Modes of Variability in E3SM and CESM Large Ensembles

John T. Fasullo[1], Julie M. Caron[1], Adam Phillips[1], Hui Li[1], Jadwiga H. Richter[1], Richard B. Neale[1], Nan Rosenbloom[1], Gary Strand[1], Sasha Glanville[1], Yuanpu Li[1], Flavio Lehner[1], Gerald Meehl[1], Jean-Christophe Golaz[2], Paul Ullrich[2], Jiwoo Lee[2], Julie Arblaster[3]

1. *National Center for Atmospheric Research, Boulder, CO*
2. *Lawrence Livermore National Laboratory, Livermore, CA*
3. *School of Earth Atmosphere and Environment, Monash, University, AUS*

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

ABSTRACT

An adequate characterization of internal modes of climate variability (MoV) is prerequisite for both accurate seasonal predictions and the attribution and detection of forced climate change in nature. Assessing the fidelity of climate models in simulating MoV is therefore essential; however, doing so is complicated by the large intrinsic variations in MoV and the limited span of the observational record. Large ensembles (LEs) provide a unique opportunity to assess model fidelity in simulating MoV and quantify inter-model contrasts. In this work, these goals are pursued in four recently produced LEs: the Energy Exascale Earth System Model (E3SM) versions 1 and 2 LEs, and the Community Earth System Model (CESM) versions 1 and 2 LEs. In general, the representation of MoV is found to improve across successive E3SM and CESM versions concurrent with improved simulation of the base state climate. The patterns of global coupled modes and many extratropical modes are well simulated by both E3SM2 and CESM2, though various persistent shortcomings are identified. The results both demonstrate the successes of these recent model versions and suggest the potential for continued improvement in the representation of MoV with advances in model physics.

SIGNIFICANCE STATEMENT

Modes of variability play a critical role in prediction of seasonal to decadal climate variability and detection of forced climate change, but historically many modes have been poorly simulated by coupled climate models. Using recently produced large ensembles, this work demonstrates the improved simulation of a broad range of internal modes in successive versions of the E3SM and CESM and discusses opportunities for further advances.

1. Introduction

Adequate representation of internal climate variability is critical for both seasonal to decadal prediction (Robertson et al. 2015; Thoma et al. 2015; Meehl et al 2016; Simpson et al. 2019; Krishnamurthy et al. 2021), multidecadal prediction (Deser et al. 2012, 2014, 2017, Kumar and Ganguly 2018; Dai and Bloecker 2019), and climate attribution and detection efforts (Santer et al. 2009; Stott et al. 2010; Schurer et al. 2013; Deser et al. 2016; McKinnon and Deser 2018). Internal variability can also mask or amplify the underlying signal of anthropogenic climate change for decades (Trenberth and Fasullo 2013; Deser et al. 2014; Medhaug et al. 2017; Lehner et al. 2018; Guo et al. 2019). However, current climate models exhibit a wide range of fidelity in reproducing observed modes of variability (MoV, Fasullo et al. 2020, Coburn and Pryor 2021, Lee et al. 2021) and fundamental questions remain regarding the capacity of the observational record to adequately sample climate modes, their intrinsic variability, and responses to forcing (Fasullo et al. 2020, Phillips et al. 2020, Deser and Phillips 2021).

It is known that many biases in model representations of MoV are systematic across many models. For example, simulation of the El Niño / Southern Oscillation (ENSO) often includes sea surface temperature (SST) anomalies that extend too far westward in the equatorial Pacific Ocean (EPO), spectra that are biased in magnitude with periodicity that is overly biennial, and teleconnections that are often too weak (Fasullo et al. 2020). Also, in many models Pacific Decadal Variability (PDV, Newman et al. 2016) fails to extend into the deep tropics with sufficient magnitude, limiting its large-scale effects and connections with coupled tropical modes. For spectra in some modes, such as in ENSO, large intrinsic variability across century-long simulations suggests an implicit limitation in the century-scale observational record for constraining simulations, with only qualitative statements being possible. At higher time frequency the simulation of the Madden Julian Oscillation (MJO) and the Quasi-Biennial Oscillation (QBO) has been particularly elusive. Often being completely absent in many models until their most recent incarnations. Across model generations, a progressive improvement in the simulation of many modes has been identified in conjunction with broader improvements in model fidelity (Fasullo 2020, Fasullo et al. 2020), suggesting a role for improvements in model resolution and physics in advancing simulation of MoV.

Challenges in evaluating simulated MoV remain. Recent work has identified shortcomings in characterizing some modes of variability through traditional techniques in the presence of forced climate change, as the modes and patterns of forced change can be conflated (Deser and Phillips 2021). For other modes, such as those in the extratropics, model biases are comparable in magnitude to observational uncertainties in some cases. Additionally, the intrinsic variability of some modes may limit the usefulness of the observational record in fully capturing their spatiotemporal character (Fasullo et al. 2020), though relatively few studies have addressed the issue. A key resource for addressing these issues and evaluating MoV are climate model large ensembles (LEs), multi-member simulations using a single climate model and specification of external climate forcings designed to allow for the estimation of forced climate changes, via the ensemble mean, and the isolation of climate variability. Four recently-produced LEs are the focus of this work, produced with versions 1 and 2 of the Energy Exascale Earth System Model (E3SM) and versions 1 and 2 of the Community Earth System Model (CESM). To benchmark many modes of variability (see Section 2c for a listing of modes assessed), a newly developed diagnostics package is used, the Climate Variability Diagnostics Package for Large Ensembles (CVDP-LE), developed at the National Center for Atmospheric Research (Phillips et al. 2020). By allowing for estimation and removal of the forced response of key modes, this approach allows for the characterization of model bias, the relative skill exhibited by E3SM and CESM, and the detectible improvements in model fidelity across model generations. For diagnosing modes not assessed in the CVDP-LE, such as the MJO and QBO, conventional methods are used (see Methods). These ensembles also allow for an exploration of the intrinsic noise of key modes and thus the usefulness of the observational record in constraining the modes.

The goal of this work is to build upon existing knowledge with four specific objectives. First, the analysis examines the intrinsic noise of MoV in century-scale records to identify those modes whose simulated bias is large relative to their intrinsic variability and observational uncertainty. The analysis then seeks to evaluate a broad range of MoV and characterize model biases in simulating them. The analysis addresses differences in model fidelity between E3SM and CESM, providing context for which modes are best represented in each model. Lastly the analysis also compares versions 1 and 2 of each model, documenting the changes across successive model versions and assessing prospects for continued model improvements. The investigation begins with a discussion of the models, data, and methods used in Section 2. The main results are presented in Section 3. These include an exploration of the rationale for choosing various modes for assessment and the examination of large-scale coupled modes such as Pacific Decadal Variability, ENSO, and the Atlantic Niño mode. The fidelity of extratropical modes is then addressed, followed by an evaluation of the QBO and MJO. A discussion of the main successes and remaining biases, and paths forward for model improvement, is contained in Section 4.

2. Materials and Methods

a. The Exascale Earth System Model

The characteristics of the ensembles used in this work are summarized in Table 1. The E3SM is the U.S. Department of Energy's (DOE) Earth system model, designed to produce climate projections on scales relevant for energy applications. The E3SM version 1 (E3SM1, Golaz et al. 2019) large ensemble (E3SM1-LE, Stevenson et al. 2023) uses the model's standard resolution of 1º for the atmosphere and land components, 0.5º for the river model, and the Model for Prediction Across Scales (MPAS) variable-resolution ocean/sea ice model that ranges from 60 km in the midlatitudes to 30 km at the equator and poles.. The E3SM1 Atmosphere Model (EAMv1, Rasch et al. 2019 doi: 10.1029/2019ms001629) shares early versions of its codebase with the Community Atmosphere Model version 5 (CAM5) used in CESM, but has been substantially modified in recent years, and therefore differs considerably from CAM5. Changes from CESM stem from the use of higher vertical atmospheric resolution (72 instead of 32 levels), modified aerosol microphysical schemes, use of the spectral element (SE) dynamical core. Cloud and convection processes have been substantially updated (Xie et al. 2018 doi: 10.1029/2018ms001350). The model employs a multi-moment representation of aerosol properties using MAM4 (Wang et al. 2020). The ocean, sea ice, and river model components are entirely different from their counterparts in CESM. At 5.3 K, climate sensitivity in E3SM1 is likely too high (Golaz et al. 2019). Produced on the compy and cori supercomputers at the DOE, the E3SM1-LE consists of 20 members spanning 1850 to 2100 initialized from a diverse range of "macro" climate states that span a broad range of ocean conditions (Stevenson et al., 2023), though only 17 members were available at the time of this study. The E3SM1-LE experiment uses the historical and SSP3-7.0 forcing scenarios specified by CMIP6 (Eyring et al., 2016).

The E3SM version 2 (E3SM2, Golaz et al. 2022) shares the same resolution of E3SM1 and features improvements in many aspects, including better simulated clouds, precipitation, and radiation, and a more realistic climate sensitivity of 4.0 K. Computational performance improvements also increase the model's suitability for the large expense of LE experiments. Produced on a high-performance computing cluster provided by the BER Earth System Modeling Program at the DOE, the E3SM2 Large Ensemble (E3SM2-LE) consists of 21 members from 1850 to 2100 (Fasullo et al., in review), initialized from a broad range of "macro" climate states with forcings that are also based on the historical and SSP3-7.0 forcing scenarios provided by CMIP6. A notable feature of the historical simulations in both the E3SM1-LE and E3SM2-LE is a relative lack of warming in the middle and late 20th century, likely due to excessive aerosol cooling (Golaz et al. 2019, 2021), though the effect on MoV is largely unknown.

b. The Community Earth System Model

The Community Earth System Model versions 1 (CESM1, Hurrell et al. 2013) and 2 (CESM2, Danabasoglu et al. 2020) use nominal 1 º horizonal resolution with 32 vertical levels. The ocean/sea ice models are the Parallel Ocean Program version 2 and CICE Version 5.1.2, respectively. Climate sensitivity in CESM1 is 4.1 K while in CESM2 is 5.3 K, though both accurately reproduce the evolution of global surface temperature over the historical era. The CESM1 large ensemble (Kay et al. 2015; CESM1-LE) uses CAM5 and therefore shares some of the codebase used by E3SM. The CESM1-LE is comprised of 40 members spanning 1920-2100 employing a "micro" initialization approach from a single climate state. The CESM2 represents a significant advance over CESM1, containing many science and infrastructure advances, with major updates to the model's cloud and aerosol schemes, and to the land and sea-ice components (Danabasoglu et al. 2020). The CESM2 Large Ensemble (Rodgers et al. 2021) uses the low-top 32 vertical level version of CESM2 generated on the IBS/ICCP supercomputer "Aleph" in Busan, South Korea spanning 1850 to 2100 with 100 members. During production of CESM2, spurious effects of prescribed biomass burning emissions were identified (Fasullo et al. 2022) leading to the production of a smoothed biomass ensemble of 50 members, which will be used in this work.

Model	**Period**	**# Mem**	**Scenarios**	**Resolution**
E3SM1	1850-2100	17	CMIP6 Historical / SSP370	1º atm/lnd, ¼-½º ocn
E3SM2	1850-2100	21	CMIP6 Historical / SSP370	1º atm/lnd, ¼-½º ocn
CESM1	1920-2100	40	CMIP5 Historical / RCP85	1º atm/lnd, 1º ocn
CESM2	1850-2100	50	CMIP6 Historical / SSP370	1º atm/lnd, 1º ocn

Table 1. Characteristics of the large ensembles used in this study, including the "future" scenario used to extend this analysis through 2020.

c. The Climate Variability Diagnostics Package for Large Ensembles

The CVDP-LE is the core diagnostics package used for this analysis (Phillips et al. 2020) and the period considered is from 1920 to 2020. Developed by NCAR's Climate Analysis Section, the CVDP-LE is an automated analysis tool for assessing internal and forced contributions to climate in coupled model LEs and observations. The package computes metrics for a wide range of modes of interannual-to-multidecadal variability in the atmosphere, ocean, and cryosphere, as well as long-term trends and key indices of global and regional climate. To avoid conflation of the patterns of forced change with those of MoV in assessing PDV and AMV, the package estimates and removes the ensemble-mean (i.e. forced response). Multiple best-estimate observations are used to estimate observational uncertainty, though the limited duration of the observational record also constitutes a sampling uncertainty that models can be used to estimate. All CVDP-LE diagnostics and metrics are saved to a data repository (see Data Availability) analysis and posted online. The major coupled modes chosen for consideration include the PDV (PDV' in Phillips et al. 2020), ENSO (Deser et al. 2010), and the Atlantic Niño mode (Zebiak 1993). In the extratropics, the Southern Annular Mode (SAM, Thompson et al. 2000), the North Atlantic Oscillation (NAO, Hurrell and Deser 2010), the Northern Annular Mode (NAM, Thompson et al. 2000), and Pacific / North America (PNA) modes are assessed. The methods for computing these modes are described in Phillips et al. (2020) and the SAM is only assessed from 1950 due to data limitations discussed below. On sub-monthly timescales the Madden Julian Oscillation (MJO, Zhang, 2005) and the stratospheric Quasi-Biennial Oscillation (QBO, Dunkerton and Delisi, 1985) are also assessed in this work, though not computed in the CVDP-LE. Methods used for these additional modes are discussed in Section 3.

d. Observational Datasets

Multiple observational products are used in this analysis to diagnose sea surface temperature (SST), sea level pressure (SLP), and near-surface air temperature (T_{2m}). Furthermore, to account for observational uncertainty, multiple datasets are used for each variable. These include SST estimates from the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST; Rayner et al 2003) dataset and the Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5, Huang et al. 2017) dataset. For T_{2m}, the Berkeley Earth Surface Temperature (BEST; Rohde et al. 2013) and Goddard Institute for Space Studies Surface Temperature (GISTEMP; Lenssen et al. 2019). SLP data from the European Center for Medium-Range Weather Forecasts (ECMWF) Twentieth Century Reanalysis (ERA20C; Poli et al. 2016) and twentieth century climate reanalysis (CERES-20C; Laloyaux et al. 2018) are used to extend into the early to mid 20th century. These data are concatenated with ECMWF's fifth generation reanalysis (ERA5, Hersbach *et al.* 2020) to extend to the present day. Mode metrics are computed from 1920 to 2020 and data sampling during this period is generally sufficient, except in the polar southern hemisphere and thus analysis of the Southern Annular Mode is limited to 1950-2020 for which the surface-based reanalyses are more consistent with the observed SAM and considered superior to the full reanalyses (Gerber and Martineau, 2018). For computing pattern correlations, both global and regional fields are used depending on the mode in question. For PDV and ENSO teleconnections, global fields are used whereas for extratropical modes, such as the NAO and SAM, regional domains are used that span the areas used for mode computation (Phillips et al. 2020). For the MJO and blocking analysis, daily data are used from NOAA-OLR, ECMWF's interim reanalysis (ERA-I, Dee et al. 2011), and NASA's MERRA2 reanalysis (Ronald Gelaro, et al., 2017). For the QBO we use zonal mean wind from ERA5.

3. Results

a. Intrinsic Mode Variability and Uncertainty

The distributions across LE members of pattern correlations of simulated MoV with observations for various modes are shown in Figure 1. Also shown is the observational uncertainty of each mode as estimated from the pattern correlations between benchmark observational datasets (red dots). Two necessary conditions for robust model evaluation are 1) that the observational uncertainty in the patterns be much less than estimated model biases and 2) that the patterns identified in simulations of comparable length to observations do not experience significant inter-member variation relative to model bias, as this would suggest excessive intrinsic mode noise and high observational sampling error. While model estimates of mode noise are used here, biases in those estimates may exist. Various key takeaways are evident from Fig. 1. First, the El Niño and La Niña T_{2m} composites contain significant uncertainty arising from intrinsic noise while the hovmoëllers exhibit high certainty. Also, the PDV can be evaluated much more robustly than the IPO, whose mean correlation is small, and ensemble spread and observational uncertainty are large. For extratropical modes (e.g. NAO, SAM, NAM, PNA), observational uncertainty is small and model fidelity is generally very high with mean correlations above 0.95 and model spread nearly encompassing the observations in many instances. Perhaps most importantly, there is generally agreement on the intrinsic noise in MoV across models, suggesting a potentially robust finding and confirming the generality of similar findings using CESM1 in Fasullo et al. (2020). The sensitivity of results to alternative definitions of extratropical modes is also expected to be small as the modes and seasons assessed are known to be robust and well-distinguishable from higher order EOF signals in general (Lee et al. 2019).

