AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters®

RESEARCH LETTER
10.1029/2021GL095041

Robust Evaluation of ENSO in Climate Models: How Many Ensemble Members Are Needed?

Jiwoo Lee[1], Yann Y. Planton[2], Peter J. Gleckler[1], Kenneth R. Sperber[1], Eric Guilyardi[3,4], Andrew T. Wittenberg[5], Michael J. McPhaden[2], and Giuliana Pallotta[1]

[1]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA, USA, [2]NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA, [3]LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France, [4]NCAS-Climate, University of Reading, Reading, UK, [5]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Key Points:

- To estimate the ensemble size required to characterize the ENSO simulation, ensemble members of CMIP6 and Large Ensemble models are analyzed
- A broad range in the relative performance of models exists with internal variability influencing the robustness of some ENSO characteristics
- The required ensemble size depends on metric, duration of observational record, and model; the size can be a small as 6 or greater than 50

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Lee,
lee1043@llnl.gov

Citation:

Lee, J., Planton, Y. Y., Gleckler, P. J., Sperber, K. R., Guilyardi, E., Wittenberg, A. T., et al. (2021). Robust evaluation of ENSO in climate models: How many ensemble members are needed? *Geophysical Research Letters*, *48*, e2021GL095041. https://doi.org/10.1029/2021GL095041

Received 1 JUL 2021
Accepted 28 SEP 2021

Abstract Large ensembles of model simulations require considerable resources, and thus defining an appropriate ensemble size for a particular application is an important experimental design criterion. We estimate the ensemble size (N) needed to assess a model's ability to capture observed El Niño-Southern Oscillation (ENSO) behavior by utilizing the recently developed International CLIVAR ENSO Metrics Package. Using the larger ensembles available from CMIP6 and the US CLIVAR Large Ensemble Working Group, we find that larger ensembles are needed to robustly capture baseline ENSO characteristics ($N > 50$) and physical processes ($N > 50$) than the background climatology ($N \geq 12$) and remote ENSO teleconnections ($N \geq 6$). While these results vary somewhat across metrics and models, our study quantifies how larger ensembles are required to robustly evaluate simulated ENSO behavior, thereby providing some guidance for the design of model ensembles.

Plain Language Summary To account for uncertainties arising from the chaotic nature of the climate system, Earth system models are often used to generate a large number of simulations under slightly different initial conditions. These large ensembles enable the consistency between models and observations to be addressed while accounting for the internal variability in the climate system. Creating a set of ensemble simulations requires substantial resources, and so in this study we diagnose what ensemble size is sufficient to robustly represent the simulated behavior of the El Niño/Southern Oscillation (ENSO), one of the most important modes of variability affecting climate worldwide.

1. Introduction

The El Nino Southern Oscillation (ENSO) is the dominant mode of tropical variability with far-reaching climatic and societal impacts (Clarke, 2008; McPhaden et al., 2006, 2020; Ropelewski & Halpert, 1987). ENSO generates large-scale sea surface temperature (SST) variations in the eastern equatorial Pacific Ocean, with SST anomalies typically between 1°C and 3°C, accompanied by changes in the oceanic thermal structure and currents, and in the atmospheric circulation and convective activity. General circulation models (GCMs) have striven to capture key observed characteristics of ENSO as documented by many previous studies (e.g., AchutaRao & Sperber, 2002; Guilyardi et al., 2020; Ham & Kug, 2014).

Evaluating GCMs against observations is essential to identify strengths and weaknesses of different models for different applications, and to track model improvements during model development and across generations of the Coupled Model Intercomparison Project (CMIP). For example, AchutaRao and Sperber (2006) compared the ENSO performance of the CMIP2 and CMIP3 models, and found improvements in representing the spatial patterns of the SST anomalies in the eastern Pacific. Later, Bellenger et al. (2014) examined the ability of the CMIP3 and CMIP5 models to simulate the tropical Pacific climatology and ENSO, and found reduced intermodel spread in ENSO amplitudes and improved ENSO lifecycles in CMIP5 relative to CMIP3. Such model improvements are key for improving forecasts and projections of future ENSO risks (Ding et al., 2020; Guilyardi et al., 2020; L'Heureux et al., 2020; Stevenson et al., 2021). Fasullo et al. (2020) and Fasullo (2020) examined ENSO-related SST variability in CMIP3, 5, 6 and large ensembles, in which

the improvement in the latest CMIP6 compared to its earlier phases is shown with the role of internal variability quantified.

The International CLIVAR Research Focus on ENSO in a Changing Climate, together with the CLIVAR Pacific Region Panel and a team of community experts on ENSO, recently developed a suite of performance metrics to evaluate ENSO simulations, and applied these metrics to the CMIP5 and CMIP6 models (Planton et al., 2021). They raised a point that the climate model evaluation depends on three aspects to focus on: (a) background climatology and basic ENSO characteristics, (b) ENSO's worldwide teleconnections, and (c) ENSO's internal processes and feedbacks represented in Historical simulations of GCMs. However, in a multi-model ensemble, it can be difficult to separate the role of internal variability versus model formulation (different physical parameterizations, resolutions, dynamical cores, representations of fluxes between ocean and atmosphere, etc.) in generating intermodel spread in the ENSO performance metrics.

To resolve this uncertainty, one can leverage ensembles of simulations from individual models to test the sensitivity of the ENSO metrics to internal variability alone. While most contributing modeling groups typically provide fewer than 10 Historical simulations (exploring different initial conditions, initialization procedures, physical parameterizations or forcings) to CMIP, some have produced 30 or more (e.g., Boucher et al., 2020; Delworth et al., 2020; Deser et al., 2020). These large ensembles offer a valuable testbed to determine the ensemble size needed to measure model performance relative to a specific skill, especially when evaluating climate variability (Deser et al., 2020). In particular, multimillennium simulations have demonstrated that ENSO's characteristics (amplitude, spectrum, irregularity, and spatial pattern) can vary substantially on multidecadal and multicentennial scales, purely due to internal variability (Stevenson et al., 2010; Wittenberg, 2009; Wittenberg et al., 2014). It is thus essential to account for this internal variability when evaluating or intercomparing models, by using a sufficient run duration under stable climate (e.g., control simulations of CMIP) and/or having an adequate ensemble size (especially when external forcings vary in time as in the CMIP Historical simulations) to robustly resolve any important differences.

Generating a large ensemble of simulations requires considerable resources, and so defining an appropriate ensemble size for a particular application has been recognized as an important step in the experimental design of both weather and climate simulations for decades (e.g., Leith, 1974). As the appropriate ensemble size is application-dependent (e.g., Branković & Palmer, 1997; Déqué, 1997; Doi et al., 2019; Pennell & Reichler, 2011; Wills et al., 2020) and likely model-dependent, CMIP has not yet defined a standard ensemble size or a standard methodology to determine the minimum ensemble size. For ENSO in GCM, Bulić and Branković (2007) concluded that a 35-member atmospheric GCM large ensemble enabled "better sampling and detection of the ENSO signal in the extratropics where atmospheric internal variability is relatively strong." Maher et al. (2018) investigated the ENSO amplitudes in two large ensembles, and argued that approximately 30–40 ensemble members from a given model were needed to robustly characterize ENSO-related SST variability in the Niño regions. Milinski et al. (2020) found that 50 members were needed to characterize winter variability in the Niño3.4 region to within ±5% error. However, gauging the ensemble size needed to robustly characterize a broad range of ENSO characteristics has not been thoroughly investigated.

