subseasonal forecasts and evaluated relative to MERRA-2. RMSE values have been calculated using the ensemble mean values over the entire course of the forecasts (up to 35 days for both GEFS and S2S), for all months and years within the climatological period 2000–2010. Horizontal bars denote the spread in error associated with both seasonal and interannual variations.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.

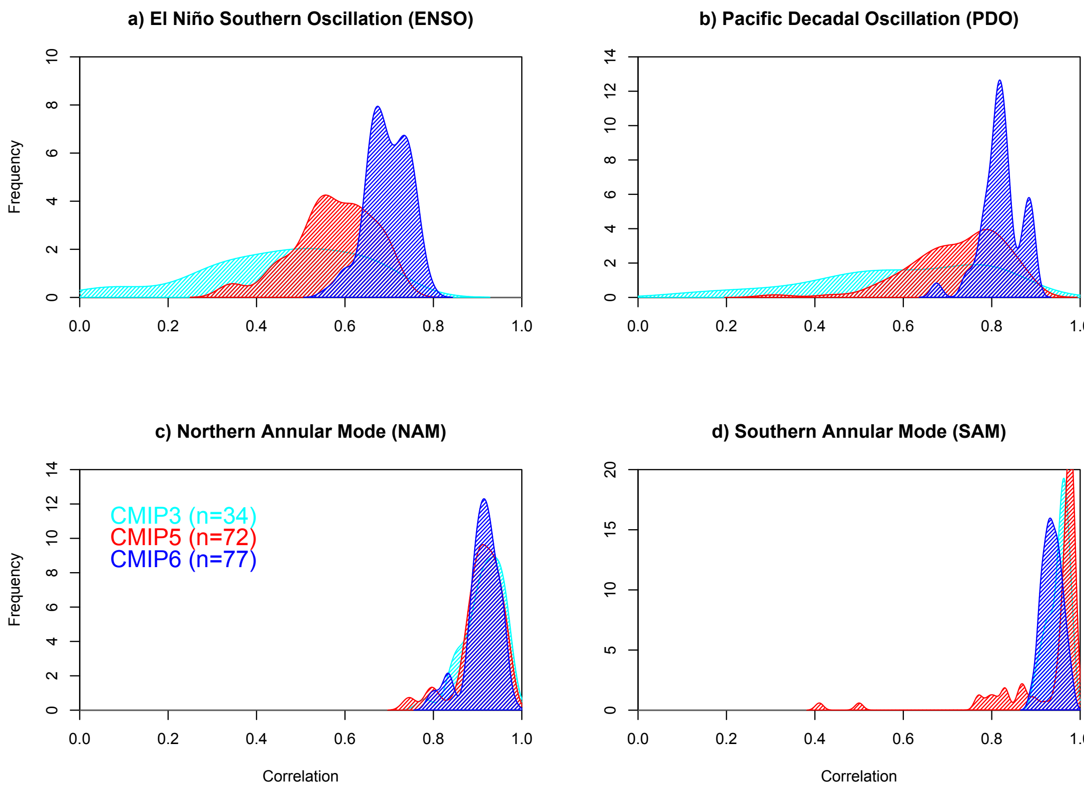

FIG. 15: Summary of correlations across the CMIP3/5/6 model ensembles (each simulation is weighted equally, with the number of simulations given in the legend) for the US models relative to observations for the a) El Niño Southern Oscillation (surface temperatures), b) the Pacific Decadal Oscillation (SLP), c) the Northern Annular Mode (SLP), and d) the Southern Annular Mode (SLP).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-19-0956.1.