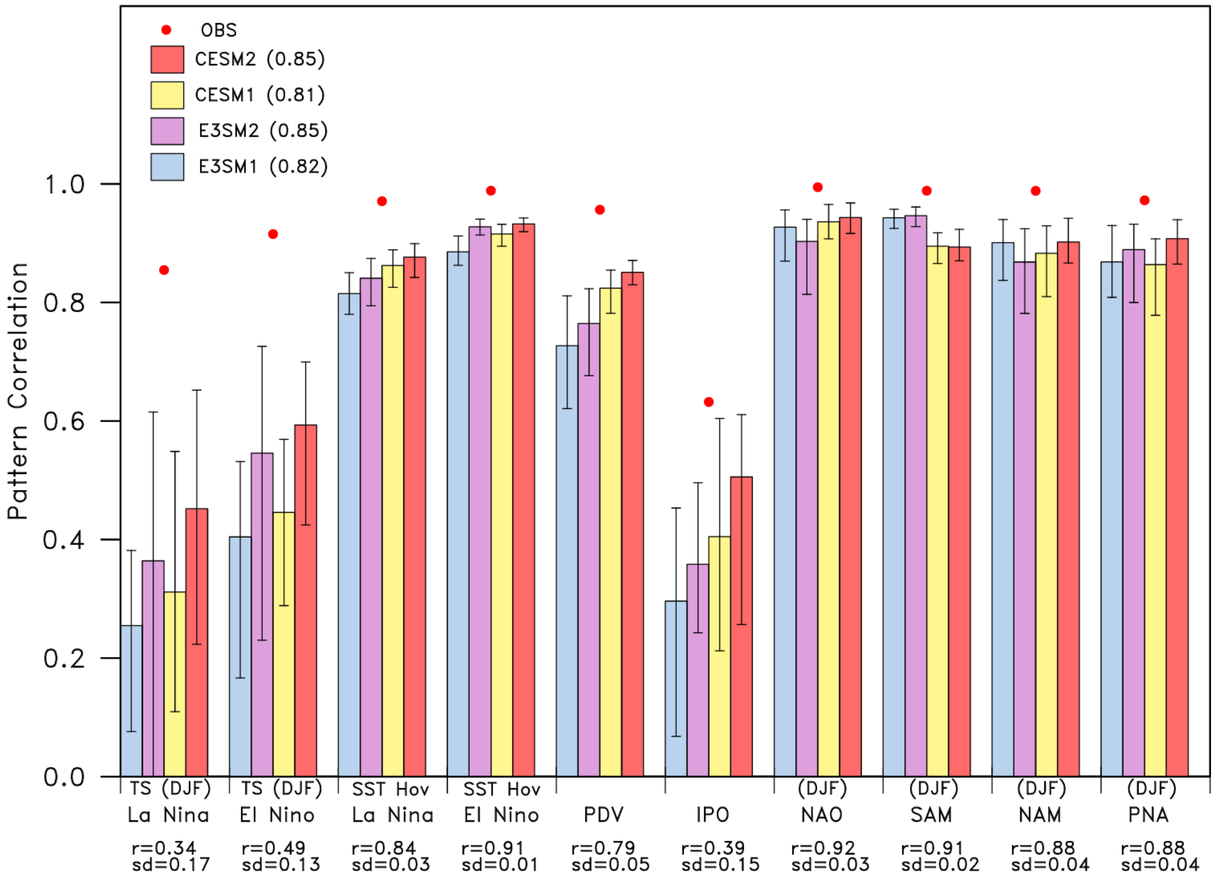

Fig. 1. Estimated mode fidelity in E3SM and CESM versions and their robustness in observations. Bars show the mean pattern correlation of various simulated modes of variability for each ensemble against observed estimates (bars) and the spread in this metric across the ensemble (whiskers). Red dots show the pattern correlation between the two observational estimates. Mean correlation magnitudes and standard deviations averaged across the ensembles are computed for each mode and specified along the abscissa.

b. Pacific Decadal Variability

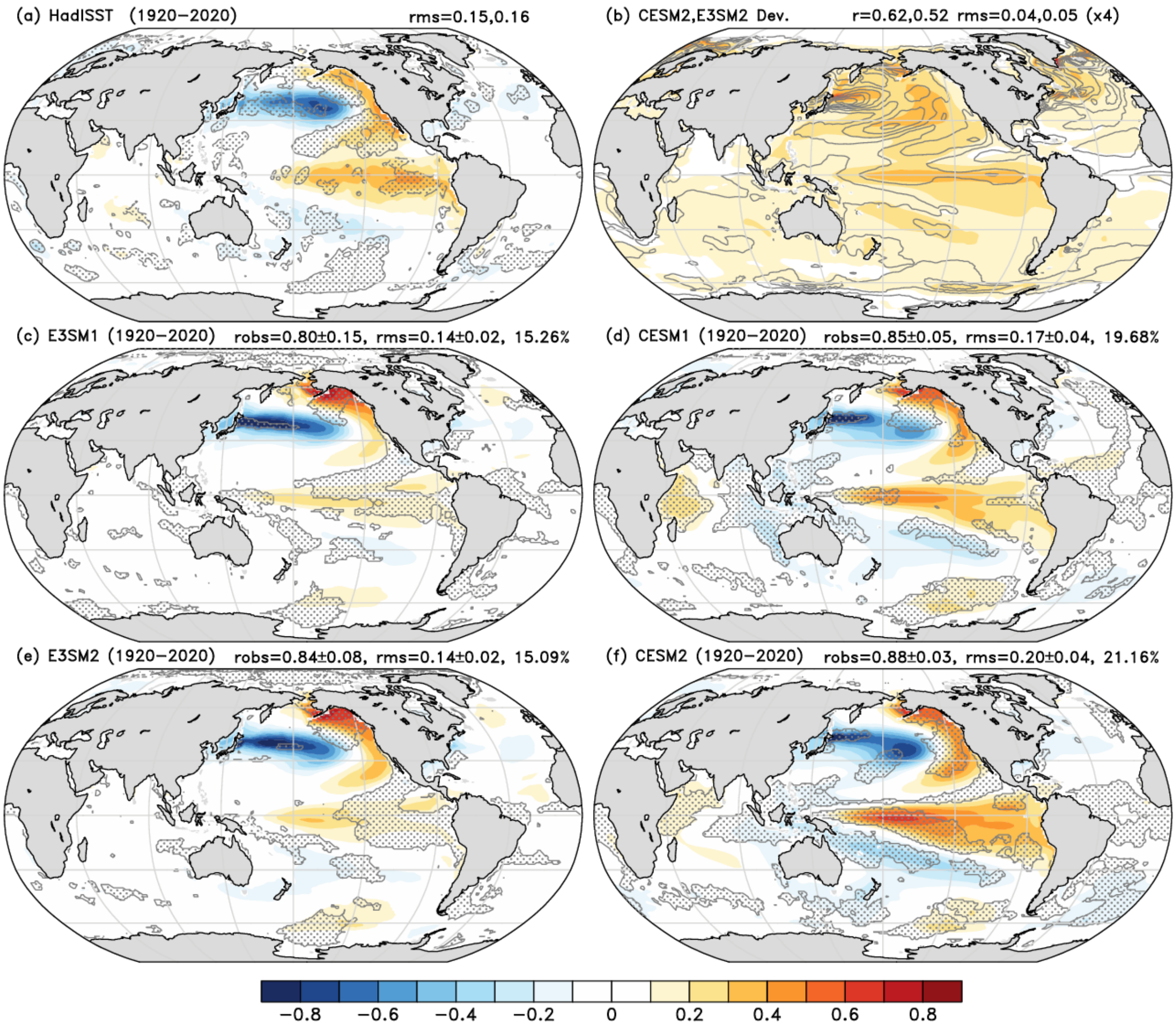

Fig. 2. The observed PDV pattern (K) estimated from SST observations (HadISST, a) and the standard deviation across patterns from members of CESM2 (contours, E3SM2 lines) and E3SM2 (lines, b). The ensemble-mean PDV patterns estimated from E3SM1 (c), CESM1 (d), E3SM2 (e), and CESM2 (f) are also shown. Stippling in (a) where differences with ERSSTv5 exceed 0.05 K and in (c-f) where observations lie outside of twice the standard deviation range of the ensemble. The pattern correlation with observations and root-mean-squared pattern intensity with two standard error ranges are also indicated along with the percent of the globe stippled.

Figure 2 shows the PDV pattern (Newman et al. 2016) estimated from HadISST (Fig. 2a), also often referred to as the Pacific Decadal Oscillation, the intra-ensemble standard deviation in the patterns from CESM2 (filled contours) and E3SM2 (contour lines, Fig. 2b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 2c-f). In observations, the PDV is characterized by a strong dipole in the North Pacific Ocean

(NPO), with the eastern flank of the dipole extending into the subtropics. Values on par with those along the coastline of North America (NA) also exist in the deep tropics, extending from the dateline to South America (SA). Weak values of the same sign as the western NPO are evident in the southwestern subtropical Pacific Ocean (SSPO). The standard deviation in the pattern across LE members shows the greatest values in the western NPO, with a local maximum in the central and eastern EPO. In E3SM1, the mean pattern correlates well with observations (r=0.8), with a root-mean-squared (RMS) magnitude of 0.14 K. The intensity of the pattern in the NPO is greater than in observations, but with values in the subtropics and deep tropics that are too weak generally, with the later centered near the dateline. The pattern is also too weak in the SSPO. In CESM1, the pattern more closely resembles the observed pattern (r=0.85), with values in the subtropics and tropics that agree favorably with observations but that also peak near the dateline rather than in the eastern EPO. The pattern in E3SM2 closely resembles that in E3SM1, with similar biases (r=0.84). In CESM2, the pattern again agrees well with observations (r=0.88) but with a magnitude (RMS=0.20) that exceeds both that in CESM1 (RMS=0.17) and that observed (RMS=0.15). As in CESM1, the tropical peak is also too strong and is located near the dateline rather than in the east. The magnitude of the pattern in the SSPO is also somewhat too strong.

c. El Niño – Southern Oscillation

Figure 3 shows El Niño teleconnection patterns with T_{2m} anomalies from December through February (DJF) estimated from composites of events in nature using HadISST (Fig. 3a), the intra-ensemble standard deviation in the analogous patterns for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 3b), and the ensemble mean composite patterns from versions 1 and 2 of both E3SM and CESM (Fig. 3c-f). The fields are based on compositing events using a 1 standard deviation threshold of the linearly detrended December Niño3.4 SST Index anomalies. In observations, the pattern is characterized by strong warming centered near 130° W that extends about 5° from the equator and is small west of the dateline. Warm anomalies also exist in many land regions, and these are particularly strong in northern NA, while some land regions (e.g., southern NA and northern Eurasia) are anomalously cool. Variability across ensemble members is particularly large over boreal land regions (Fig. 3b). In E3SM1, the T_{2m} anomalies are similar to those observed, however they extend too far westward and are also too strong in the eastern EPO. Remote anomalies, and particularly those in northern NA, are weaker than those observed, consistent with a general weakness of ENSO teleconnections in models when normalized by Niño3.4 SST variability

(Fasullo et al. 2020). The weakness in Fig. 3c-f is particularly notable given the excessive strength of SST anomalies in the EPO. In all ensembles, the ensemble mean biases are similar to those in E3SM1, with anomalies in the EPO that extend too far westward and remote anomalies that are generally too weak. An exception is the E3SM2 anomaly pattern, which exhibits extratropical teleconnections that are stronger generally than the other LEs over boreal land.

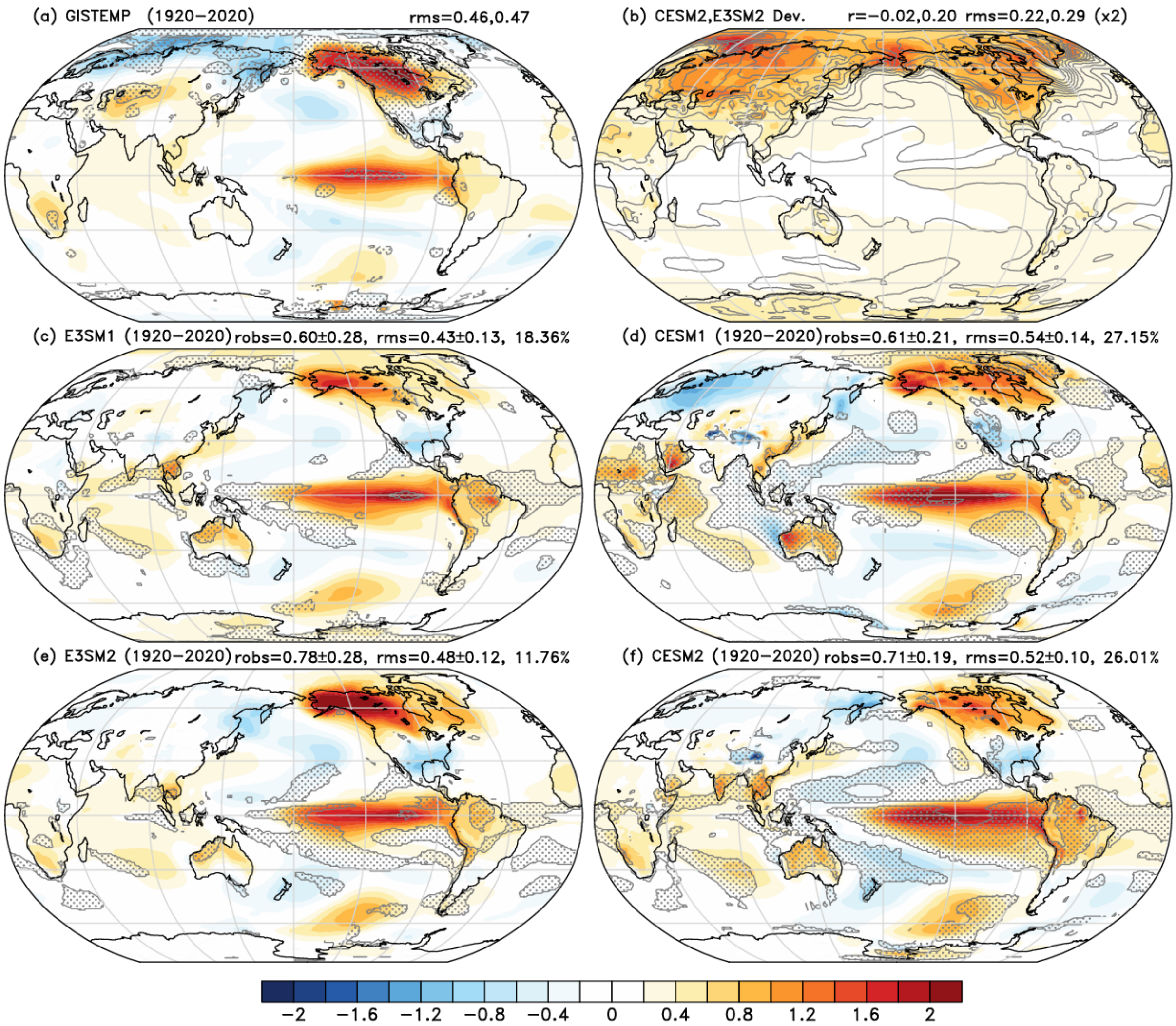

Fig. 3. The observed El Niño teleconnection pattern in T_{2m} (K) in DJF estimated from observations (GISTEMP, **a**) and the standard deviation (σ) of patterns across members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Stippling in (**a**) where differences with GISTEMP exceed 0.2 K and in (**c-f**) where observations lie outside of the two σ range of each ensemble. Values in the headers include (a) the RMS magnitude of the patterns in GISTEMP and BEST, (b) the pattern correlations with the respective simulated

patterns and the RMS magnitudes, (c-f) the pattern correlation of ensemble members with the GISTEMP pattern and the 2 σ range, the mean RMS magnitude and 2 σ range, and the percentage of global area in which the observations lie outside of the 2 σ range.

The La Niña composite patterns, generated in an analogous fashion to the El Niño composite but for cool anomalies exceeding the standard deviation threshold, are shown in Figure 4 for DJF T_{2m} anomalies estimated from composites of events in nature using HadISST (Fig. 4a), the intra-ensemble standard deviation in the analogous patterns for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 4b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 4c-f). In observations the pattern is characterized by a zonally broad band of cool anomalies centered near 200 ^{o} E that extends further westward than warm anomalies during El Niño and about 5^{o} from the equator. Cool anomalies also exist in northern NA, and many subtropical land regions. As for El Niño teleconnections, intra-ensemble variability is particularly large over boreal land regions. In E3SM1, the anomalies are similar to those observed however they extend slightly too far west and are also too strong in the eastern EPO. Remote anomalies, and particularly those in NA are similar in strength and structure to those observed. Composite anomalies in all LEs share aspects of the E3SM1 biases, including the excessive westward extension of anomalies in the EPO. Teleconnections in E3SM2 and CESM2 strengthen relative to E3SM1 and CESM1 however anomalies in NA however observed teleconnections in most regions fall within the ensemble range. Pattern correlations with observations increase between both E3SM1 (r=0.50) and E3SM2 (r=0.61), and CESM1 (r=0.45) and CESM2 (r=0.64).

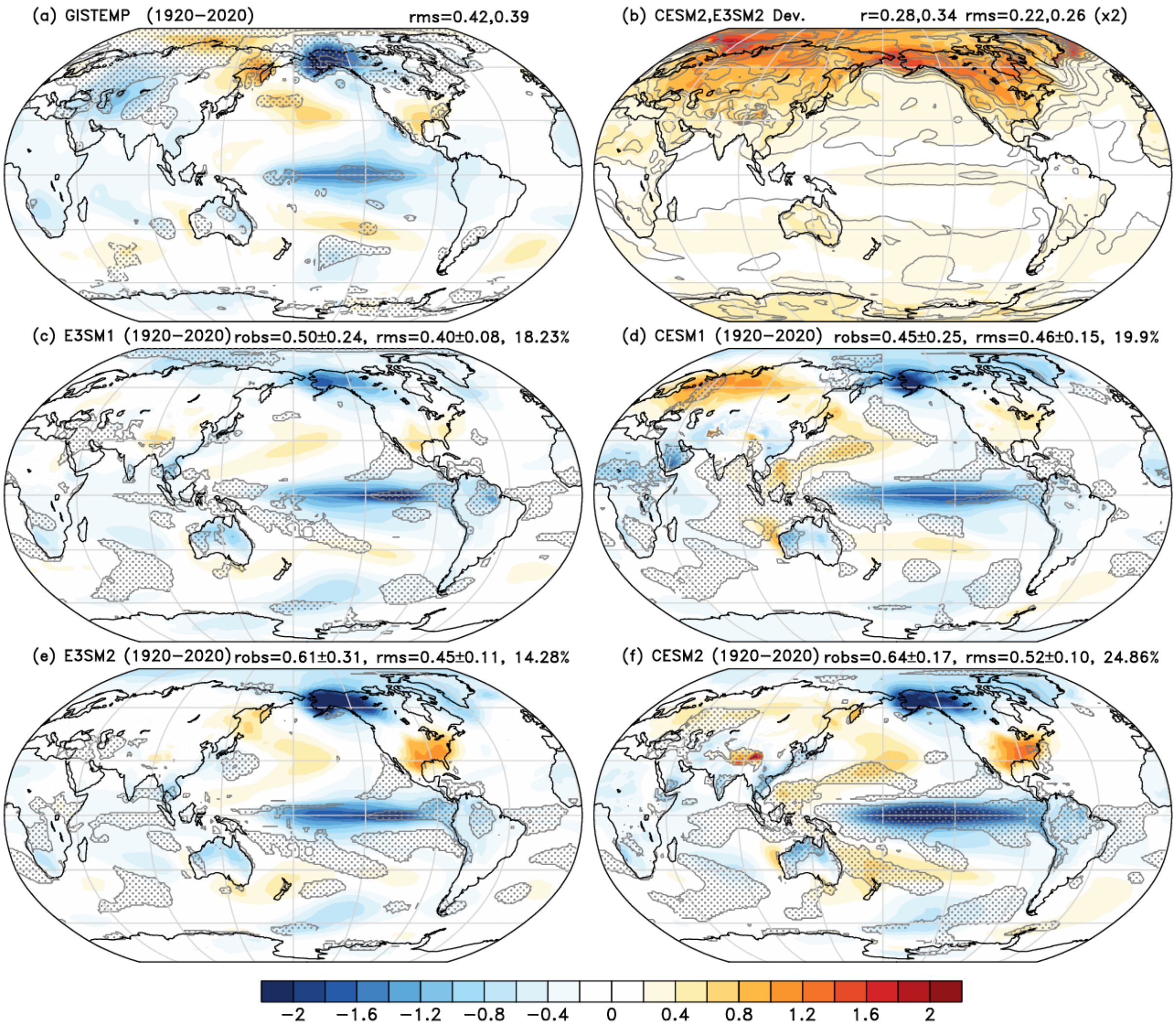

Fig. 4. The observed La Niña teleconnection pattern in T_{2m} (K) in DJF estimated from observations (GISTEMP, **a**) and the standard deviation in patterns across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Stippling in (**a**) where differences with GISTEMP exceed 0.2K and in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