In this study, we apply the CLIVAR ENSO Metrics Package metrics to all available Historical ensemble members of the CMIP6, to assess the robustness of model skill and address the following question: *What is the minimum number of ensemble members needed to obtain robust results for characterizing ENSO performance in GCMs?* We examine the models' ability to capture the elements of the background climatology relevant to ENSO, the emergent tropical Pacific behavior of ENSO, ENSO's remote teleconnections outside the tropical Pacific, and key ENSO processes and feedbacks, by applying the CLIVAR ENSO Metrics Package (Planton et al., 2021).

2. Data and Methods

We use all currently available simulations from the most recent generation of the Coupled Model Intercomparison Project (CMIP6) and several large ensembles made available by a few modeling groups. The CMIP6 coupled Historical experimental protocol (Eyring et al., 2016) is well-suited for evaluating the ENSO simulations against observations. The Historical simulations are initialized in 1,850 and run to 2,014 with close to observed time-varying natural and anthropogenic forcings (Durack et al., 2018). We use all available Historical members from 58 CMIP6 models obtained through Earth System Grid Federation (ESGF; Wil-

Table 1
List of Models That Provided Historical Simulations (1850–2014) to CMIP6 or Multi-Model Large Ensemble Archive (MMLEA), and Their Ensemble Sizes

Participation	Model	Members	Model	Members
CMIP6	ACCESS-CM2	3	GFDL-CM4	1
	ACCESS-ESM1-5	**30**[a]	GFDL-ESM4	3
	AWI-CM-1-1-MR	5	GISS-E2-1-G	47[c]
	AWI-ESM-1-1-LR	1	GISS-E2-1-G-CC	1
	BCC-CSM2-MR	3	GISS-E2-1-H	25[c]
	BCC-ESM1	3	HadGEM3-GC31-LL	5
	CAMS-CSM1-0	3	HadGEM3-GC31-MM	4
	CanESM5	**(25, 40)**[b]	INM-CM4-8	1
	CanESM5-CanOE	3	INM-CM5-0	10
	CESM2	11	IPSL-CM5A2-INCA	1
	CESM2-FV2	3	**IPSL-CM6A-LR**	**32**[a]
	CESM2-WACCM	3	IPSL-CM6A-LR-INCA	1
	CESM2-WACCM-FV2	3	KACE-1-0-G	3
	CMCC-CM2-HR4	1	KIOST-ESM	1
	CMCC-CM2-SR5	1	MIROC-ES2H	3
	CMCC-ESM2	1	**MIROC-ES2L**	**31**[a]
	CNRM-CM6-1	**29**[a]	**MIROC6**	**50**[a]
	CNRM-CM6-1-HR	1	MPI-ESM-1-2-HAM	3
	CNRM-ESM2-1	10	MPI-ESM1-2-HR	10
	E3SM-1-0	5	MPI-ESM1-2-LR	10
	E3SM-1-1	1	MRI-ESM2-0	7
	EC-Earth3	**22**[a]	NESM3	5
	EC-Earth3-AerChem	2	**NorCPM1**	**30**[a]
	EC-Earth3-CC	1	NorESM2-LM	3
	EC-Earth3-Veg	9	NorESM2-MM	3
	EC-Earth3-Veg-LR	3	SAM0-UNICON	1
	FGOALS-f3-L	3	TaiESM1	2
	FGOALS-g3	6	UKESM1-0-LL	19
	FIO-ESM-2-0	3		
MMLEA	**CESM1-CAM5**	**40**[a]	**CanESM2**	**50**[a]

Note. Models having 20 or more *initial condition* ensemble members (i.e., varying initial condition but fixed physical parameterizations) are marked in **bold** and with ([a]) and used for determining the required ensemble size in Section 3.2. Models marked with a hash ([c]) are excluded despite having 20 or more members because of varying physical parameterizations. CMIP6 models that are available as of June 2021 are applied in this study. Further information on each CMIP6 Model is available at https://es-doc.org/cmip6/.

[a]Models having 20 or more "initial condition" ensemble members with varying initial conditions but fixed physical parameterizations. [b]CanESM5 has 25 and 40 initial condition ensembles under different physical parameterization configurations (*p1* and *p2*, respectively), which are considered as different sets of ensembles in this study. [c]Models having 20 or more members but less than 20 *initial condition ensemble members* because the ensemble was composed by varying physical parameterizations.

liams et al., 2016) and 2 Single-Model Initial condition Large Ensemble (SMILE) models obtained through the Multi-Model Large Ensemble Archive (MMLEA) collected by the US CLIVAR Large Ensemble Working Group (Deser et al., 2020). Models used in this study are listed in Table 1. We note that the two SMILE ensembles for CESM1-CAM5 and CanESM2 have been conducted using CMIP5 forcing. It is possible that

differences in CMIP5 and CMIP6 forcings may have an impact on some of the CLIVAR ENSO Metrics, however, we chose to include them to make use of as many large ensembles as possible.

To gauge how well models simulate the observed characteristics of ENSO, we apply the CLIVAR ENSO Metrics Package (hereafter CEM2021; Planton et al., 2021) to examine intermodel and intermember spread of the metrics results. The metrics in CEM2021 are divided into three Metrics Collections: *Performance* (i.e., background climatology and basic ENSO characteristics), *Teleconnections* (ENSO's worldwide teleconnections), and *Processes* (ENSO's internal processes and feedbacks). Each metric is computed using monthly mean simulated and observed fields. We use the same observations as in Planton et al. (2021), including AVISO, ERA-Interim (Dee et al., 2011), GPCPv2.3 (Adler et al., 2003), and TropFlux (Praveen Kumar et al., 2012, 2013), and refer to these as our reference data sets (list of variables and epochs are provided in supplement, as Table S1 in Supporting Information S1). In addition, the following observation-based reference data sets were used as alternatives to measure the range of observational discrepancies: The Twentieth Century Reanalysis (20CR, Compo et al., 2011), Clouds and the Earth's Radiant Energy System Energy Balanced and Filled (CERES-EBAF, Kato et al., 2018), CPC Merged Analysis of Precipitation (CMAP, Xie & Arkin, 1997), ERA-20C (Poli et al., 2016), ECMWF Reanalysis v5 (ERA-5, Hersbach et al., 2020), Hadley Center Sea Ice and Sea Surface Temperature data set (HadISST, Rayner et al., 2003), TRMM-3B43v7 (Huffman et al., 2007), and Optimum Interpolation SST (OISST, Huang et al., 2021). The analysis is conducted using the PCMDI Metrics Package (PMP, Gleckler et al., 2016) framework in which the CEM2021 is implemented. Extending the study of Planton et al. (2021), in which the CEM2021 metrics were applied to CMIP6 simulations using one ensemble member per model and ensemble members of a selected model, in this study we apply the CEM2021 metrics to all available ensemble members of all available CMIP6 models and additional SMILE models to assess the robustness of model skill.