Figure 5 shows El Niño composite patterns analogous to Fig. 3 but with DJF precipitation (P) estimated from GPCP (Fig. 5a), the intra-ensemble standard deviation in the analogous patterns for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 5b), and the ensemble mean patterns for versions 1 and 2 of both E3SM and CESM (Fig. 5c-f). In observations the pattern is characterized by a zonally broad band of positive P anomalies across the eastern EPO centered near 200°E that extends about 5°N from the equator and into the southern subtropics near 220°E. The anomalies are surrounded to the north, west, and south by negative P anomalies. Positive anomalies exist off the west coast of NA, southeastern SA, and southern NA, where cool anomalies are associated with high P (Fig. 3a). Negative precipitation anomalies span South Africa, much of Australia, and northern SA and thus these regions are relatively warm and dry during El Niño (Fig. 3a). Ensemble spread is particularly large in the tropics and south of the equator where the major precipitation zones migrate during DJF. In all LEs the broad pattern of anomalies is captured but significant biases exist. These include anomalies that are too narrowly confined to the EPO and extend too far westward, and extratropical anomalies that are too weak. Large biases also exist in the Indian Ocean, particularly in CESM1/2. While pattern correlations are generally strong ($r>0.64$), they are strongly influenced by anomalies in the tropics that dominate the overall pattern variance. The overall strength of the patterns increases between versions 1 and 2 of both E3SM ($r=0.70$ to $r=0.81$) and CESM ($r=0.64$ to $r=0.81$), in improved agreement with GPCP, however in CESM2 the pattern strength is excessive (RMS=1.31), likely in association with the excessive strength of SST anomalies (Fig. 3).

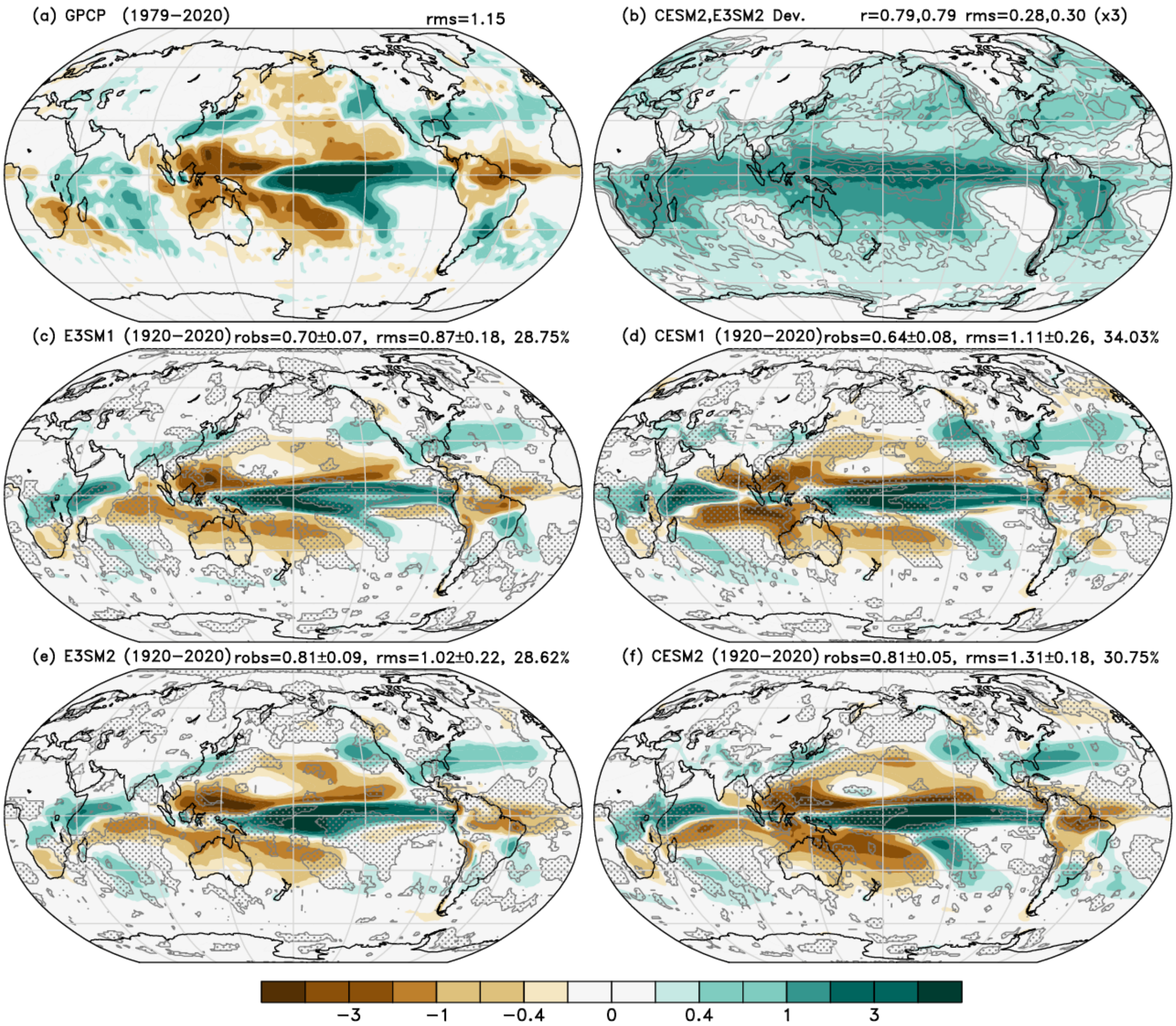

Fig. 5. The observed El Niño teleconnection pattern in PR (mm day^{-1}) in DJF estimated from observations (GPCP, **a**) and the standard deviation in patterns across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

Figure 6 shows La Niña composite patterns with precipitation (P) estimated from composites of events in nature using GPCP (Fig. 6a), the intra-ensemble standard deviation in the analogous patterns for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 6b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 6c-f). Observed precipitation teleconnections are characterized by a zonally broad band of strongly negative precipitation anomalies in the central and eastern EPO centered just west of the dateline that extends about 5° from the equator and into the southern subtropics near 220°E. It is surrounded to the north, west, and south by positive anomalies that extend into the midlatitudes. Negative precipitation anomalies exist off the west coast of NA, and over southeastern SA, and southern NA, where warm anomalies are associated with reduced precipitation (Fig. 4a). Positive precipitation anomalies span South Africa, much of Australia, and northern SA and thus these regions are also relatively cool during La Niña (Fig. 3a). Intra-ensemble variability in precipitation is distributed more evenly about the equator than during El Niño and is large over southern Africa, Australia, and SA. In all LEs, the anomalies are too narrowly confined to the EPO and extend too far to the west. Simulated pattern correlations increase between versions 1 and 2 of both E3SM (r=0.68 to r=0.73) and CESM (r=0.54 to r=0.68), as does the strength of simulated patterns. As for El Niño, the pattern strength in CESM2 is excessive.

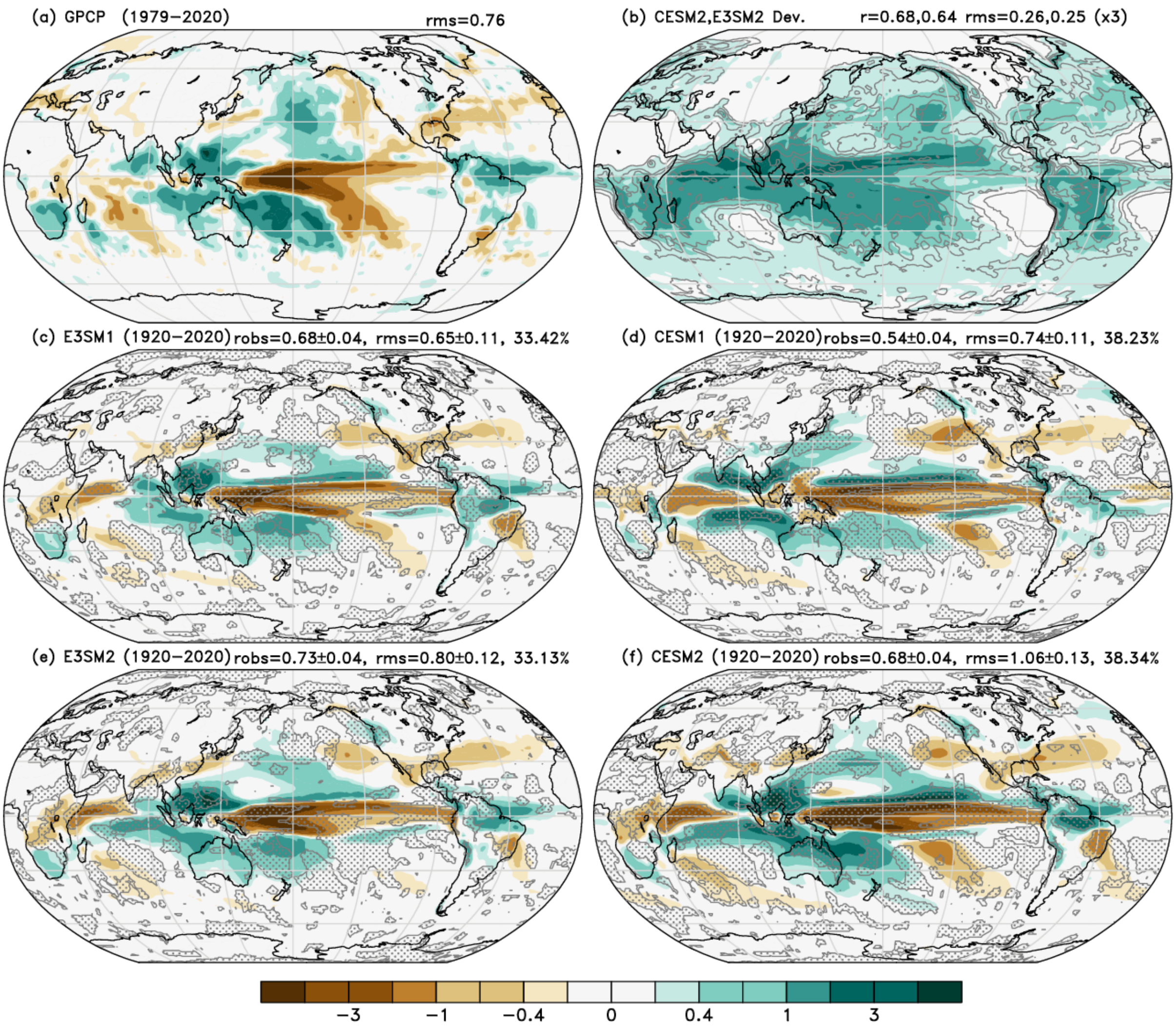

Fig. 6. The observed La Niña teleconnection pattern in PR (mm day^{-1}) in DJF estimated from observations (GPCP, **a**) and the standard deviation in the pattern across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

To diagnose ENSO's space-time structure, hovmoëllers are shown in Figure 7 for El Niño's composite anomalies in observed near surface air temperature (Fig. 7a), the intra-ensemble standard deviation in the analogous composites for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 7b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 7c-f). In observations, the pattern is characterized by strong warming east of the dateline extending to 80°W that emerges in June through September (JJAS), peaks in December, and recedes from March through May (MAM). Often El Niño events transition to La Niña conditions, leading to slightly cool anomalies in the composite in DJF of the following year. Significant variability exists from event to event, however, such that large spread across the ensembles also exists at that time (Fig. 7b). In the LEs, a similar overall pattern exists but with notable differences, such as the significantly greater magnitude and broader westward extent of anomalies, as also evident in Figs. 3 and 4. Other features are also evident in Fig. 7, such as the strong zonal contrast of anomalies between the central and eastern EPO in versions 1 and 2 of E3SM. In these models, there is also rapid anomaly growth near 115°W that is not observed. Biases in the temporal structure are also evident, including excessive cool anomalies in E3SM1, E3SM2, and CESM2 in the year following El Niño events, suggesting excessive biennial power in ENSO.

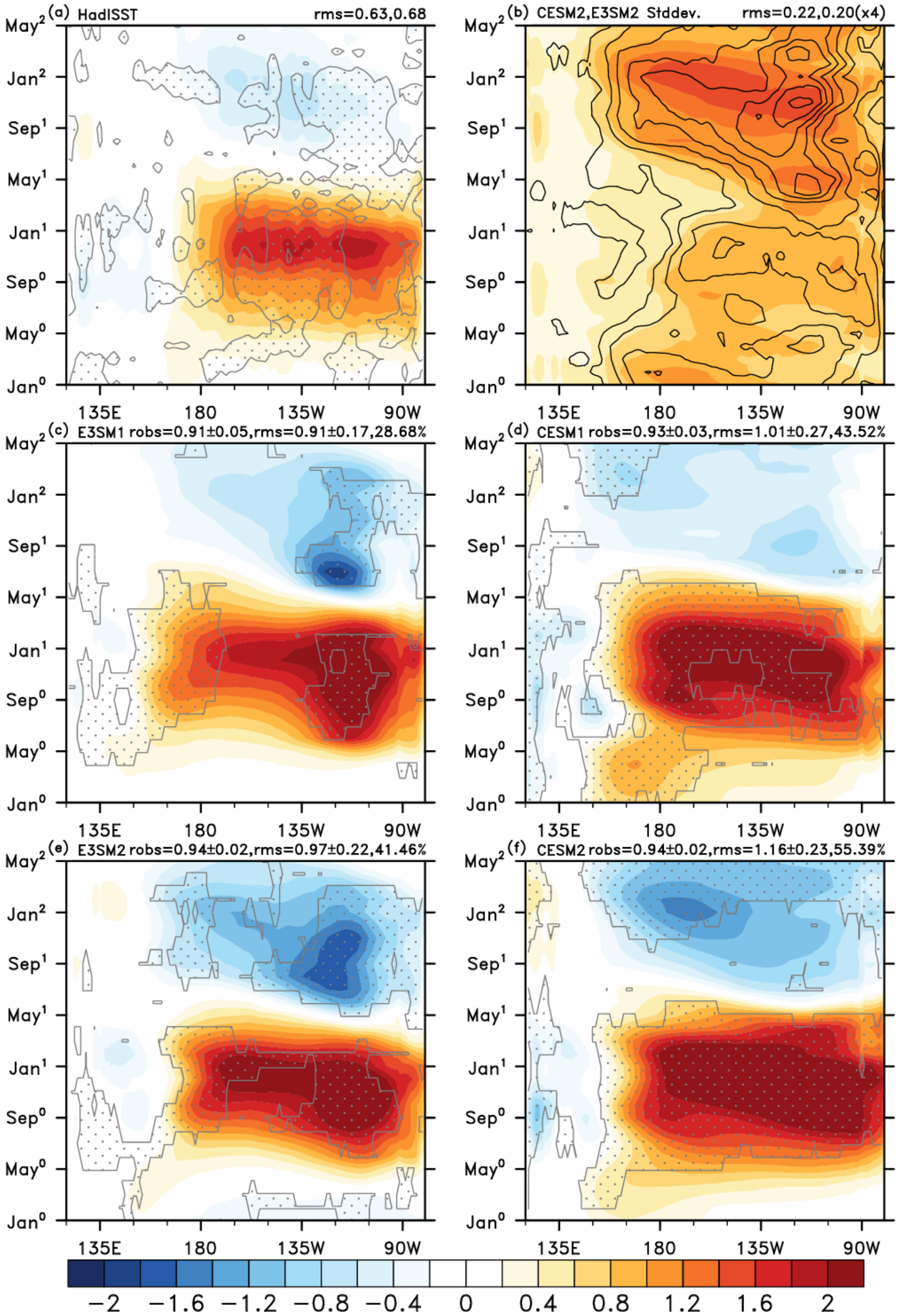

Fig. 7. The observed El Niño hovmoëller pattern in SST (K) estimated from observations (HadISST, **a**) and the standard deviation in the pattern across members of CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with ERSSTv5 exceed 0.1 K and in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

Figure 8 shows La Niña's time-space structure in T_{2m} anomalies estimated from composites of events in nature using HadISST (Fig. 8a), the intra-ensemble standard deviation in the analogous structures for CESM2 (filled contours) and E3SM2 (contour lines,

Fig. 8b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 8c-f). In observations, the pattern is characterized by cooling that extends from east of the dateline (~165ºE) to 110ºW that emerges in boreal summer, peaks from December to January (depending on longitude) and recedes in March through May. Unlike El Niño, La Niña exhibits an evolution that suggests a westward propagation of anomalies. While La Niña events often transition from El Niño events, their multi-year character can also limit the expression of El Niño in the surrounding years of composite events, as evident by neutral anomalies in the January prior to La Niña and cool conditions in the following January (Fig. 8a). As a result, significant inter-member spread is also evident at the onset and termination of events (Fig. 8b). While the general character of composite anomalies in the LEs is similar to that observed, including their westward propagation, significant biases exist in both the magnitude and temporal evolution of events. Events in E3SM1 and E3SM2 are generally too biennial, as evident by the strong warm anomalies in the DJF periods prior to and following events in the composites. As for El Niño, a strong zonal gradient in variability is also evident in E3SM1 and E3SM2 that contrasts with the observed structure. A similar but weaker bias exists in CESM1 and CESM2 with event magnitude that is excessive, particularly in CESM2. Thus, while the pattern correlation of the composite improves between versions 1 and 2 of both E3SM (r=0.84 to r=0.85) and CESM (r=0.88 to r=0.89), the strength of the patterns shown in the panel headers generally worsens.

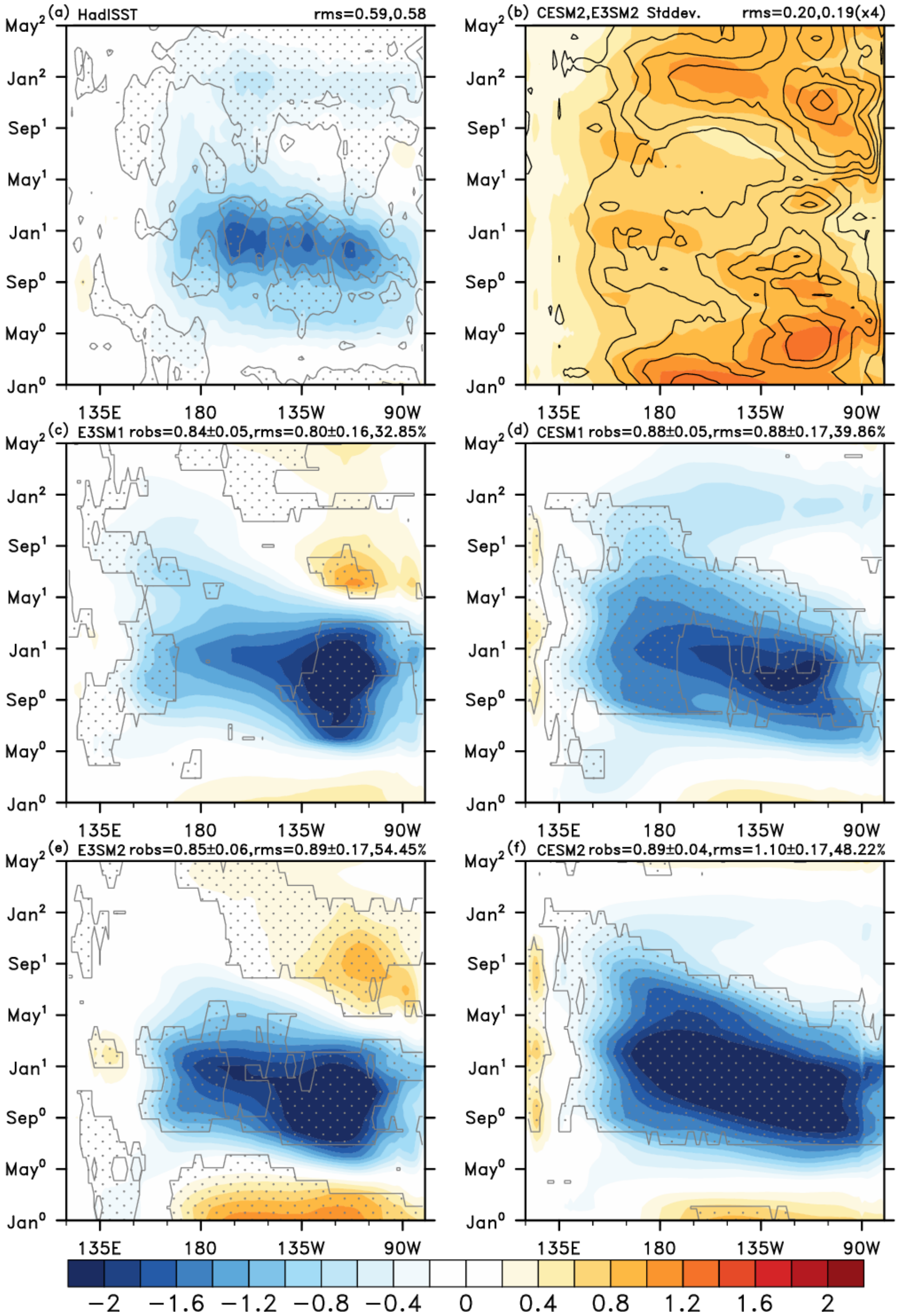

Fig. 8. The observed La Niña hovmoëller pattern in SST (K) estimated from observations (HadISST, **a**) and the standard deviation in the pattern across members of CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with ERSSTv5 exceed 0.1 K and in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

d. Spectra of ENSO and PDV

Figure 9 shows the distribution of bandpass power for ENSO and the PDV across four frequency bands: high (<3 years), interannual (4-6 years), decadal (6-10 years), and multidecadal (> 10 years), assessing ENSO power from SST in the Niño3, Niño4, and Niño3.4 regions. Interannual power in Niño3.4 and Niño4 is generally consistent with observations in E3SM1 and E3SM2 but is stronger than observed in CESM1 and CESM2. In the Niño3 region, power is generally too large in all ensembles except in E3SM2. In the high, decadal, and multidecadal bands model-observation agreement is improved relative to Niño3.4, with the main bias being the excessive power in CESM, and particularly in CESM2. Bandpass power of the PDV index (Fig. 9d) is characterized by greatest magnitudes in the multidecadal band. The large intra-ensemble spread exhibited in all LEs leads to observed power falling within the ensemble range in all cases. Simulated power in the multidecadal band is slightly less for CESM1, while in the decadal band CESM1 exhibits the greatest ensemble mean power. Differences between the models are small however in the context of the large intra-member spread, suggesting a fundamental limit in constraining PDV spectra with the observational record.

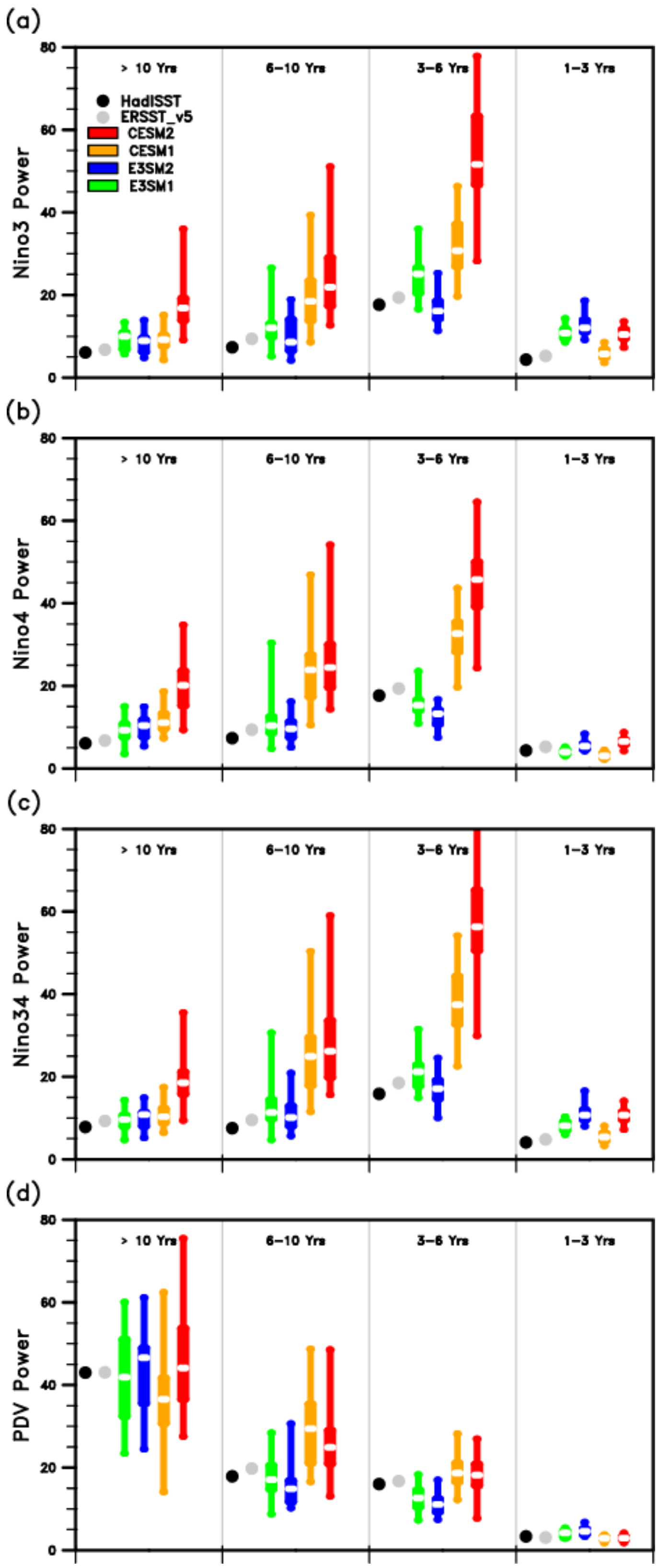

Fig. 9. Spectral power (K^2) in various bands and regions for the (**a**) Niño3, (**b**) Niño4, and (**c**) Niño3.4 indices, and for (**d**) the PDV index for observations (HadISST-black dot, ERSSTv5-grey dot) and E3SM1 (green), E3SM2 (blue), CESM1 (orange), and CESM2 (red).

e. Atlantic Niño

The Atlantic Niño is a dominant mode of interannual variability in the equatorial Atlantic, with many similarities to El Niño in the Pacific (Zebiak, 1993; Xie and Carton, 2004; Keenlyside and Latif, 2007; Ding et al., 2012). The Atlantic Niño index was computed as the averaged SST anomaly over the equatorial Atlantic Ocean (3°N–3°S, 0–20W). The positive phase of Atlantic Niño is characterized by anomalously warm sea surface temperatures in the eastern equatorial Atlantic, relaxed trade winds, and a weakened east-west SST gradient. Similar to ENSO, the interannual variability of the Atlantic Niño is mainly the result of a set of coupled ocean-atmosphere interactions, known collectively as the Bjerknes feedback. The Atlantic Niño has a significant impact on the local climate in the tropical Atlantic sector and can affect remote regions through teleconnections.

Figure 10 shows the composite Atlantic Niño patterns averaged in June to August (JJA) in HadISST (Fig. 10a), the intra-ensemble standard deviation in the patterns from CESM2 (filled contours) and E3SM2 (contour lines, Fig. 10b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 10c-f). In the observations, the anomalous warming is centered around the equator within 5˚S to 5˚N in the central-eastern basin and extends southward along the west coast of Africa to around 15˚S. The strongest SST warming occurs in the eastern equatorial Atlantic between 15˚W-0˚ and along the Africa coast between 10˚S-15˚S.

Compared to the observations, the ensemble mean surface warming patterns in the central-eastern equatorial Atlantic are narrower in their meridional extent in all four models. The strength of the pattern is generally comparable among the models and the observations. In E3SM1 and E3SM2, the ensemble-mean warming patterns exhibit a southward shift of about 2˚ towards the Southern Hemisphere, with an anomalous cooling signal north of the equator, creating a meridional dipole pattern that is absent from observations. There is also spurious cooling in the western equatorial Atlantic near Cape Branco and over the larger area in the North Atlantic. The pattern correlation with the observations is 0.48 for E3SM1 and 0.43 for E3SM2. The ensemble mean warming is stronger than observed in E3SM1, while it is generally consistent with the observed strength in E3SM2.

The CESM1 ensemble mean captures the observed symmetric warming pattern near the equator, although there is an overestimation of warming on its east side near western Africa. Similar to E3SM1, a spurious cooling is present in the western equatorial Atlantic and the

North Atlantic subtropics. The ensemble mean of CESM2 exhibits a weaker and more dispersed warming pattern compared to the other models, though the spatial pattern of warming is generally consistent with observations, and it is the only model that does not show spurious cooling in the subtropical North Atlantic. Unlink observations and the other models, it does not however show significant South Atlantic coastal warm anomalies. The spatial pattern correlations with the observations are higher in CESM1 (r=0.58) and CESM2 (r=0.55) than in either E3SM version.

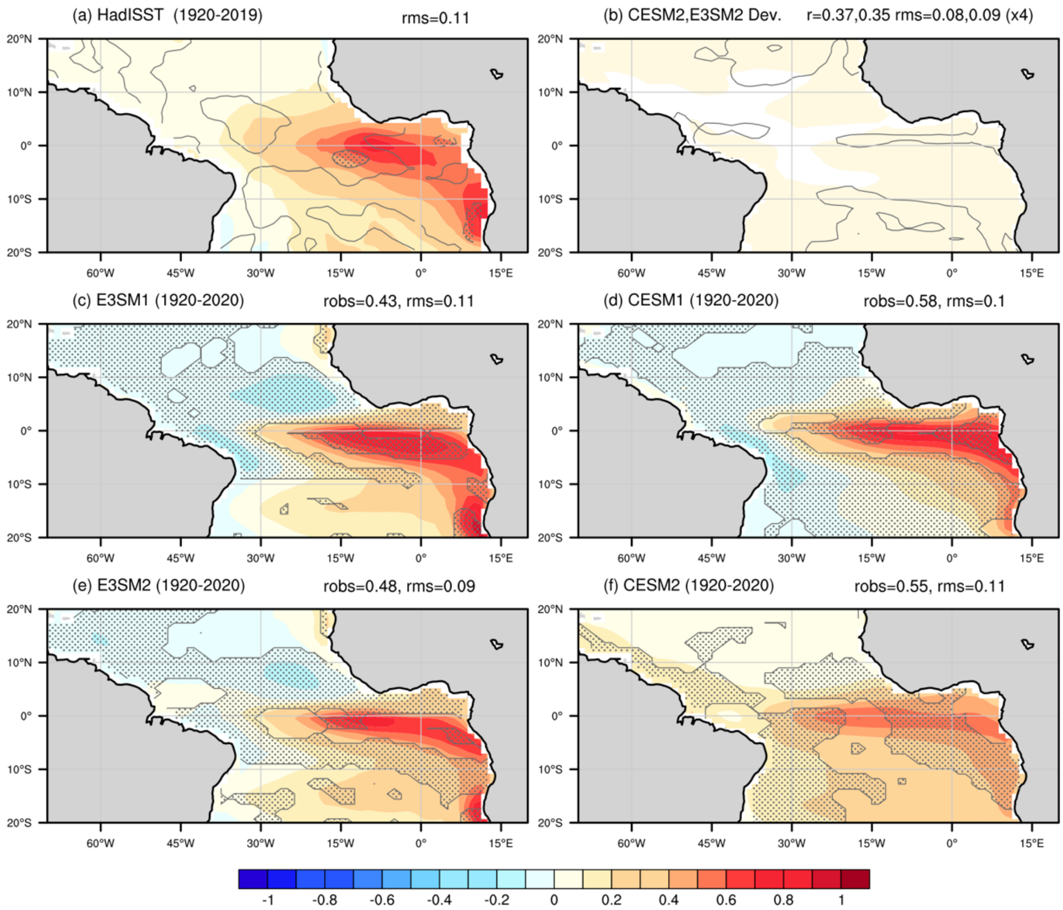

Fig. 10. The observed Atlantic Niño pattern in SST estimated from observations (HadISST, **a**) and the standard deviation in patterns across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Stippling in (**a**) where differences with ERSSTv5 exceed 0.1 ˚C and in (**c-f**) where observations lie outside of each ensemble. Pattern correlations and pattern magnitudes are indicated in the panel captions.

Figure 11 shows Atlantic Niño teleconnection patterns with JJA precipitation estimated from composites of events in nature using GPCP (Fig. 11a), the intra-ensemble standard deviation in the analogous patterns for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 11b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM

(Figs. 11c-f). In the observations, the precipitation anomaly associated with Atlantic Niño is characterized by enhanced rainfall across the equatorial Atlantic basin, located slightly to the north of the equator at ~5˚N.

The simulated precipitation patterns are closely related to their respective patterns of surface temperature. Compared to observations, E3SM1 and E3SM2 exhibit significant differences in their ensemble mean precipitation patterns, with pattern correlations against observations of -0.06 for E3SM1 and -0.01 for E3SM2. Both models display a noticeable southeastward bias in the rainfall maxima towards the eastern equatorial Atlantic. Moreover, the magnitude of the ensemble mean precipitation is too strong at the equator, while a band of negative precipitation anomalies exists across the basin to the north of the equator, in contrast to the observed precipitation maximum in that region. The magnitude of the pattern is also weaker (rms = 0.19 for E3SM1, rms = 0.17 for E3SM2) than observed (rms = 0.27). CESM1 has the highest pattern correlation with observations (r = 0.59), though areas of negative precipitation anomalies are also present at around 10˚N and to the west of the basin. The ensemble mean pattern of precipitation in CESM2 is confined in its meridional extent, and it has a pattern correlation of 0.48 with observations. Additionally, its magnitude (rms = 0.21) is the closest of the LEs to the observations.