To robustly evaluate a model, we need a large enough ensemble size to indicate whether the model could have plausibly simulated the observed realization, which however can be challenging when the metric is strongly modulated, when the model or its forcings are bad, or when the observations are uncharacteristic of the true long-term behavior of nature (e.g., too short observation period)—not all of which may be known in advance. To estimate the "large enough" (or "at minimum") ensemble size needed to gauge ENSO performance, we apply a Monte-Carlo approach as proposed by Milinski et al. (2020). We apply CEM2021 results from models with large ensembles (LEs) of 20 or more members (with varying initial conditions, but fixed physical parameterizations, and forcings), to capture the ensemble spread caused by internal variability. The LEs include ACCESS-ESM1-5 (Ziehn et al., 2020), CanESM5 (Swart et al., 2019), CNRM-CM6-1 (Voldoire et al., 2019), EC-Earth3 (Döscher et al., 2021), IPSL-CM6-LR (Boucher et al., 2020), MIROC-ES2L (Hajima et al., 2020), MIROC6 (Tatebe et al., 2019), and NorCPM1 (Bethke et al., 2021) of CMIP6, as well as CESM (Kay et al., 2015) and CanESM2 (Kirchmeier-Young et al., 2017) of the SMILE models (models marked with asterisk in Table 1). For each LE model and metric, a random sample of N members (pseudoensemble or *PE*), with N ranging from 1 to the full ensemble size, is drawn from the ensemble. We generate 1,000 PEs to estimate the sampling distribution for each metric and model, resampling "with replacement" (each PE member is drawn from the full ensemble each time, thus independent to previous draws) or "without replacement" (each new member is drawn only from members not previously selected for that PE). We consider a PE of size N sufficient if at least 95% of the resampled PE means from the "with replacement" are within ±10% of the "true" ensemble mean estimated from the full ensemble. Additional details are provided in Supporting Information S1.

3. Results

3.1. Performance Overview

Figure 1 provides a quick-look summary of CMIP6 results using a *portrait plot* (Gleckler et al., 2008) for each of the three metrics collections defined as part of the CEM2021. This figure resembles Figure 1 of Planton et al. (2021), except here we include multiple members from individual CMIP6 models, to assess the level of variation arising from internal climate variability. Objectively summarizing results across all metrics is achieved via a common normalization, to ensure that results from each metric span a similar range. Simple normalizations like the one we use, calculated relative to the multi-model mean error (MMME) for each metric, are well-established and have been applied in analogous figures for the mean