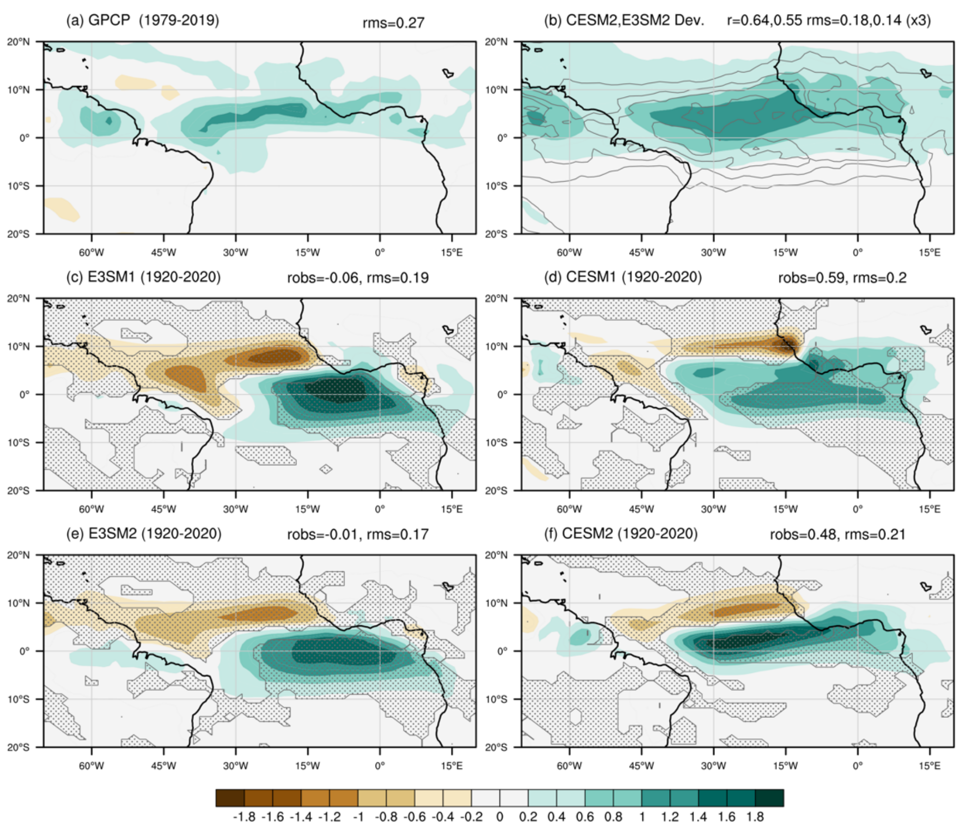

Fig. 11. The observed Atlantic Niño pattern in PR estimated from observations (GPCP, **a**) and the standard deviation in patterns across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Stippling in (**c-f**) where observations lie outside of each ensemble. Pattern correlations and pattern magnitudes are indicated in the panel captions.

To examine the seasonal timing of the Atlantic Niño, Figure 12 shows the space-time hovmoëllers for Atlantic Niño's composite anomalies in surface temperature in HadISST (Fig. 12a), the intra-ensemble standard deviation in the analogous composites for CESM2 (filled contours) and E3SM2 (contour lines, Fig. 12b), and the ensemble mean patterns from versions 1 and 2 of both E3SM and CESM (Fig. 12c-f). The observed Atlantic Niño peaks in the boreal summer around JJA with warming spreading across the central and eastern equatorial Atlantic basin.

In both E3SM1 and E3SM2, the timing of the Atlantic Niño peak season lags the observations by about a month, with the seasonal peak occurring in August. The ensemble mean pattern of E3SM1 is characterized by a more rapid increase of surface temperature in June. Note that the anomalous cooling in the western basin (also seen in Figure 10) in E3SM1 is present through March to September, lying east of the warm anomalies from March to June and then retreating west afterwards. The potential model biases leading to these cool anomalies are worthy of further investigation. The space-time structure has a pattern correlation of 0.69 and 0.79 in E3SM1 and E3SM2, respectively. The pattern strength is slightly stronger than observed (rms = 0.20) in E3SM1 (rms = 0.21) and slightly weaker in E3SM2 (rms = 0.18).

The seasonal timings in CESM1 and CESM2 are broadly consistent with the observations. Interestingly, despite having different seasonal timing, CESM1 and E3SM1 share several similarities as they both feature a rapid increase of temperature with a stronger peak than observed and both have a similar cooling bias in the western basin. In CESM2, the ensemble mean space-time structure agrees best with the observations, with a pattern correlation of 0.93 and the same pattern strength as the observed (rms = 0.20).

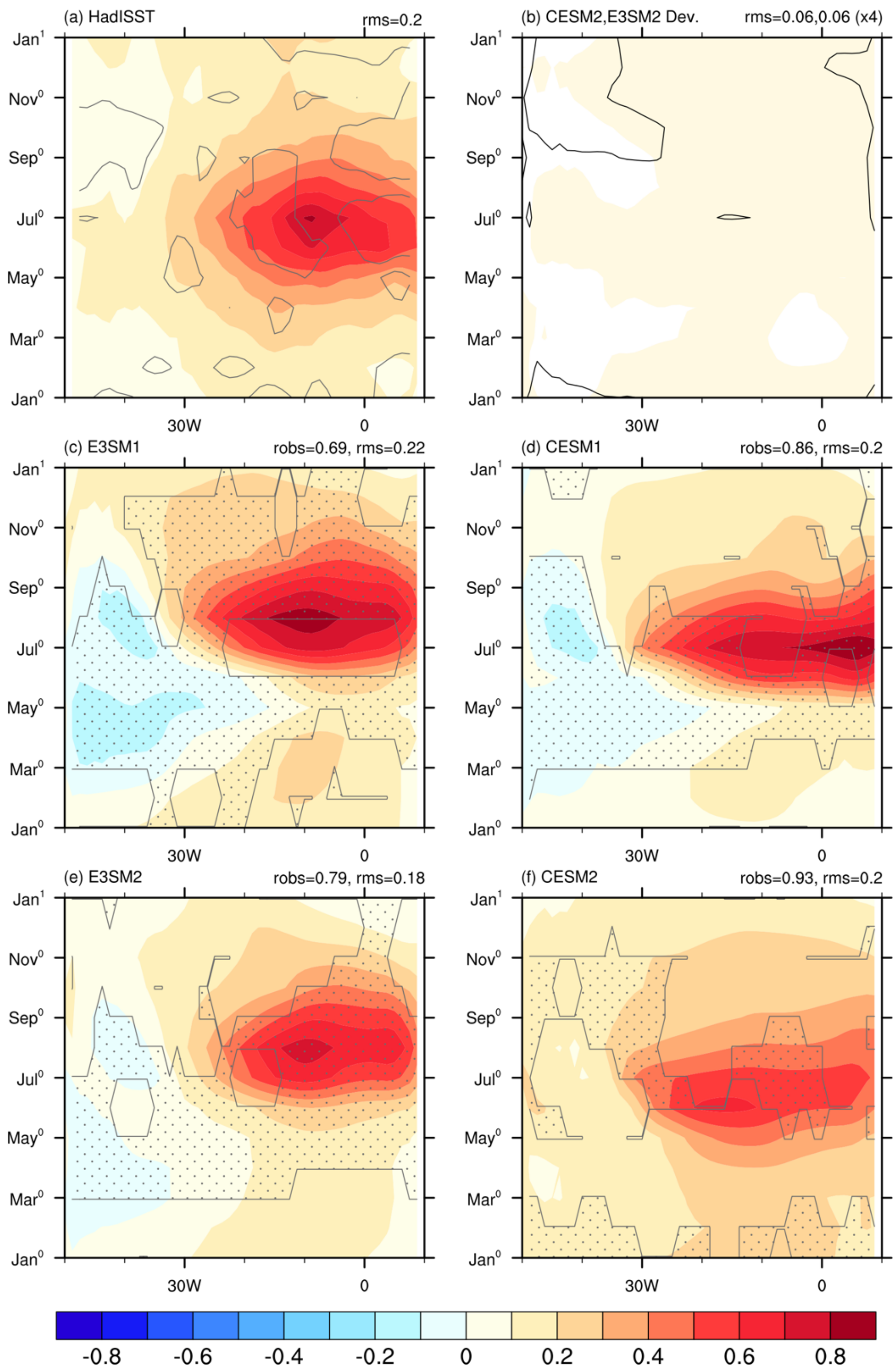

Fig. 12. The observed Atlantic Niño hovmoëller pattern in SST estimated from observations (HadISST, **a**) and the standard deviation in the pattern across members of CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with ERSSTv5 exceed 0.1˚C and in (**c-f**) where observations lie outside of each ensemble. Pattern correlations and pattern magnitudes are indicated in the panel captions.

f. Extratropical Modes

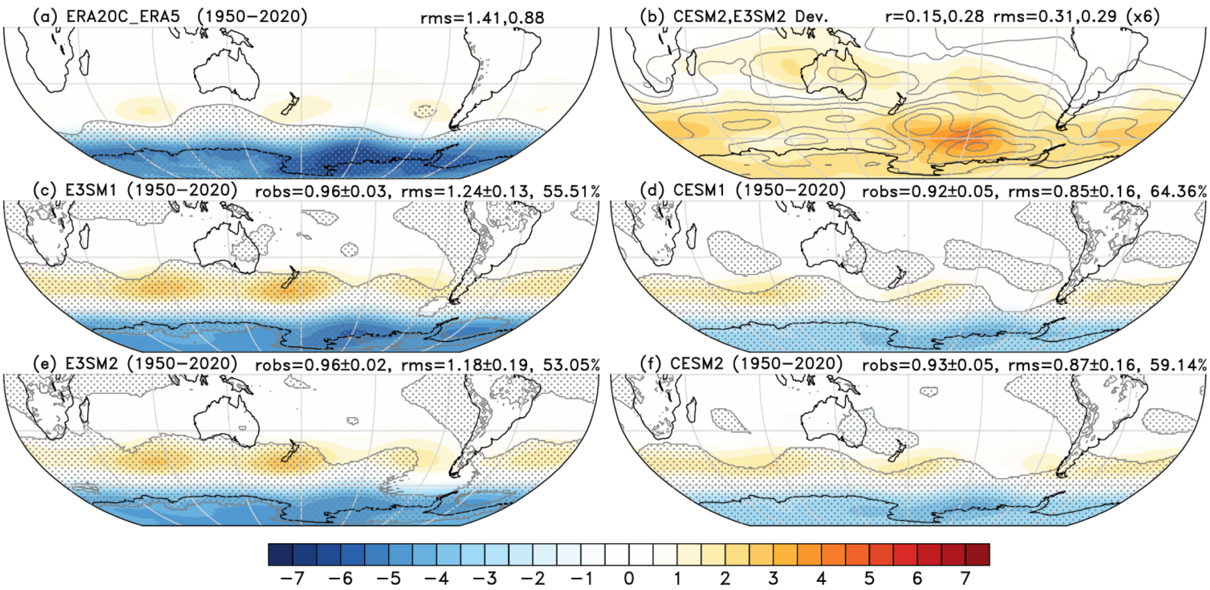

Fig. 13. The SAM teleconnection pattern in sea level pressure (PSL, hPa) in DJF estimated from surface-based reanalyses since 1950 (ERA20C/ERA5, **a**) and the standard deviation in the pattern across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with CERA20C/ERAI exceed 0.5 hPa and stippling in (**c-f**) where observations lie outside of the 2 σ range. Metrics in the panel headers are consistent with those in Fig. 3.

The spatial structure of the SAM teleconnection, computed as the leading empirical orthogonal function (EOF) from monthly PSL south of 20ºS (Thompson et al. 2000), is shown in Figure 13. Due to poor data sampling in the early 20th Century, the pattern from the reanalyses is only computed over 1950-2019. It is characterized by moderate negative values over the Antarctic continent, with maxima along the coast of west Antarctica and in the Ross Sea, and weak positive values in the Southern Indian Ocean and near New Zealand (Fig. 13a). The pattern of intra-ensemble spread in CESM2 (Fig. 13b) is characterized by the greatest values northeast of the Ross Sea, with a regional minimum south of Australia. It differs slightly from the pattern of spread in E3SM2, which shows a maximum southeast of New Zealand (lines in Fig. 13b). The ensemble mean pattern in E3SM1 (Fig. 13c) captures the observed pattern, as do all the ensembles ($0.92 < r < 0.96$), with values that are weaker (rms=1.24) than ERA20C/ERA5 (rms=1.41) but stronger than CERA20C/ERAI (rms=0.88). In CESM1 the ensemble mean pattern is weaker than both observational estimates (rms=0.85), though the CERA20C/ERAI estimate lies within the ensemble spread, while in E3SM2 the ensemble mean lies between them (rms=1.16). In CESM2, the pattern is similar to that in CESM1, with the CERA20C/ERAI estimate lying within the ensemble spread. Stippling at low latitudes indicates a weaker connection of the SAM to the tropics in the models, consistent with broader model evaluation (Lim et al. 2016).

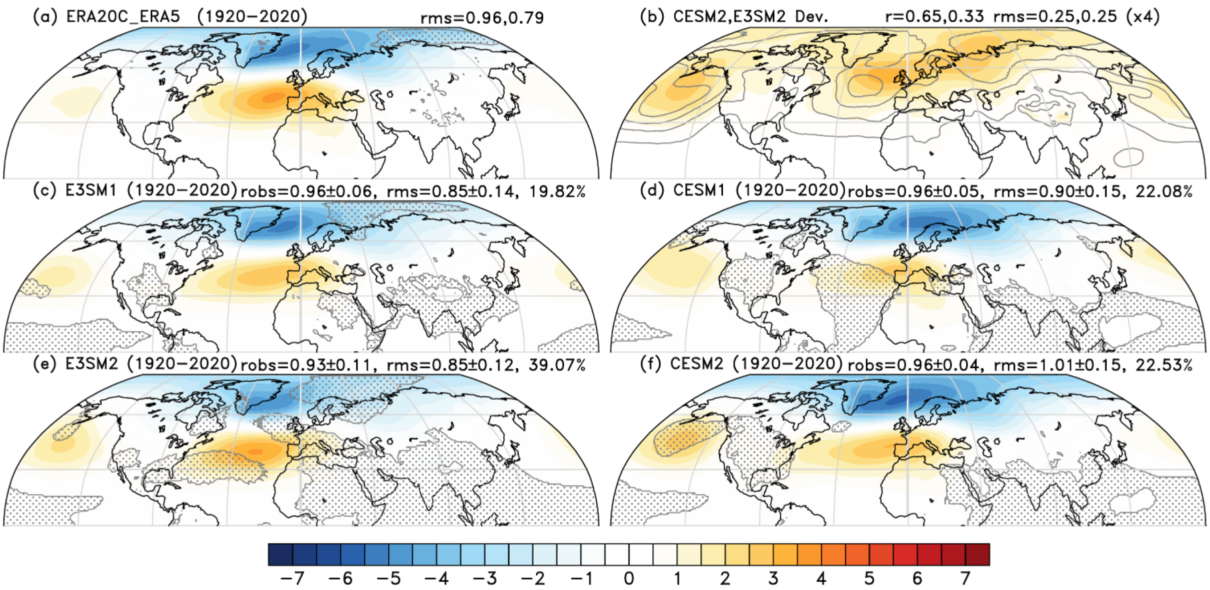

Fig. 14. The observed NAO teleconnection pattern in PSL in DJF estimated from observations (ERA20C/ERA5, **a**) and the standard deviation in the pattern across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with CERA20C/ERAI exceed 0.3 hPa and stippling in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

The spatial structure of the NAO teleconnection, estimated from the leading EOF of PSL in the region 20º-80ºN, 90ºW-40ºE following Hurrell and Deser (2009), is explored in Figure 14. The observed pattern (Fig. 14a) is characterized by strong negative values across the Arctic with a regional maximum east of Greenland and positive values west of Spain. The pattern of intra-ensemble spread in CESM2 (Fig. 14b) is characterized by the greatest values in the North Atlantic, North Pacific, and western Arctic Oceans, and is similar to the pattern of spread in E3SM2. The ensemble mean pattern in E3SM1 (Fig. 13c) correlates strongly with observations (r=0.96) with a magnitude (rms=0.85) that is slightly weaker than in ERA20C/ERA5 (rms=0.96) but stronger than in CERA20C/ERAI (rms=0.79). In CESM1 the strength of the pattern (rms=0.90) also lies between the observational estimates, and the pattern correlation is again high (r=0.96). In E3SM2, that pattern is very similar to E3SM1 (r=0.93, rms=0.85) while CESM2 exhibits the closest agreement with observations (r=0.96).

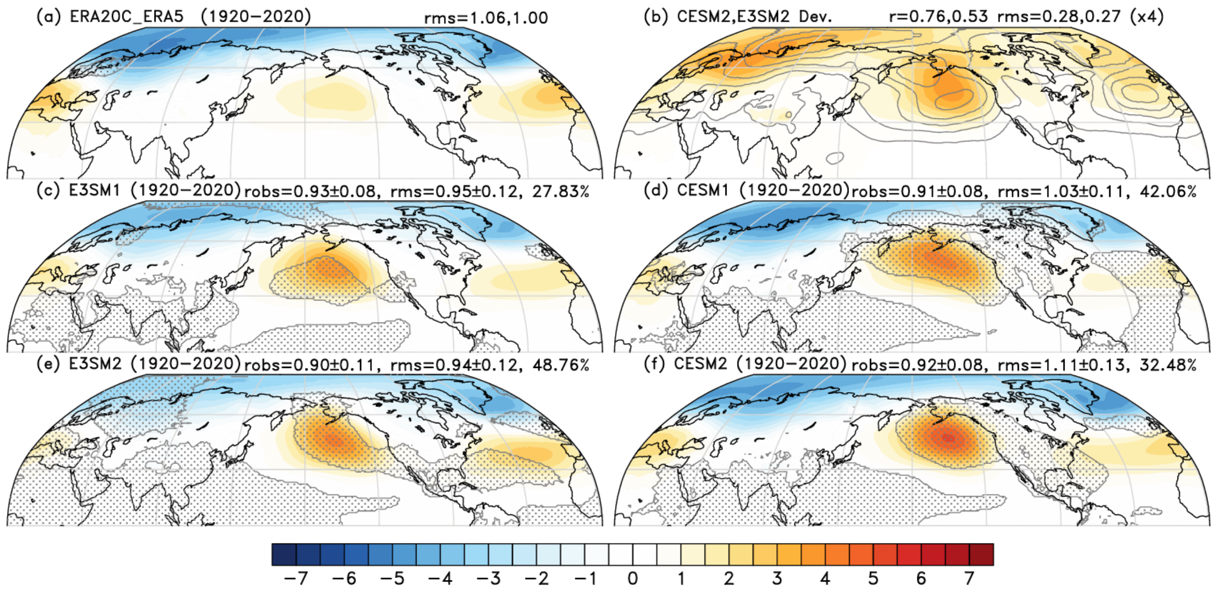

Fig. 15. The observed NAM teleconnection pattern in PSL (hPa) in DJF estimated from observations (ERA20C/ERA5, **a**) and the standard deviation in the pattern across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with CERA20C/ERAI exceed 0.5 hPa and stippling in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

The spatial structure of the NAM teleconnection, computed as the leading empirical orthogonal function (EOF) from PSL north of 20ºN (Thompson et al. 2000), is shown in Figure 15. The observed pattern (Fig. 15a) is characterized by strong negative values across the Arctic with weak regional maxima in the North Pacific and Atlantic Oceans and over Europe. The pattern of intra-ensemble spread in CESM2 (Fig. 15b) is characterized by the greatest values in the North Pacific and western Arctic Oceans and is similar to the pattern of spread in E3SM2. The ensemble mean pattern in E3SM1 (Fig. 15c) correlates strongly with observations (r=0.93) with a magnitude (rms=0.95) that is weaker than observed (rms=1.09) but observed strength is encompassed by the ensemble spread. For all ensembles, the magnitude of the ensemble-mean pattern in the North Pacific is much greater than that observed, and the observed values lie outside of the model ensemble ranges (stippling). In CESM1 the strength of the NAM pattern (rms=1.03) falls between the observational estimates, and the pattern correlation is again high (r=0.91). In E3SM2, the NAM pattern is very similar to that in E3SM1 (r=0.90, rms=0.94) while CESM2 exhibits the closest agreement with the observed pattern (r=0.92) but a magnitude that may be too large (rms=1.11) though the ensemble spread encompasses observed estimates. As for all models, biases in the North Pacific in CESM2 are strong.

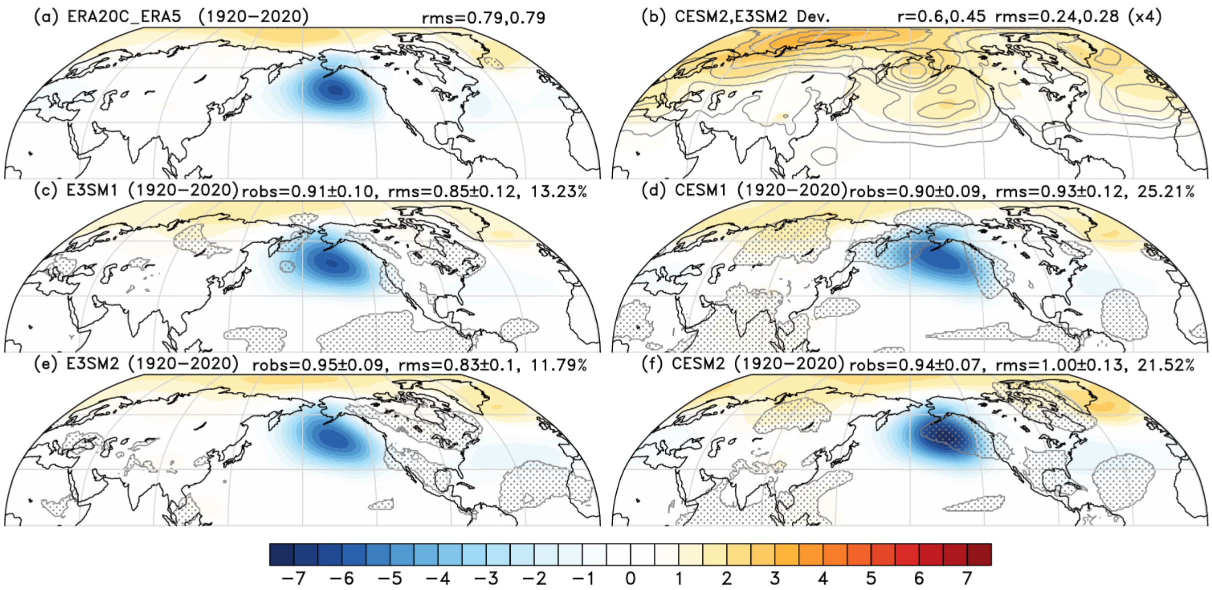

Fig. 16. The observed PNA teleconnection pattern in PSL (hPa) in DJF estimated from observations (ERA20C/ERA5, **a**) and the standard deviation in the pattern across ensemble members from CESM2 (contours) and E3SM2 (lines, **b**). The ensemble-mean patterns estimated from E3SM1 (**c**), CESM1 (**d**), E3SM2 (**e**), and CESM2 (**f**) are also shown. Contours in (**a**) where differences with CERA20C/ERAI exceed 0.3 hPa and in (**c-f**) where observations lie outside of each ensemble. Metrics in the panel headers are consistent with those in Fig. 3.

The spatial structure of the PNA teleconnection is shown in Figure 16. The observed pattern (Fig. 16a) is characterized by strong negative values in the North Pacific's Aleutian Low region and weak negative values spanning the Arctic. The pattern of intra-ensemble spread in CESM2 (Fig. 16b) is characterized by the greatest values in the western Arctic and the patterns of spread in CESM2 and E3SM2 agree closely. The ensemble mean pattern in E3SM1 (Fig. 16c) correlates strongly with observations (r=0.91) with a magnitude (rms=0.85) that is stronger than observed (rms=0.80), though the observations fall within the ensemble spread. In CESM1 the strength of the pattern (rms=0.93) is even stronger than in E3SM1, and the pattern correlation is again high (r=0.90). In E3SM2 (r=0.95) and CESM2 (r=0.94), the PNA patterns correlate strongly with observations, though the pattern in CESM2 is stronger than observed (rms=1.00), particularly in the North Pacific Ocean.

High pressure blocking patterns in the mid-latitudes are large-scale modes of extratropical variability most commonly associated with extremes of surface temperature and precipitation (Pfahl, 2014). However, unlike mobile baroclinic weather systems, atmospheric blocking characteristics: onset, duration, and decay, are difficult for models to simulate and predict. This is particularly true during northern winter when events are most prevalent, and the zonal jet interactions are most complex.

Figure 17 shows the meridional distribution of blocking frequency in models and reanalysis. Based on the measure used here, sectors in western Europe and the central Pacific can be ‘blocked’ about a quarter of the time. The general pattern of model performance is that the frequency of blocking is well captured in the Pacific more poorly simulated over western Europe. The general eastward shift of the maximum frequency into Eastern Europe, leaving an underestimate in the eastern Atlantic is a typical shortcoming in CMIP models, with little improvement between CMIP3 and CMIP6 (Davini, P., and F. D’Andrea, 2020; Schiemann et al., 2020). Among the models presented here the greatest disagreement and biases exist within the Euro-Atlantic sector. It is worth noting that during MAM, the agreement and accuracy across all the ensembles is much greater (not shown). Over the Greenland region ensemble disagreement is greatest, with E3SM1 and especially E3SM2 showing a pronounced maximum. This may be due to the mean climate bias in the region, where model SSTs are substantially cooler then observed (Golaz et al., 2019, Golaz et al., 2022). However, the pronounced intra-ensemble spread shows that blocking frequency in some multi-decadal periods (1979-2005) can easily overlap with ensemble members from other models. In the Pacific sector, blocking is often associated with downstream Rossby wave radiation from the tropics. With multiple ensemble members we are able to stratify the 2 modeling systems, such that E3SM maximum blocking frequency is too strong, but positioned correctly, whereas CESM frequency is accurate, but positioned too far east. This would not be discernible if using a single or just a few ensemble members, illustrating an additional benefit of model evaluation using LEs.

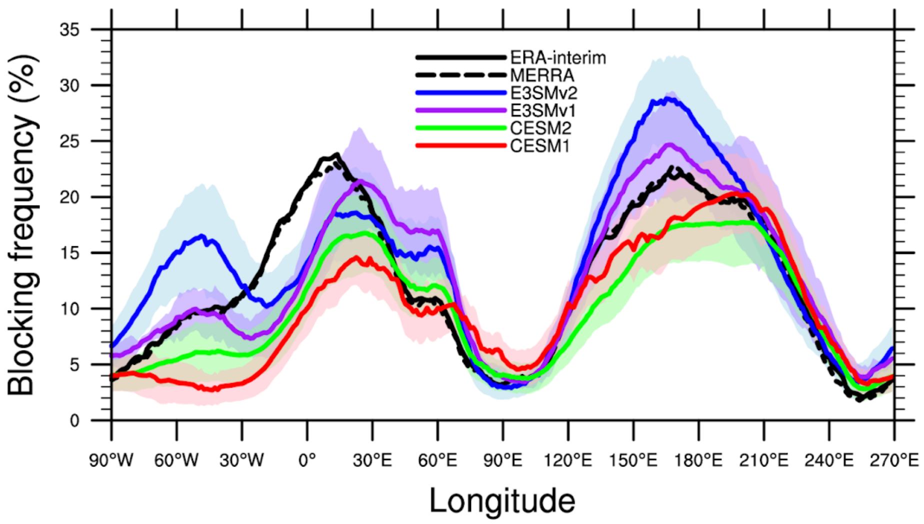

Fig. 17. Ensemble averaged blocking for E3SM1 (5 members), E3SM2 (20 members), CESM2 (50 members) and CESM1 (40 members) and reanalyses (ERA-interim and

MERRA2) for the period 1979-2005. The diagnostic is based on the frequency of a daily 500-hPa height field, meridional gradient threshold metric from D'Andrea et al., (1998). The shaded regions are -/+ 1 standard deviation across each ensemble. Because of limited model data availability, the version of E3SM1 analyzed here is from the 5-member CMIP6 submission for the period 1980-2006.

A) Pattern Correlations

	ENSO TAS (DJF^{+1})	ENSO PSL (DJF^{+1})	El Niño Hovmöller	La Niña Hovmöller	AMV' Low-Pass	PDV'	NAO (JFM)	PNA (DJF)	SAM (DJF)	SST std dev (Ann)	PR std dev (Ann)	Mean Score
E3SM1 (10%)	0.54	0.67	0.68	0.74	0.28	0.69	0.90	0.91	0.95	0.63	0.81	0.79
E3SM1 (Avg)	0.63	0.76	0.75	0.77	0.39	0.80	0.94	0.95	0.96	0.66	0.83	0.82
E3SM1 (90%)	0.72	0.82	0.79	0.81	0.51	0.87	0.96	0.96	0.97	0.67	0.83	0.83
CESM1 (10%)	0.51	0.74	0.83	0.74	0.15	0.84	0.85	0.89	0.87	0.72	0.77	0.79
CESM1 (Avg)	0.60	0.79	0.86	0.80	0.29	0.87	0.90	0.92	0.89	0.74	0.77	0.80
CESM1 (90%)	0.67	0.84	0.88	0.84	0.37	0.89	0.93	0.94	0.91	0.75	0.78	0.82
E3SM2 (10%)	0.56	0.74	0.87	0.83	0.30	0.74	0.88	0.91	0.95	0.61	0.82	0.84
E3SM2 (Avg)	0.69	0.81	0.90	0.90	0.43	0.84	0.91	0.94	0.97	0.64	0.83	0.85
E3SM2 (90%)	0.76	0.85	0.92	0.92	0.52	0.88	0.94	0.96	0.97	0.67	0.84	0.86
CESM2 (10%)	0.66	0.79	0.85	0.81	0.14	0.87	0.88	0.94	0.87	0.76	0.84	0.84
CESM2 (Avg)	0.72	0.85	0.88	0.84	0.34	0.88	0.92	0.96	0.90	0.77	0.85	0.85
CESM2 (90%)	0.77	0.88	0.90	0.87	0.48	0.90	0.95	0.97	0.92	0.78	0.86	0.86

B) RMS Differences

	ENSO TAS °C	ENSO PSL hPa	El Niño Hovmöller °C	La Niña Hovmöller °C	AMV' Low-Pass °C	PDV' °C	NAO (JFM) hPa	PNA (DJF) hPa	SAM (DJF) hPa	SST std dev (Ann) °C	PR std dev (Ann) mm day^{-1}	Mean Score
E3SM1 (10%)	0.57	1.08	0.39	0.30	0.74	0.08	0.45	0.38	0.79	0.18	0.18	0.55
E3SM1 (Avg)	0.70	1.27	0.44	0.35	0.90	0.10	0.61	0.48	0.84	0.19	0.18	0.58
E3SM1 (90%)	0.77	1.43	0.49	0.43	1.01	0.11	0.76	0.57	0.90	0.20	0.18	0.61
CESM1 (10%)	0.71	1.17	0.53	0.46	0.80	0.07	0.64	0.50	1.19	0.10	0.20	0.63
CESM1 (Avg)	0.82	1.43	0.63	0.59	1.06	0.09	0.78	0.63	1.29	0.11	0.21	0.69
CESM1 (90%)	0.93	1.70	0.78	0.70	1.31	0.10	0.91	0.73	1.40	0.12	0.21	0.75
E3SM2 (10%)	0.59	0.99	0.29	0.24	0.73	0.07	0.53	0.41	0.72	0.18	0.18	0.53
E3SM2 (Avg)	0.70	1.21	0.37	0.32	0.87	0.09	0.74	0.51	0.82	0.19	0.18	0.56
E3SM2 (90%)	0.81	1.43	0.46	0.39	1.04	0.10	0.84	0.57	0.90	0.20	0.18	0.60
CESM2 (10%)	0.61	1.06	0.52	0.53	0.76	0.09	0.59	0.47	1.15	0.13	0.19	0.62
CESM2 (Avg)	0.71	1.30	0.62	0.70	1.08	0.10	0.78	0.58	1.30	0.14	0.21	0.70
CESM2 (90%)	0.78	1.44	0.74	0.83	1.48	0.12	0.94	0.74	1.40	0.16	0.22	0.79

Table 2. CVDP-LE summary metrics of model performance for E3SM1, E3SM2, CESM1, and CESM2 based on pattern correlations (**a**) and root-mean-squared pattern magnitudes of differences with observed modes (**b**).

To provide a broad overview of model performance in simulating MoV, the skill scores from the CVDP-LE are summarized in Table 2. They demonstrate many aspects of the simulated modes already discussed, including the improvements in fidelity across generations of E3SM and CESM whose average pattern correlation scores increase from 0.82 to 0.85 and

0.80 to 0.85, respectively. The strength of the patterns also improves across successive generations of E3SM, from 0.58 to 0.56, but degrades slightly in CESM (0.69 to 0.70) though observational uncertainty in magnitude is also large as discussed in Section 3.f. The challenges faced by each model in simulating some modes (e.g. PDV in CESM) is also evident and the broad spread of the model ensembles in simulating some modes (e.g. between 10% and 90% ranges of PDV) illustrates the inherent limitations of model evaluation with an observational record of limited length, as intrinsic variability drives a broad range of skill scores within each ensemble (0.69 to 0.87 in E3SM1).

g. The Quasi-Biennial Oscillation

The QBO is the dominant mode of variability in the tropical lower stratosphere (Baldwin et al. 2001) and is characterized by descending easterlies and westerlies from about 10 to 100 hPa with an average period of 28 months. The QBO is important to surface climate through its various teleconnections (see Anstey et al. 2022 for review), with an influence deep convection, the MJO, and precipitation (e.g. Collimore et al., 2003, Liess and Geller 2012, Yoo and Son 2016, Son et al. 2017). It also influences the subtropical jet (Garfinkel and Hartmann, 2011, Wang et al. 2018), and the stratospheric polar vortex which subsequently affects the NAO (Holton and Tan 1980, Anstey and Shepherd 2014).

Simulation of the QBO in global models is improving. Only 5 models out of CMIP5 were able to simulate the QBO, however 15 models in CMIP6 succeeded to do so (Richter et al. 2020). Biases in its representation have persisted however with models underestimating its amplitude in the lowermost stratosphere, a likely reason for underrepresentation of many of the QBO's teleconnections and its MJO coupling (Richter et al. 2020, Kim et al. 2020).

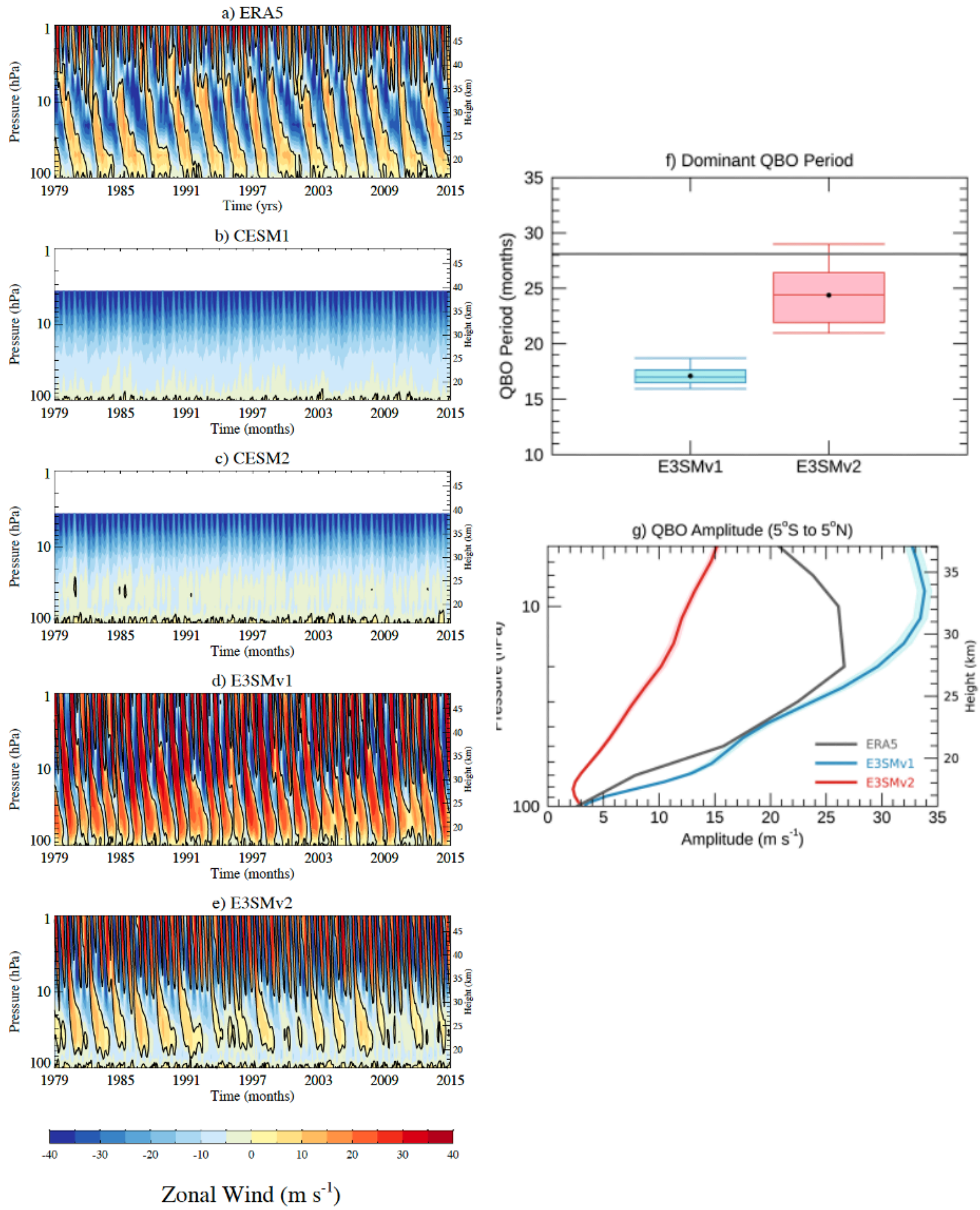

Fig. 18. Zonal-mean zonal wind averaged between 5°S and 5°N between 1979 and 2014 for a) ERA5, b) CESM1, c) CESM2, d) E3SM1, e) E3SM2. The first ensemble member is shown for the models. f) Distribution of dominant QBO periods across E3SM1 and E3SM2 ensembles (whiskers) The distribution median is depicted by horizontal line in each box, box edges mark the lower and upper quartiles, and box whiskers mark the minimum and maximum values. Black dots represent mean values. The dominant QBO period for ERA5 is marked by the gray line. g) QBO amplitude calculated using the DD method for ERA5, E3SM1 and E3SM2. Shading represents the +/- standard error.