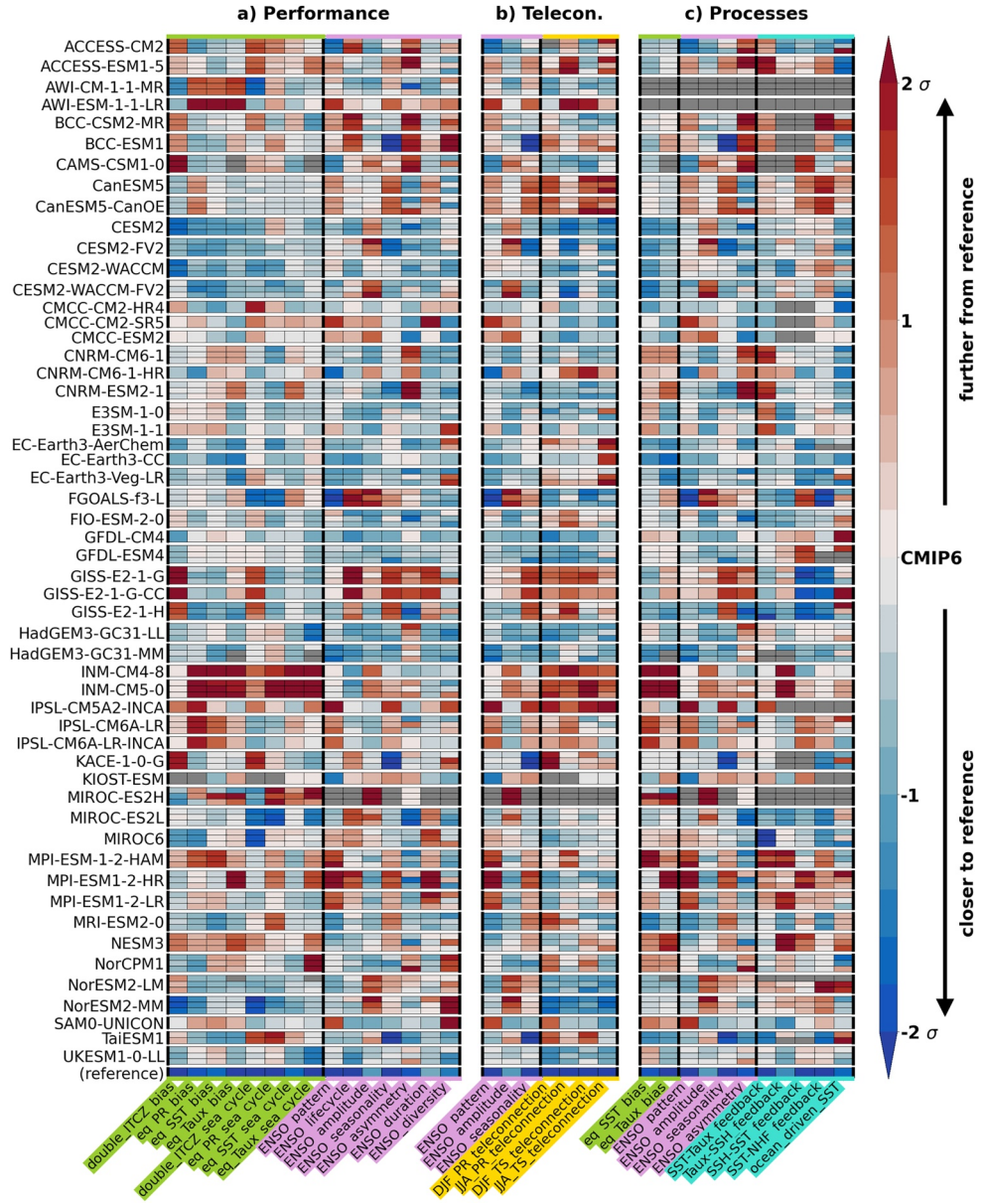

Figure 1. El Niño-Southern Oscillation (ENSO)metrics *portrait plot* for Coupled Model Intercomparison Project (CMIP6) with results for *Performance*, *Teleconnection*, and *Processes* Metrics Collections defined as part of CEM2021 (Planton et al., 2021). Multiple realizations are shown as available, with a maximum of three per model for brevity. The initial error metrics are positive-definite measures of distance from the reference observations (e.g., root-mean-square error or percent absolute error), for a given physical field of interest (see Table B1 of Planton et al., 2021 for definitions). To aid comparison across models and metrics, the metrics are displayed nondimensionally, as a difference from the multi-model mean error (MMME) computed from all CMIP6 divided by the intermodel standard deviation (σ) within each metric column. A displayed value of 0 (bright color) corresponds to the MMME; a value of 2 (dark red) corresponds to a model error two standard deviations greater (worse) than the MMME; and a value of −2 (dark blue) a model error that is two standard deviations less (better) than the MMME. Missing metric (for which the calculation was not available due to unpublished variable(s) or technical issues in the data set) is indicated in gray. To weight the models equally in the MMME, the error metrics of each model are first averaged across its own ensemble members before averaging across all models. Metrics are grouped according to their application (Metrics Collection or MC), while individual metrics highlighted using color-codes according to its category: evaluating background climatology (light green), basic ENSO characteristics (magenta), teleconnections (yellow), or physical processes (cyan).

climate (Flato et al., 2014; Gleckler et al., 2008), indices of temperature and precipitation extremes (Kim et al., 2020; Sillmann et al., 2013), extratropical modes of variability (Lee et al., 2019, 2021), and ENSO (Bellenger et al., 2014; Planton et al., 2021). The color scale in Figure 1 (±2 standard deviation from the MMME in each column) is expressed relative to the range of errors in the CMIP6 multi-model ensemble. Figure 1 thus highlights the strengths and weaknesses of each model relative to the multi-model distribution. For most models the relative performance is mixed across the metrics, including smaller (blue) and larger (red) errors relative to the MMME. Figure 1 indicates that the members for a given model and metric generally have similar errors relative to the multi-model distribution, suggesting that each model's relative performance is fairly insensitive to internal variability. There are exceptions, however, for some of the ENSO performance metrics (lifecycle, amplitude, asymmetry, and diversity), and feedback metrics (in particular the ocean-driven SST tendency), which show substantial spread due to internal variability when assessed over the epochs of the reference data sets.

Figure 2 is based on the same statistics used in Figure 1, but without normalization. The circles in each panel represent the average error across all members as compared to our reference data set, with vertical line markers showing the results for individual members. These plots collectively illustrate the intermodel skill differences, as well as the intermember (internal) variability in the errors for each model, for those selected three example metrics (analysis for other metrics are available in Figure S1 in Supporting Information S1). For the *Equatorial SST Bias* metric (Figure 2a), as well as others based on mean state characteristics (Figure S1 in Supporting Information S1), the intermember spread due to internal variability is very narrow. The internally generated spread is larger for *ENSO Amplitude* (Figure 2b), as large as 1σ of intermodel spread in general. For *ENSO Asymmetry* (Figure 2c), there are some members that nearly match the observations while others differ strongly from observations (e.g., CanESM2). For metrics with such behavior, multiple members are needed to obtain an accurate assessment of skill relative to observations. It is worth noting that for most of the metrics the spread among observational products is smaller than 1 standard deviation of intermodel spread (Figure 2 and Figure S1 in Supporting Information S1). However, there are a few cases that the observational spread exceeds 1 standard deviation of intermodel spread (e.g., *ENSO seasonality*, *diversity*, and *teleconnection* metrics as shown in Figure S1 in Supporting Information S1), in which our confidence in the reliability of the results is limited. Figure 2 also shows that the intermember spread is model-dependent. While it is important to examine whether the model reproduces the realistic aspects of the internal variability, it is however challenging because of the short observational record. Further analysis is required to better understand the role of observational uncertainty and the model-dependent internal variability.

3.2. Estimating the Required Ensemble Size

We now estimate how many members are needed for each metric, to ensure that the results are reasonably representative of any given model's overall performance. We use results from the 10 models contributed to CMIP6 or SMILEs that have 20 or more ensemble members with varying initial conditions but fixed physical parameterizations, thus focusing on the ensemble spread caused by internal variability. These models are ACCESS-ESM1-5, CanESM5, CNRM-CM6-1, EC-Earth3, IPSL-CM6-LR, MIROC-ES2L, MIROC6, and NorCPM1 of CMIP6, and CESM and CanESM2 of the SMILEs (Table 1).

Figure 3 depicts the distribution of sampling errors for IPSL-CM6A-LR as a function of ensemble size (N). Results are shown for metrics that vary little from one member to another (*Equatorial SST Bias*), moderately (*ENSO Amplitude*) and substantially (*ENSO Asymmetry*) relative to other metrics, for an epoch of the length of the reference data set. The pseudoensemble means from the "without replacement" sampling results converges to the full ensemble mean. On the contrary, pseudoensemble means from the "with replacement" sampling does not converge to the mean when the entire sample size is considered, which approximates what would happen if the samples had been drawn from the underlying infinite-member distribution. We define our estimate of a minimum ensemble size needed to resolve differences in skill between the models, N_{min}, as the smallest value of n (i.e., number of samples in subset) where at least 95% of the "with replacement" pseudoensemble means fall within 10% of the mean of the full ensemble. The N_{min} is estimated to be 1 for *Equatorial SST Bias* (Figure 2a), and 8 for *ENSO Amplitude* (Figure 2b), while entire ensemble size (32) is not large enough for *ENSO Asymmetry* (Figure 2c), for the IPSL-CM6A-LR model and for the epoch lengths of the reference data set.