Panels (a) - (e) of Figure 18 illustrate the basic features of tropical stratospheric variability in ERA5 and in the CESM and E3SM large ensembles. Tropical zonal mean winds in ERA5 show alternating descending easterlies and westerlies with clearly defined phases between 10 and 100 hPa (Fig 18 a). CESM1 and CESM2 have a relatively low model top, and do not employ a non-orographic gravity wave parameterization, hence the tropical stratospheric mean winds are mostly easterly (Fig 18b, c) and do not simulate the QBO. On the other hand, E3SM has a higher model top and includes a non-orographic gravity wave parameterization following Richter et al. 2010 with different tunings (Rasch et al. 2019, Golaz et al. 2022). E3SM1 produces an oscillation in the tropical stratospheric zonal mean wind between 10 and 100 hPa that primarily consists of westerlies that often reach 30 to 40 m/s in amplitude, much stronger than in observations (Fig 18d). In addition, the period of the oscillation is shorter than in observations, and the easterly phases are weak and brief. On the other hand, the tropical stratospheric zonal mean wind in E3SM2 shows much weaker westerlies ,which propagate downward from 10 to 50 hPa, and alternate with very weak westerlies. Hence, while E3SM1/2 are able to simulate the QBO, their biases are substantial.

In Figure 18, the QBO period and amplitude are also assessed in E3SM1 and E3SM2 across the large ensembles (Fig. 18f, g). The dominant QBO period is computed by applying the fast Fourier transform to the deseasonalized time series of zonal-mean zonal wind at 20 hPa, averaged between 5°S and 5°N, following Richter et al. 2020. QBO amplitude is calculated using the Dunerketon and Delisi (1985) method which was also used in Richter et al. (2020) and Richter et al. (2022). Using this method, the mean amplitude of the QBO is approximated by the square-root of 2 times the standard deviation of the deseasonalized time series of the monthly mean westerly winds. This method is particularly useful when the phases of the QBO are not coherent, as in the simulation of the QBO in E3SM2. Figure 18f shows that the average QBO period in E3SM1 over the historical period from 1979 to 2014 is 17.1, lower than the QBO period in ERA5 of 28.1 months. The mean QBO period in E3SM2 is much closer to that in ERA5 at 24.4 months. The inter-ensemble spread of QBO periods is approximately twice as large in E3SM2 as compared to E3SM1. The amplitude of the QBO in E3SM1 is similar to that derived from ERA5 in the lowermost stratosphere, between 30 and 100 hPa (Fig 18f). On the other hand, E3SM2 largely underestimates the QBO amplitude throughout the depth of the stratosphere.

An analysis of the reasons for the differences between the QBO in E3SM1 and E3SM2 are undergoing detailed examination and are beyond the scope of this work. The differences likely result from both different gravity wave parameterization tunings and changes in convection parameterization that occurr from E3SM1 to E3SM2 (Golaz et al. 2022). Additionally, we examine the E3SM1 and E3SM2 models for the observed boreal wintertime QBO-MJO relationship, where the easterly phase of the QBO has been shown to increase MJO activity over the maritime continent (Yoo and Son, 2016). We find that the models do not produce a robust relationship between these two modes of variability, which is consistent with other CMIP6 models (Kim et al 2020).

Figure 19a shows that in ERA5 the boreal stratospheric polar vortex is weakened during the easterly QBO (EQBO) winter and strengthened during the westerly QBO (WQBO) winter, known as the Holton-Tan effect (Holton and Tan, 1980; Lu et al., 2020). E3SM1 produces the Holton-Tan effect in the boreal winter, although the stratospheric polar vortex during EQBO weakens less than in observations (Figure 19b). E3SM2 fails to produce the Holten-Tan effect (Figure 19c), which may be related to the low amplitude of the QBO in the lower stratosphere in E3SM2 as compared to ERA5. The subtropical pathway is proposed as a possible route for the QBO in the lower stratosphere to modulate the tropospheric subtropical jet (Garfinkel and Hartmann, 2011; Haynes et al., 2021). Although E3SM1 and E3SM2 produce statistically significant responses to the QBO in the subtropical jet region, the patterns of the responses in models are biased. Figure 19d shows that there is a negative phase PNA pattern corresponding to EQBO in the boreal winter in the ERA5. E3SM1 produces a similar wave train pattern in the North Pacific and America (Figure 19e), but the magnitude is much weaker than that depicted in ERA5. For other regions, such as the Atlantic Ocean, the geopotential height response to the QBO in E3SM1 is opposite to that observed. The response in E3SM2 has an even larger bias (Figure 19f).

Lastly, the teleconnection of the QBO to surface temperature is assessed. Figure 19g shows that the surface temperature anomalies have a La Niña-like pattern during EQBO and El Niño-like pattern during WQBO in ERA5. A higher frequency of El Niño during the WQBO and of La Niña during the EQBO since 1979 has been observed (Taguchi, 2010; Liess and Geller, 2012). Both E3SM1 and E3SM2 produce a La Niña-like pattern during the EQBO phase (Figure 19h, i), but the amplitude in E3SM2 is larger than in E3SM1. However, both simulated patterns are weaker than observed.

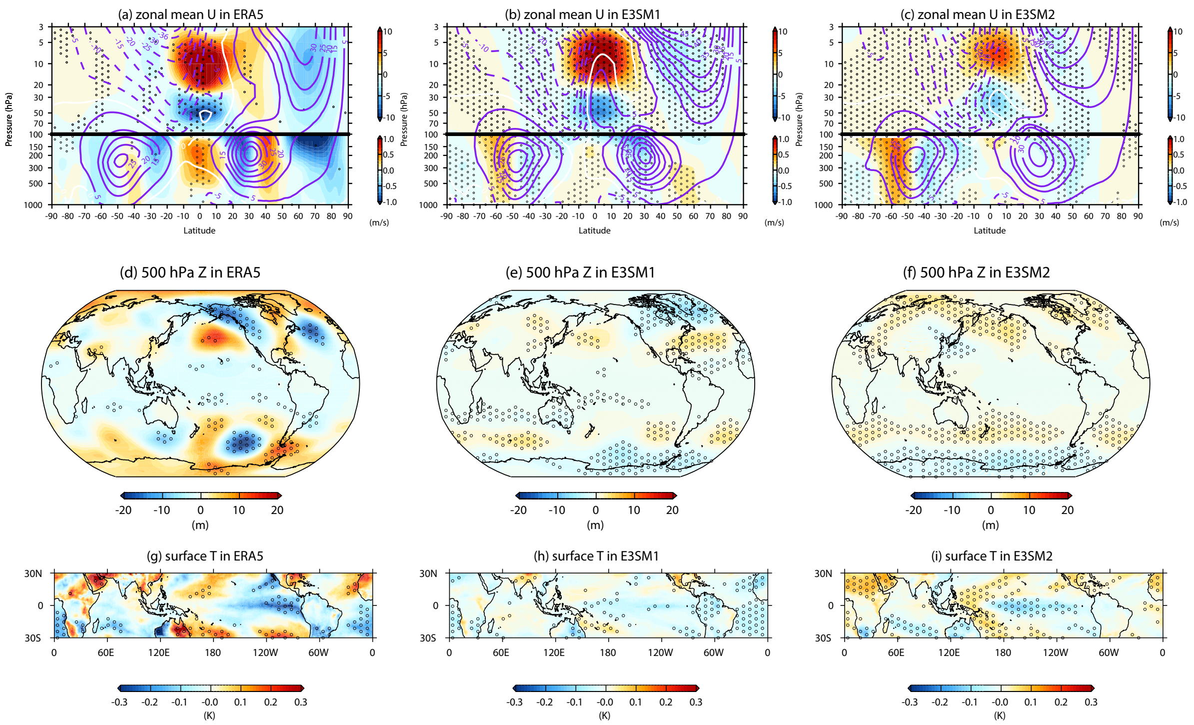

Figure 19. The zonal mean zonal wind (a-c), geopotential height at 500 hPa (d-f), and surface temperature (g-i) in boreal winter (DJF) are regressed to the normalized QBO index by the simple linear regression model using an ordinary least-squares (OLS) method from ERA5 (left column), E3SM1 (middle column) and E3SM2 (right column). The QBO index is defined as the equatorial zonal mean zonal wind at 50 hPa. The analysis covers from 1979 to 2014. All the panels correspond to the EQBO situation. The contour lines in (a-c) depict the climatological mean. The one-tailed t-test is utilized and the dotted region is significant at the 90% confidence level.

h. The Madden Julian Oscillation

The Madden Julian Oscillation (MJO) is the largest mode of variability and the leading source of predictability on intraseasonal timescales (Zhang, 2005). It also has large, but complex interactions with slow modes of variability such as El Niño (e.g, Moon et al., 2011). While skill has improved recently (Li et al., 2022), the MJO remains a difficult mode of variability to simulate (Jiang et al., 2020). As a consequence, this restricts the predictive skill of models that exploit the inherent predictability associated with the MJO (Wang et al., 2018). Figure 20 shows the percentage of the total intraseasonal (20-100-day filtered) variance explained by contribution from the first and second combined EOFs. On the whole, the MJO has experienced an increase in simulation skill from the baseline CESM1 model to the successive CESM2 and E3SM1/2 models. The overall bias in this metric has been reduced by around half, on average, and the improvements have been attributed to modifications to the physical parameterization in both modeling systems (Danabasoglu et al., 2020; Golaz et al . 2019).

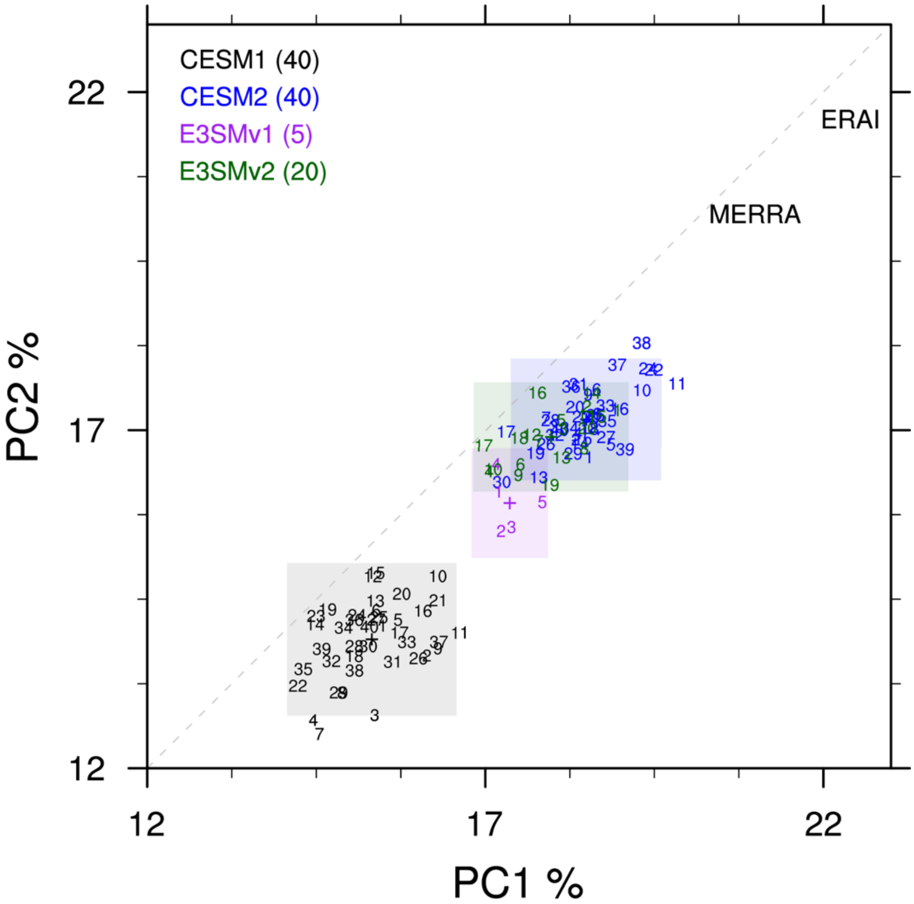

Fig. 20. The percentage of the total intraseasonal (20-100-day filtered) variance explained by contribution from the first and second combined EOFs (for global anomalies of 850-mb/200-mb zonal wind and outgoing longwave radiation). The domain of the shaded boxes are -/+ 1 standard deviation along each axis. The '+' represents the ensemble mean for each ensemble set. Because of limited model data availability, the version of E3SM1 analyzed here is from the 5-member CMIP6 submission for the period 1980-2006.

Although the average strength of the explained variance is greatest in CESM2, the ensemble variability over this multi-decadal time frame (1979-2005) is substantial. Variations in the underlying tropical SSTs can dominate the MJO response, such that comparisons of single simulations across models would not be able to determine the correct skill ordering reliably. For example, E3SM1 member 4, E3SM2 10 and CESM2 member 10 are statistically indistinguishable and yet the ensemble mean behavior is quite well separated. Such a large intra-ensemble variability over these multi decadal timescales prompts further understanding of the MJO's relationship to ENSO and the PDV, which have been seen in the observational record (e.g., Dasgupta et al. 2020; Tang and Yu, 2008).

4. Discussion and conclusions

The importance of MoV to a range of climate science questions and applications motivates an assessment of their fidelity across climate models. Here the performance of versions 1 and 2 of E3SM and CESM is evaluated using LEs. The availability of these LEs allows an assessment of the intrinsic variability in MoV and estimation of inter-model differences. Contrasts between model generations and between E3SM and CESM are identified in multiple modes that enable an assessment of suitability for purpose of the models in a broad range contexts. For example, biases are identified in patterns of PDV that are too weak in the Tropics in E3SM and that are displaced westward in both E3SM and CESM (Fig. 2). For El Niño and La Niña, the pattern of anomalies in the tropics are again displaced westward from those observed and composite El Niño teleconnections are too weak (Figs. 3, 4), particularly given the excessive power of Niño3.4 SST variability in CESM2 (Fig. 9). While simulated La Niña teleconnections agree more closely with those observed, the excessive magnitude of Niño3.4 SST anomalies calls into question both the depiction of conditions in the Tropics and their linkages to the extratropics. The temporal structure of both El Niño and La Niña in E3SM suggests tropical variability that is overly biennial, though a strong zonal gradient in variability is also identified, precluding a simple characterization of spectral bias in E3SM1 or E3SM2 (Figs. 7, 8).

E3SM1 and E3SM2 are able to simulate an internally generated QBO (Fig. 18). However, the representation of it in both versions of the model is biased from observations in various ways. The QBO in E3SM1 has a period much too short as compared to observations, although the amplitude is reasonable. On the other hand, the QBO in E3SM2 has a period closer to the observed QBO, however the amplitude is too weak. E3SM1 is able to capture the teleconnection of the QBO to the polar vortex (Fig. 19). Both E3SM versions underestimate the connection to the subtropical jet, however they are able to capture the pattern of teleconnections to surface temperatures although that connection is also weak.

Finally, MoV based on higher frequency phenomena generally show similar improvements across model versions for E3SM and CESM (Fig. 20). For the MJO this improvement is dramatic in CESM, where there is very limited activity in CESM1. For E3SM the improvement is less dramatic, consistent with the similar model configurations of CESM2 and E3SM1/v2. Among all models the decadal variation in the MJO strength across ensembles is substantial and often overlapping. Blocking frequency in E3SM is, overall, greater than in CESM. Common to many models the largest NH biases are in the Atlantic sector. In particular, the large positive biases in E3SM may reflect the cool ocean biases in this region.

Model	Overall	Energy	Water	Dynamics	Mean	Annual	ENSO
E3SM1	0.778±0.008	0.782±0.009	0.745±0.008	0.802±0.009	0.875±0.001	0.874±0.002	0.583±0.023
E3SM2	0.801±0.008	0.821±0.008	0.767±0.008	0.816±0.009	0.885±0.001	0.873±0.001	0.653±0.024
CESM1	0.803±0.004	0.809±0.004	0.762±0.004	0.839±0.004	0.889±0.001	0.887±0.000	0.640±0.011
CESM2	0.814±0.004	0.827±0.003	0.772±0.004	0.843±0.004	0.909±0.001	0.893±0.001	0.647±0.010

Table 3. CMATv1 summary metrics of model performance for E3SM1, E3SM2, CESM1, and CESM2 (Fasullo, 2020). Uncertainties correspond to the 2 standard error range.