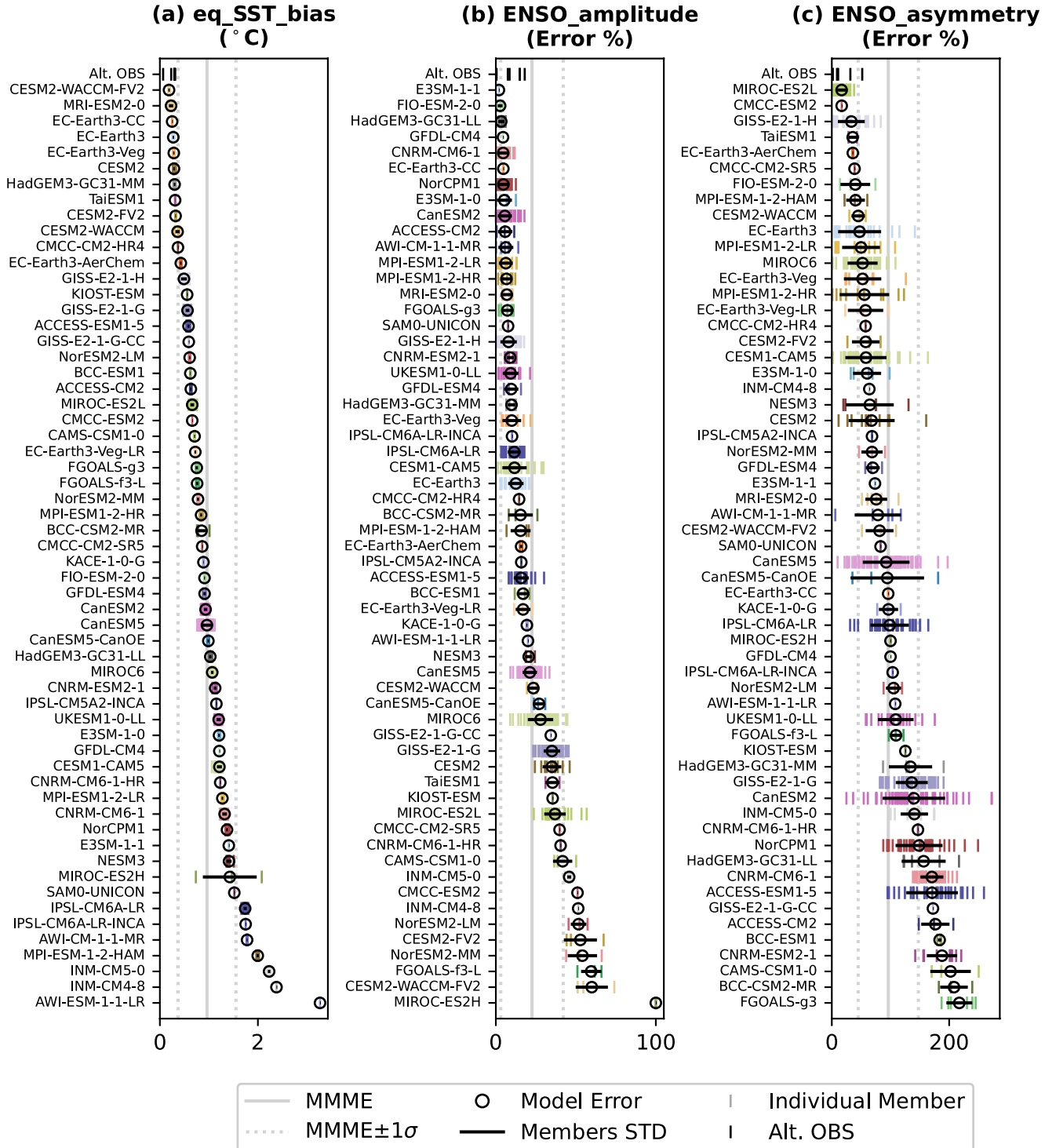

Figure 2. Error metrics calculated for ensemble simulations from the Coupled Model Intercomparison Project (CMIP6) models and large ensembles of CESM1-CAM5 and CanESM2. Lines represent standard deviation of error metrics for individual model ensembles, with circles denoting the average of all members for any given model. Three representative metrics are shown: (a) Equatorial sea surface temperature (SST) Bias, (b) El Niño-Southern Oscillation ENSO Amplitude, and (c) Asymmetry, with results from other metrics in Figure S1 in Supporting Information S1. In each panel, a corresponding unit is given in the subtitle. Models are sorted by their metric values (smaller metric value for better performance). Vertical solid and dotted lines in light gray are for multi-model mean error and its ±1 standard deviation, respectively. Error metrics calculated for alternative observation-based data sets (Alt. OBS) are shown at the top row of each panel.

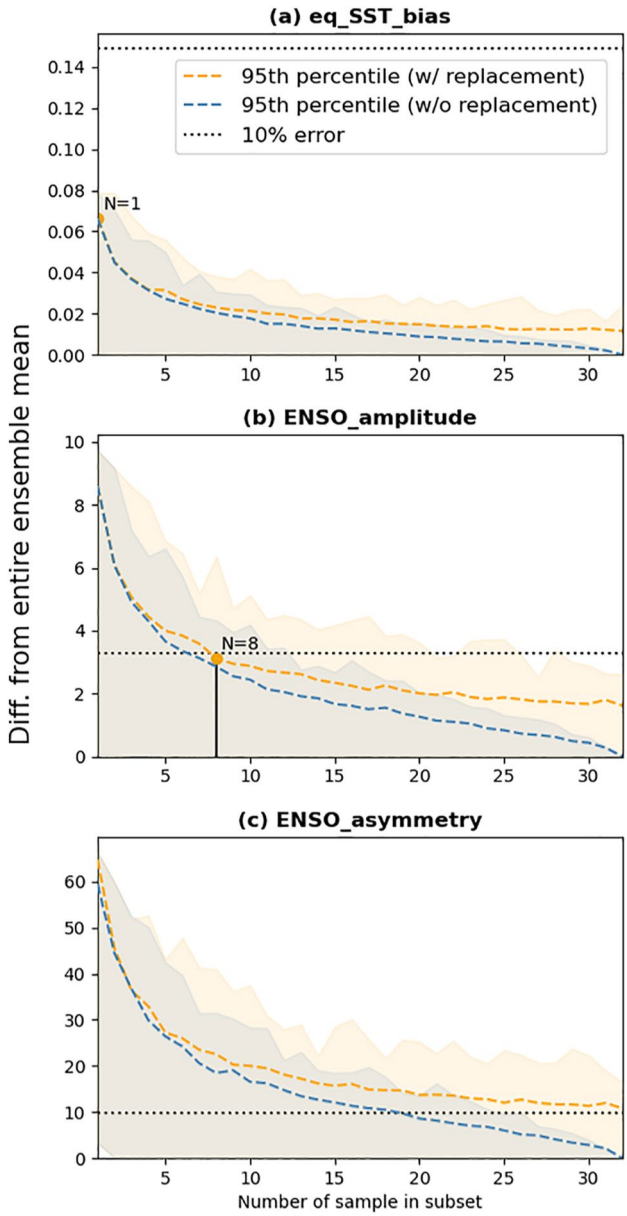

Figure 3. Absolute difference of the sample mean from the actual mean of the entire IPSL-CM6A-LR ensemble (ordinate) for pseudoensembles sampled with (orange) or without (blue) replacement at different sample sizes (abscissa). Three representative metrics are shown: (a) Equatorial sea surface temperature (SST) Bias, (b) El Niño-Southern Oscillation (ENSO) Amplitude, and (c) Asymmetry. Annotated N indicates the minimum ensemble size (N_{min}) for which at least 95% of the "with replacement" pseudoensemble means fall within 10% of the mean metric value from the full ensemble. Shaded area indicates the full min-max range of the sample distribution, long-dashed lines indicate 95th percentiles of the sample distribution, and short-dashed horizontal lines indicate a difference of 10% from the mean of the full ensemble. Note that by definition the distribution of the pseudoensemble without replacement (blue) converges toward the mean of the full ensemble.

We repeated the aforementioned analysis to estimate the N_{min} for individual metrics, and from the 10 ensemble models listed above (i.e., models highlighted in Table 1). The height of each bar in Figure 4 shows the maximum N_{min} for each metric, selected conservatively as the largest value of N_{min} among the 10 models. As anticipated, the background climatology metrics (light green) and teleconnection metrics (yellow) require smaller ensembles (1–12 members and 1–6 members, respectively) than metrics evaluating basic ENSO characteristics (magenta, 17–50 members). Note that in the CEM2021 the teleconnection metrics measure the skill on global spatial pattern, while if a metric targets regional analysis then it may show larger spread (e.g., AchutaRao & Sperber, 2006). The *ENSO Asymmetry, Duration* and *Diversity* metrics require the largest N_{min}, 50. For the metrics evaluating physical processes (cyan), the N_{min} varies across from 3 to 50. The two metrics requiring an $N_{min} > 50$ include the *SST-Taux Feedback* metric, which examines the sensitivity of sea surface temperature anomalies in the eastern equatorial Pacific to zonal wind stress anomalies in the western equatorial Pacific, and the *Ocean-driven SST* metric, which gauges how much anomalous heating by local ocean advection and mixing is associated with a 1 K change in SST in the eastern equatorial Pacific Niño3 region (5°N–5°S, 150°–90°W).

4. Summary and Discussion

We applied the CLIVAR ENSO Metrics Package (CEM2021; Planton et al., 2021) to all available ensemble members of the models in the CMIP6 Historical experiment database plus two additional large ensembles with CMIP5 forcing. By using several ensembles exceeded 20 members (ACCESS-ESM1-5, CanESM5, CNRM-CM6-1, EC-Earth3, IPSL-CM6-LR, MIROC-ES2L, MIROC6, and NorCPM1 from CMIP6, and CESM and CanESM2 from MMLEA), we estimated the minimum number of members needed to diagnose how well climate models simulate a diverse suite of ENSO characteristics. We find that the results vary across metrics and are somewhat model-dependent. Models require a larger ensemble to constrain baseline ENSO characteristics ($N > 50$) and physical processes ($N > 50$) than they do for the background climatology ($N \geq 12$) and ENSO-related teleconnections ($N \geq 6$). We have shown how estimates of an N_{min} can vary from one model to the next, and thus we encourage future investigators to apply the same tests to other large ensembles as they become available. With the approach we have applied, however, the minimum effective ensemble size is constrained by the size of the full ensemble (i.e., N_{min} cannot exceed the size of the largest ensemble) and can be biased low if the available ensemble size is too small. Nonetheless, where gauging the simulation of ENSO may be of interest, we recommend these estimates be considered in the design of new coordinated experiments, including the Historical simulations in the next phase of CMIP. Early studies of how climate change affects future ENSO amplitude were based on CMIP model simulations with far fewer than 10 ensemble members, and often only one (e.g., Collins et al., 2010; Meehl et al., 2007; van Oldenborgh et al., 2005). In contrast, for the CMIP6 Historical simulations a minimum of three ensemble members was encouraged (Eyring et al., 2016). Further increasing ensemble size in future MIPs could help to further strengthen the robustness of these comparisons.

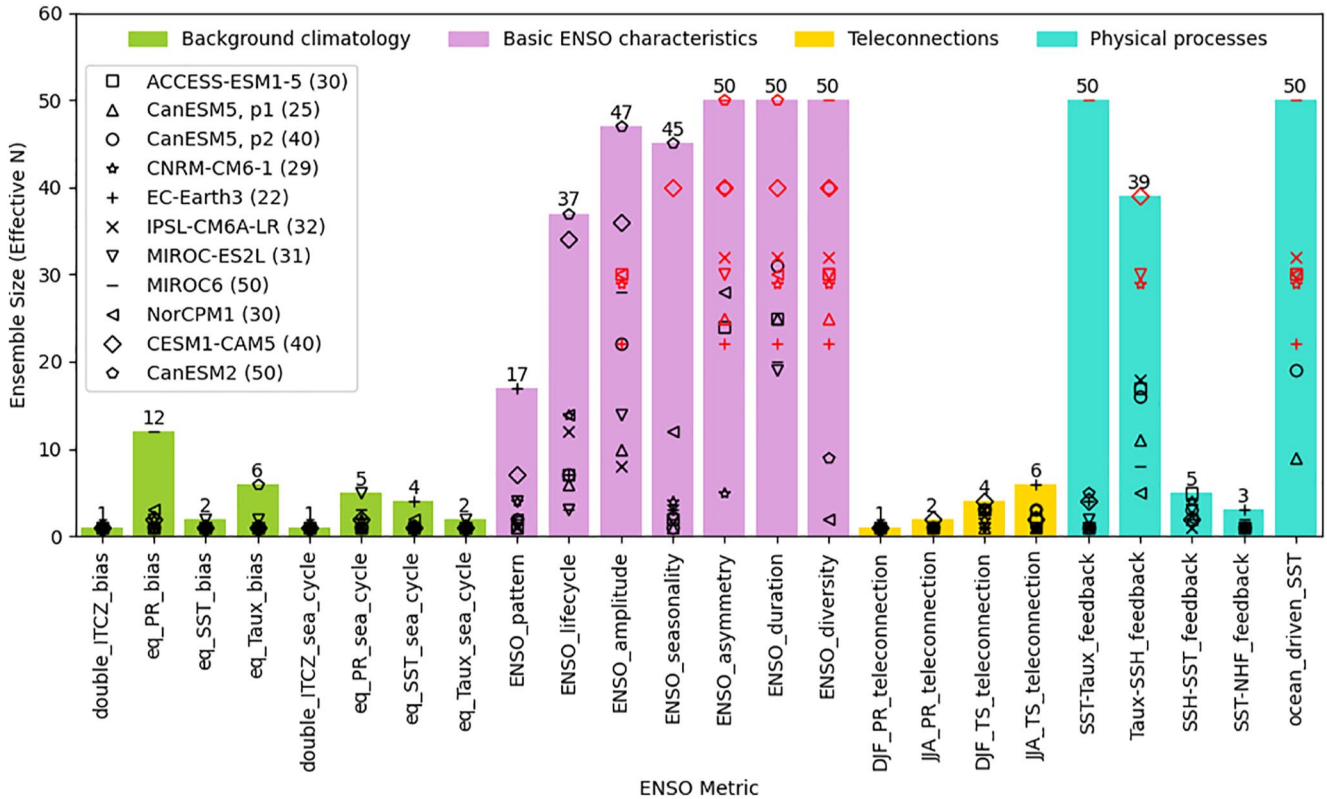

Figure 4. Minimum number of required ensemble members (N_{min} of Figure 3) for individual metrics obtained from models with at least 20 initial condition ensemble members (see Table 1). Each vertical bar indicates the maximum N_{min} (across the 11 ensembles from 10 models) for the given metric. Metrics are listed along the abscissa. Ordinate indicates the *minimum required* ensemble size for 95% of the ensemble means (so estimated) to fall within 10% of the actual mean of the full ensemble, as shown in Figure 3. Markers in red indicate cases where N_{min} exceeds the full ensemble size. Metrics are color coded as in Figure 1, for the background climatology (light green), basic El Niño-Southern Oscillation (ENSO) characteristics (magenta), teleconnections (yellow), and physical processes (cyan). In the legend box at upper right of the plot, models used are listed with their ensemble sizes in parentheses.