The improvements identified in simulating the patterns and magnitudes of many major MoV have been accompanied by improvements in the depiction of the mean climate state, as summarized with version 1 of the Climate Model Analysis Tool (CMATv1, Fasullo 2020, Table 3). CMATv1 computes pattern correlations between simulated fields and benchmark observations for fields such as radiative fluxes and atmospheric energy transports, precipitable water and precipitation, and near surface wind speed and 500 hPa eddy geopotential height. It also considers a range of time domains such as the mean climatology, seasonal contrasts, and variability during ENSO. It is not known whether the association between improved MoV and base state climate is causal, and indeed the intrinsic variability of modes identified here (e.g. Fig. 1), highlights the difficulty in tracking mode fidelity between single realizations during model development. However, the association between improvements in modes and model mean state is suggestive of improvements in the physical process that govern each. Notably, the model versions assessed here use identical atmospheric grids, and similar ocean grids, and thus contrasts between the model versions are certainty not the result of model resolution. Rather, inter-model contrasts suggest a role for improved physical representations, which in CESM include major changes to clouds and convection (Danabasoglu et al. 2020), and in E3SM include significant changes to radiation and precipitation (Table 3; Golaz et al. 2019). Continued improvements in these fields are suggested as a plausible path forward for continued depiction of MoV in models.

Acknowledgments.

This work was supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological & Environmental Research (BER) under Award Number DE-SC0022070. Development of the E3SM model is supported as part of the Energy Exascale Earth System Model (E3SM) project, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. This work also was supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. The efforts of Dr. Fasullo in this work were supported by NASA Awards 80NSSC17K0565 and 80NSSC22K0046, and NSF Award 2103843. Computing resources (ark:/85065/d7wd3xhc) were provided by the Climate Simulation Laboratory at NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation and other agencies. The CESM project is supported primarily by the National Science Foundation. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC awards BER-ERCAP0020104 (E3SM2-LE-SSP370) and BER-ERCAP0025058 (E3SM1-LE). Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.

Data Availability Statement.

The CVDP-LE output and analysis metrics of other modes presented in this study have been made available via NCAR’s Global Data Exchange (https://gdex.ucar.edu/dataset/378_caron.html).

REFERENCES

Anstey, J. A., and T. G. Shepherd, 2014: Review Article High-latitude influence of the quasi-biennial oscillation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1–21, https://doi.org/10.1002/qj.2132.

Anstey, J.A., Osprey, S.M., Alexander, J. et al., 2022: Impacts, processes and projections of the quasi-biennial oscillation. *Nat Rev Earth Environ,* **3**, 588–603, https://doi.org/10.1038/s43017-022-00323-7.

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., et al., 2001: The quasi-biennial oscillation. *Review of Geophysics*, ***39***, 179–229, doi: 10.1029/1999RG000073.

Coburn, J., and S. C. Pryor, 2021: Differential credibility of climate modes in CMIP6. *J. Climate*, **34**, 8145–8164, https://doi. org/10.1175/JCLI-D-21-0359.1.

Collimore, C. C., D. W. Martin, M. H. Hitchman, A. Huesmann, and D. E. Waliser, 2003: On The Relationship between the QBO and Tropical Deep Convection. *J. Climate*, **16**, 2552–2568, doi: 10.1175/1520-0442(2003)016<2552:OTRBTQ>2.0.CO;2.

D'Andrea, F., S. Tibaldi, M. Blackburn, G. Boer, M. Déqué, M. R. Dix, B. Dugas, L. Ferranti, T. Iwasaki, A. Kitoh, V. Pope, D. Randall, E. Roeckner, D. Strauss, W. Stern, H. Van den Dool and D. Williamson, 1998: Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988. *Climate Dynamics,* **14,** 385-407, doi: https://doi.org/10.1007/s003820050230.

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

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D. A., Du- Vivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto- Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), 2020: *J. Adv. Model. Earth Sy.,* **12**, 35, https://doi.org/10.1029/2019ms001916.

Dasgupta, P., Metya, A., Naidu, C.V. et al. 2020: Exploring the long-term changes in the Madden Julian Oscillation using machine learning. *Sci. Rep*., **10**, 18567. https://doi.org/10.1038/s41598-020-75508-5

Davini, P., and F. D'Andrea, 2020: From CMIP3 to CMIP6: Northern Hemisphere Atmospheric Blocking Simulation in Present and Future Climate. *J. Climate*, **33**, 10021–10038, https://doi.org/10.1175/JCLI-D-19-0862.1.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., … and Vitart, F., 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, ***137***(656), 553-597, doi: 10.1002/qj.828.

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

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

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

Deser, C., I. R. Simpson, K. A. McKinnon, and A. S. Phillips, 2017: The Northern Hemisphere extra-tropical atmospheric circulation. response to ENSO: How well do we know it and how do we evaluate models accordingly? *J. Climate*, **30**, 5059–5082, https://doi.org/10.1175/JCLI-D-16-0844.1.

Deser, C., and A.S. Phillips, 2021: Defining the internal component of Atlantic multidecadal variability in a changing climate. *Geophysical Research Letters*, **48**(22), doi: 10.1029/2021GL095023.

Ding, H., N. Keenlyside, and M.Latif, 2012: Impact of the equatorial Atlantic on the El Niño Southern Oscillation. *Clim. Dynam.* **38**, 1965–1972, https://doi.org/10.1007/s00382-011-1097-y.

Dunkerton, T.J. and Delisi, D.P., 1985: Climatology of the equatorial lower stratosphere. *Journal of the Atmospheric Sciences*, **42**, 376–396.

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 de- sign and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Fasullo, J. T., 2020: Evaluating simulated climate patterns from the CMIP archives using satellite and reanalysis datasets using the Climate Model Assessment Tool (CMATv1). *Geoscientific Model Development*, **13**(8), 3627-3642, doi: 10.5194/gmd-13-3627-2020.

Fasullo, J. T., A. Phillips, and C. Deser, 2020: Evaluation of leading modes of climate variability in the CMIP archives. *J. Climate*, **33**, 5527–5545, https://doi.org/10.1175/JCLI-D-19-1024.1.

Fasullo, J. T., Lamarque, J. F., Hannay, C., Rosenbloom, N., Tilmes, S., DeRepentigny, P., ... and Deser, C., 2022: Spurious late historical-era warming in CESM2 driven by prescribed biomass burning emissions. *Geophysical Research Letters*, **49**, doi: 10.1029/2021GL097420.

Garfinkel, C. I., and D. L. Hartmann, 2011: The Influence of the Quasi-Biennial Oscillation on the Tro- posphere in Winter in a Hierarchy of Models. Part II: Perpetual Winter WACCM Runs. *Journal of the Atmospheric Sciences*, **68** (9), 2026–2041, https://doi.org/10.1175/2011JAS3702.1.

Gerber, E. P., and P. Martineau, 2018: Quantifying the variability of the annular modes: reanalysis uncertainty vs. sampling uncertainty. *Atmospheric Chemistry and Physics*, **18**(23), 17099-17117, doi: 10.5194/acp-18-17099-2018.

Golaz, J. C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., ... and Zhu, Q., 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, doi: 10.1029/2018MS001603

Golaz, J. C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., ... & Bader, D. C., 2022: The DOE E3SM Model Version 2: overview of the physical model and initial model evaluation. *Journal of Advances in Modeling Earth Systems*, **14**(12), doi: 10.1029/2022MS003156.

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

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., … & Thépaut, J. N., 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**(730), 1999-2049, doi: 10.1002/qj.3803.

Holton, J. R., and H.-C. Tan, 1980: The Influence of the Equatorial Quasi-Biennial Oscillation on the Global Circulation at 50 mb. *Journal of the Atmospheric Sciences*, **37** (10), 2200–2208, https://doi.org/10. 1175/1520-0469(1980)037<2200:TIOTEQ>2.0.CO;2, URL http://journals.ametsoc.org/doi/abs/10.1175/ 1520-0469%281980%29037%3C2200%3ATIOTEQ%3E2.0.CO%3B2.

Huang, B., Peter W. Thorne, et. al, 2017: Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5), Upgrades, validations, and intercomparisons. *J. Climate*, **30** (20), doi: 10.1175/JCLI-D-16-0836.1

Hurrell, J. W., and C. Deser, 2009: North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.*, **78**, No. 1, 28-41, doi:10.1016/j.jmarsys.2008.11.026.

Hurrell, J., and Coauthors, 2013: The Community Earth System Model: A framework for collaborative research. *Bull. Amer. Meteor. Soc.*, **94**, 1339–1360, doi:10.1175 /BAMS-D-12-00121.1.

Jiang, X., Adames, Á. F., Kim, D., Maloney, E. D., Lin, H., & Kim, H., et al., 2020: Fifty years of research on the Madden-Julian Oscillation: Recent progress, challenges, and perspectives. *J. Geo. Res.: Atm.,* **125**, e2019JD030911. https://doi.org/10.1029/2019JD030911.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M., 2015: The Community Earth System Model (CESM) Large Ensemble Project: A Community Re- source for Studying Climate Change in the Presence of Inter- nal Climate Variability, *B. Am. Meteorol. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/bams-d-13-00255.1.

Keenlyside, N. and M. Latif, 2007: Understanding equatorial Atlantic interannual variability. *J. Clim.* **20**, 131–142, https://doi.org/10.1175/JCLI3992.1.

Kim, H., Caron, J. M., Richter, J. H., and Simpson, I. R., 2020: The lack of QBO-MJO connection in CMIP6 models. *Geophysical Research Letters*, **47**, e2020GL087295. https://doi.org/ 10.1029/2020GL087295.

Krishnamurthy, V., and Coauthors, 2021: Sources of subseasonal predictability over CONUS during boreal summer. *J. Climate*, **34**, 3273–3294, https://doi.org/10.1175/JCLI-D-20-0586.1.

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

Laloyaux, P., E. de Boisseson, M. Balmaseda, J-R Bidlot, S. Broennimann, R. Buizza, P Dalhgren, D. Dee, L. Haimberger, H. Hersbach, Y. Kosaka, M. Martin, P. Poli, N. Rayner, E. Rustemeier, D. Schepers, 2018: CERA-20C: A couple reanalysis of the Twentieth Century, *JAMES*, **10**, 5, **https://doi.org/10.1029/2018MS001273**.

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

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, no. 12, 6307-6326, doi:10.1029/2018JD029522.

Li, Y., J. Wu, J. Luo, and Y. M. Yang, 2022: Evaluating the Eastward Propagation of the MJO in CMIP5 and CMIP6 Models Based on a Variety of Diagnostics. *J. Climate*, **35**, 1719–1743, https://doi.org/10.1175/JCLI-D-21-0378.1.

Liess, S., and M. A. Geller, 2012: On the relationship between qbo and distribution of tropical deep convection. *Journal of Geophysical Research: Atmospheres*, 117 (D3), https://doi.org/https://doi.org/ 10.1029/2011JD016317, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD016317,https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JD016317.

Lim, E.-P., H. H. Hendon, J. M. Arblaster, F. Delage, H. Nguyen, S.-K. Min, and M. C. Wheeler, 2016: The impact of the Southern Annular Mode on future changes in Southern Hemisphere rainfall, *Geophys. Res. Lett.*, **43**,7160–7167, doi:10.1002/2016GL069453

Lee, J., K. Sperber, P. Gleckler, C. Bonfils, and K. Taylor, 2019: Quantifying the Agreement Between Observed and Simulated Extratropical Modes of Interannual Variability. *Climate Dynamics,* 52, 4057-4089, doi: 10.1007/s00382-018-4355-4

Lee, J., K. Sperber, P. Gleckler, K. Taylor, and C. Bonfils, 2021: Benchmarking performance changes in the simulation of extratropical modes of variability across CMIP generations. *J. Climate*, **34**, 6945–6969, https://doi.org/10.1175/JCLI-D-20-0832.1.

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

Medhaug, I., Stolpe, M. B., Fischer, E. M., and Knutti, R., 2017: Reconciling controversies about the ‘global warming hiatus’. *Nature*, **545**(7652), 41-47, doi: doi.org/10.1038/nature22315.

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

Moon, JY., Wang, B. and Ha, KJ. 2011: ENSO regulation of MJO teleconnection. *Clim. Dyn.,* **37**, 1133–1149. https://doi.org/10.1007/s00382-010-0902-3

Newman, M. and coauthors, 2016: The Pacific Decadal Oscillation, Revisited, *J. Clim.*, **29**(12), 4399–4427 https://journals.ametsoc.org/view/journals/clim/29/12/jcli-d-15-0508.1.xml

Pfahl, S. 2014: Characterising the relationship between weather extremes in Europe and synoptic circulation features, *Nat. Hazards Earth Syst. Sci*., **14**, 1461–1475, https://doi.org/10.5194/nhess-14-1461-2014.

Phillips, A. S., C. Deser, J Fasullo, D. P. Schneider and I. R. Simpson, 2020: Assessing Climate Variability and Change in Model Large Ensembles: A User’s Guide to the “Climate Variability Diagnostics Package for Large Ensembles”, doi:10.5065/h7c7-f961.

Poli, P., H. Hersbach, D. P. Dee, P. Berrisford, A. J. Simmons, F. Vitart, P. Laloyaux, D. G. H. Tan, C. Peubey, J-N. Thépaut, Y. Trémolet, E.V. Hólm, M. Bonavita, L. Isaksen,

& M. Fisher, 2016: ERA-20C: An atmospheric reanalysis of the Twentieth Century, *J. Climate*, **29**, 11, pp. 4083-4097.

Rasch, P. J., Xie, S., Ma, P.-L., Lin, W., Wang, H., Tang, Q., et al., 2019: An overview of the atmospheric component of the Energy Exascale Earth System Model. *Journal of Advances in Modeling Earth Systems*, ***11***, https://doi.org/10.1029/ 2019MS001629

Rayner, N. A., D. E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Climate and Dynamics*, **108**, D14, **https://doi.org/10.1029/2002JD002670**.

Richter, J. H., Sassi, F., & Garcia, R. R. (2010). Toward a physically based gravity wave source parameterization in a general circulation model. *Journal of the Atmospheric Sciences*, **67**(1), 136–156. https://doi.org/10.1175/2009JAS3112.1.

Richter, J. H., Anstey, J. A., Butchart, N., Kawatani, Y., Meehl, G. A., Osprey, S., and Simpson, I. R., 2020: Progress in simulating the quasi-biennial oscillation in CMIP models. *Journal Geophysical Research: Atmospheres*, **125**, e2019JD032362. https://doi.org/ 10.1029/2019JD032362

Richter, J.H., Butchart, N., Kawatani, Y., Bushell, A.C., Holt, L., Serva, F., Anstey, J., Simpson, I.R., Osprey, S., Hamilton, K., Braesicke, P., Cagnazzo, C., Chen, C.-C., Garcia, R.R., Gray, L.J., Kerzenmacher, T., Lott, F., McLandress, C., Naoe, H., Scinocca, J., Stockdale, T.N., Versick, S., Watanabe, S., Yoshida, K. and Yukimoto, S., 2022: Response of the Quasi-Biennial Oscillation to a warming climate in global climate models. *Q.J.R. Meteorol. Soc*, **148**: 1490-1518. https://doi.org/10.1002/qj.3749

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

Rodgers, K. B., Lee, S. S., Rosenbloom, N., Timmermann, A., Danabasoglu, G., Deser, C., ... and Yeager, S. G., 2021: Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, **12**(4), 1393-1411, doi: 10.5194/esd-12-1393-2021.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., et al., 2013//; Berkeley Earth temperature averaging process. *Geoinformatics & Geostatistics: An Overview*, **01**(02). https://doi.org/10.4172/2327-4581.1000103.

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

Schiemann, R., Athanasiadis, P., Barriopedro, D., Doblas-Reyes, F., Lohmann, K., Roberts, M. J., Sein, D. V., Roberts, C. D., Terray, L., and Vidale, P. L., 2020: Northern Hemisphere blocking simulation in current climate models: evaluating progress from the Climate Model Intercomparison Project Phase 5 to 6 and sensitivity to resolution, *Weather Clim. Dynam*., **1**, 277–292, https://doi.org/10.5194/wcd-1-277-2020.

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

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

Son, S. W., Y. Lim, C. Yoo, H. H. Hendon, and J. Kim, 2017: Stratospheric control of the Madden-Julian Oscillation. *Journal of Climate*, **30**, 1909–1922, https://doi.org/10.1175/JCLI-D-16-0620.1.

Stevenson, S., X. Huang, Y. Zhao, E. Di Lorenzo, M. Newman, L. van Roekel, T. Xu, and A. Capotondi, Ensemble Spread Behavior in Coupled Climate Models: Insights from the Energy Exascale Earth System Model version 1 Large Ensemble, *J. Advances Modeling Earth Systems*, **15**, https://doi.org/10.1029/2023MS003653.

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

Tang, Y., and Yu, B., 2008: MJO and its relationship to ENSO, *J. Geophys. Res*., **113**, D14106, doi:10.1029/2007JD009230.

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

Thompson, D.W.J. & Wallace, J.M., 2000: Annular modes in the extratropical circulation. Part I: Month-to-month variability. *J. Climate*. **13**. 1000-1016. https://doi.org/10.1175/1520- 0442(2000)01360;1000:amitec62;2.0.co;2

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

Wang, J., H.-M. Kim, E. K. M. Chang, and S.-W. Son, 2018: Modulation of the MJO and North Pacific Storm Track Relationship by the QBO. *Journal of Geophysical Research: Atmospheres*, https://doi.org/ 10.1029/2017JD027977.

Wang, H., Easter, R. C., Zhang, R., Ma, P.-L., Singh, B., Zhang, K., et al. (2020). Aerosols in the E3SM version 1: New developments and their impacts on radiative forcing. *Journal Advances Modeling Earth Systems*, **12**(1), e2019MS001851. https://doi.org/10.1029/2019MS001851.

Xie, S.-P., and J. A. Carton, 2004: Tropical Atlantic variability: Patterns, mechanisms, and impacts. Earth's Climate: The Ocean–Atmosphere Interaction, *Geophys. Monogr*., **147**, 121–142, https://doi.org/10.1029/147gm07.

Yoo, C., and S. W. Son, 2016: Modulation of the boreal wintertime Madden-Julian oscillation by the stratospheric quasi-biennial oscillation. *Geophysical Research Letters*, **43** (3), 1392–1398, https://doi.org/ 10.1002/2016GL067762.

Zebiak, S, 1993: Air–sea interaction in the equatorial Atlantic region. *J. Clim*. **6**, 1567–1586, https://doi.org/10.1175/1520-0442(1993)006<1567:AIITEA>2.0.CO;2.

Zhang, C., 2005: Madden-Julian Oscillation, *Rev. Geophys*., **43**, RG2003, doi:10.1029/2004RG000158.