It is clear that improvement of ENSO in models is not an easy task. The diverse range of model performance within each of the process metrics is indicative of the complex nature of the model biases, and the tolerance level will depend on application and the signal-to-noise ratio (i.e., how large of a difference matters for a given metric). The requirement for robustness also depends on the metric and ultimately the science question being asked. The CEM2021 grouped metrics into three *Metrics Collections* (MCs) to address the following three science questions identified in Planton et al. (2021): (a) "How well are background climatology and basic ENSO characteristics simulated in Historical simulations?," (b) "How well are ENSO's global teleconnections represented in Historical simulations?," and (c) "How well are ENSO's internal processes and feedbacks represented in Historical simulations?." In each *Metrics Collection* some of the baseline ENSO characteristics are included (as shown in Figure 1 that all MCs include metrics highlighted in magenta color), and our conservative approach therefore suggests that to fully address each question requires a substantial ensemble size ($N > 50$ for *Performance* and *Process*, $N \geq 47$ for *Teleconnection Metrics Collections* of CEM2021), reinforcing the importance of the large ensembles.

Having found a substantial range in N_{min} across models for some metrics, a next step could involve more targeted guidance, perhaps estimating a suitable ensemble size on a per model basis *before* a large ensemble is generated. This might be possible by examining characteristics of a sufficiently long control run, treating nonoverlapping segments as proxies for Historical simulations. Further research quantifying the impact of time-varying external forcings (lacking in control runs) on the CEM2021 should help determine to what extent control runs can be used as a guide for this purpose. The approach and methodology used in this study can also be applied to estimate ensemble sizes for robustly evaluating other characteristics simulated

by Earth System Models, including those associated with internal variability, such as extratropical modes of variability (e.g., Fasullo et al., 2020; Lee et al., 2019, 2021; Orbe et al., 2020).

It must also be kept in mind that multiple century-long control runs span a more diverse set of ENSO regimes than sampled in the limited record length of available observations (Wittenberg, 2009). For these diverse regimes, it is entirely likely that different balances of processes are in effect (e.g., Atwood et al., 2017; Chen et al., 2017). One possible avenue of evaluation is to subsample the simulated ENSO variability over the observed epoch, as a basis for more rigorous assessment for GCMs. Since ENSO is modulated on multidecadal and longer time scales, high-quality observational records and reanalyses for the tropical Pacific must be sustained to support help improve understanding of longer time scale changes in the behavior of ENSO and its evaluation in climate models (Cravatte et al., 2016; Kessler et al., 2019). Inclusion of more regionally based metrics may also influence the assessment of model performance (e.g., AchutaRao & Sperber, 2006; Cai et al., 2018). Further work is also needed to establish how the selection of reference data may influence any conclusions derived from the CEM2021.

Data Availability Statement

The CMIP6 simulations were obtained through the ESGF (https://esgf-node.llnl.gov/projects/cmip6/) and the Large ensemble simulations of CESM1-CAM5 and CanESM2 models were obtained through the Multi-Model Large Ensemble Archive (https://www.cesm.ucar.edu/projects/community-projects/MMLEA). Observation-based reference data sets are available at their providers' websites: AVISO (https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products.html), CERES-EBAF (https://ceres.larc.nasa.gov/data/), CMAP (https://www.cpc.ncep.noaa.gov/products/global_precip/html/wpage.cmap.html), ERA-Interim, ERA-20C, ERA-5 (https://www.ecmwf.int/en/forecasts/datasets), GPCPv2.3 (https://psl.noaa.gov/data/gridded/data.gpcp.html), HadISST (https://www.metoffice.gov.uk/hadobs/hadisst/), OISST (https://www.ncei.noaa.gov/products/optimum-interpolation-sst), TRMM-3B43v7 (https://disc.gsfc.nasa.gov/datasets/TRMM_3B43_7/summary), TropFlux (https://incois.gov.in/tropflux/), and 20CR (https://psl.noaa.gov/data/20thC_Rean/).

Acknowledgments
Work of LLNL-affiliated authors was performed under the auspices of the U.S. Department of Energy (DOE) by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 via the Regional and Global Climate Modeling Program. The authors thank the CLIVAR Pacific Panel members and ENSO experts for a number of fruitful discussions. The authors acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the output and providing access, and the multiple funding agencies who support CMIP and ESGF. The U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) provides coordinating support and led development of software infrastructure for CMIP. The authors thank the US CLIVAR Working Group on Large Ensembles for making the Multi-Model Large Ensemble Archive (MMLEA) publically available. The altimeter products were produced by Ssalto/Duacs and distributed by Aviso+, with support from Cnes. ERA-Interim data are provided by ECMWF. The TropFlux data are produced under a collaboration between Laboratoire d'Océanographie: Expérimentation et Approches Numériques (LOCEAN) from Institut Pierre Simon Laplace (IPSL; Paris, France) and National Institute of Oceanography/CSIR (NIO; Goa, India), and supported by Institut de Recherche pour le Développement (IRD; France). TropFlux relies on data provided by the ECMWF interim reanalysis (ERA-Interim) and ISCCP projects. Support for the Twentieth Century Reanalysis Project version 3 data set is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER), by the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Physical Sciences Laboratory. This is PMEL Contribution No. 5276. The authors acknowledge the support from the Agence Nationale de la Recherche ARISE project, under Grant ANR-18-CE01-0012, and the Belmont project GOTHAM, under Grant ANR-15-JCLI-0004-01, the European Commission's H2020 Programme "Infrastructure for the European Network for Earth System Modelling Phase 3 (IS-ENES3)" project under Grant Agreement 824084.

References

AchutaRao, K., & Sperber, K. R. (2002). Simulation of the El Niño southern oscillation: Results from the coupled model intercomparison project. *Climate Dynamics*, *19*, 191–209. https://doi.org/10.1007/s00382-001-0221-9

AchutaRao, K., & Sperber, K. R. (2006). ENSO simulation in coupled ocean-atmosphere models: Are the current models better? *Climate Dynamics*, *27*, 1–15. https://doi.org/10.1007/s00382-006-0119-7

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P. P., Janowiak, J., et al. (2003). The Version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *Journal of Hydrometeorology*, *4*(6), 1147–1167. https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2

Atwood, A. R., Battisti, D. S., Wittenberg, A. T., Roberts, W. H., & Vimont, D. J. (2017). Characterizing unforced multi-decadal variability of ENSO: A case study with the GFDL CM2.1 coupled GCM. *Climate Dynamics*, *49*(7), 2845–2862. https://doi.org/10.1007/s00382-016-3477-9

Bellenger, H., Guilyardi, É., Leloup, J., Lengaigne, M., & Vialard, J. (2014). ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dynamics*, *42*(7), 1999–2018. https://doi.org/10.1007/s00382-013-1783-z

Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., et al. (2021). NorCPM1 and its contribution to CMIP6 DCPP. *Geoscientific Model Development Discussions*, 1–84. https://doi.org/10.5194/gmd-2021-91

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS002010. https://doi.org/10.1029/2019MS002010

Branković, Č., & Palmer, T. N. (1997). Atmospheric seasonal predictability and estimates of ensemble size. *Monthly Weather Review*, *125*, 859–874. https://doi.org/10.1175/1520-0493(1997)125<0859:ASPAEO>2.0.CO;2

Bulić, I. H., & Branković, Č. (2007). ENSO forcing of the Northern Hemisphere climate in a large ensemble of model simulations based on a very long SST record. *Climate Dynamics*, *28*(2–3), 231–254. https://doi.org/10.1007/s00382-006-0181-1

Cai, W., Wang, G., Dewitte, B., Wu, L., Santoso, A., Takahashi, K., et al. (2018). Increased variability of eastern Pacific El Niño under greenhouse warming. *Nature*, *564*, 201–206. https://doi.org/10.1038/s41586-018-0776-9

Chen, C., Cane, M. A., Wittenberg, A. T., & Chen, D. (2017). ENSO in the CMIP5 simulations: Life cycles, diversity, and responses to climate change. *Journal of Climate*, *30*(2), 775–801. https://doi.org/10.1175/JCLI-D-15-0901.1

Clarke, A. J. (2008). *An introduction to the dynamics of El Niño and the Southern Oscillation*. Elsevier.

Collins, M., An, S. I., Cai, W., Ganachaud, A., Guilyardi, E., Jin, F. F., et al. (2010). The impact of global warming on the tropical Pacific Ocean and El Niño. *Nature Geoscience*, *3*(6), 391–397. https://doi.org/10.1038/ngeo868

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., et al. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28. https://doi.org/10.1002/qj.776

Cravatte, S., Kessler, W. S., Smith, N., Wijffels, S. E., Ando, K., Cronin, M., et al. (2016). *First Report of TPOS 2020 (GOOS-215)*. Retrieved from http://tpos2020.org/first-report

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Delworth, T. L., Cooke, W. F., Adcroft, A., Bushuk, M., Chen, J. H., Dunne, K. A., et al. (2020). SPEAR: The next generation GFDL modeling system for seasonal to multidecadal prediction and projection. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001895. https://doi.org/10.1029/2019MS001895

Déqué, M. (1997). Ensemble size for numerical seasonal forecasts. *Tellus*, *49*(1), 74–86. https://doi.org/10.3402/tellusa.v49i1.12212

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277–286. https://doi.org/10.1175/JCLI-D-16-0844.1

Ding, H., Newman, M., Alexander, M. A., & Wittenberg, A. T. (2020). Relating CMIP5 model biases to seasonal forecast skill in the tropical Pacific. *Geophysical Research Letters*, *47*, e2019GL086765. https://doi.org/10.1029/2019GL086765

Doi, T., Behera, S. K., & Yamagata, T. (2019). Merits of a 108-member ensemble system in ENSO and IOD predictions. *Journal of Climate*, *32*(3), 957–972. https://doi.org/10.1175/JCLI-D-18-0193.1

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., et al. (2021). The EC-Earth3 earth system model for the climate model intercomparison project 6. *Geoscientific Model Development Discussions*, 1–90. https://doi.org/10.5194/gmd-2020-446

Durack, P. J., Taylor, K. E., Eyring, V., Ames, S. K., Hoang, T., Nadeau, D., et al. (2018). Toward standardized data sets for climate model experimentation. *Eos*, *99*(10), 1029. https://doi.org/10.1029/2018EO101751

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 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. https://doi.org/10.5194/gmd-13-3627-2020

Fasullo, J. T., Phillips, A. S., & Deser, C. (2020). Evaluation of leading modes of climate variability in the CMIP archives. *Journal of Climate*, *33*(13), 5527–5545. https://doi.org/10.1175/JCLI-D-19-1024.1

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2014). Evaluation of climate models. In T. Stocker (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the Fifth assessment report of the Intergovernmental panel on climate change* (pp. 741–866). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020

Gleckler, P. J., Doutriaux, C., Durack, P. J., Taylor, K. E., Zhang, Y., Williams, D. N., et al. (2016). A more powerful reality test for climate models. *Eos*, *97*(12), 20–24. https://doi.org/10.1029/2016eo051663

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*, D06104. https://doi.org/10.1029/2007JD008972

Guilyardi, E., Capotondi, A., Lengaigne, M., Thual, S., & Wittenberg, A. T. (2020). ENSO modeling: History, progress, and challenges. In *El Niño Southern Oscillation in a changing climate* (pp. 199–226). American Geophysical Union. https://doi.org/10.1002/9781119548164.ch9

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., et al. (2020). Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, *13*(5), 2197–2244. https://doi.org/10.5194/gmd-13-2197-2020

Ham, Y. G., & Kug, J. S. (2014). ENSO phase-locking to the boreal winter in CMIP3 and CMIP5 models. *Climate Dynamics*, *43*, 305–318. https://doi.org/10.1007/s00382-014-2064-1

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

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., et al. (2021). Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1. *Journal of Climate*, *34*(8), 2923–2939. https://doi.org/10.1175/JCLI-D-20-0166.1

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55. https://doi.org/10.1175/JHM560.1

Kato, S., Rose, F. G., Rutan, D. A., Thorsen, T. J., Loeb, N. G., Doelling, D. R., et al. (2018). Surface irradiances of edition 4.0 Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) data product. *Journal of Climate*, *31*(11), 4501–4527. https://doi.org/10.1175/JCLI-D-17-0523.1

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kessler, W. S., Wijffels, S. E., Cravatte, S., Smith, N., Kumar, A., & Fujii, Y. (2019). *Second report of TPOS 2020 (GOOS-234)*. Retrieved from http://tpos2020.org/project-reports/second-report

Kim, Y. H., Min, S. K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes*, *29*, 100269. https://doi.org/10.1016/j.wace.2020.100269

Kirchmeier-Young, M. C., Zwiers, F. W., & Gillett, N. P. (2017). Attribution of extreme events in Arctic sea ice extent. *Journal of Climate*, *30*(2), 553–571. https://doi.org/10.1175/JCLI-D-16-0412.1

L'Heureux, M. L., Levine, A. F. Z., Newman, M., Ganter, C., Luo, J.-J., Tippett, M. K., & Stockdale, T. N. (2020). Chapter 10: ENSO prediction. In *El Niño Southern Oscillation in a changing climate* (pp. 227–246). American Geophysical Union. https://doi.org/10.1002/9781119548164.ch10

Lee, J., Sperber, K. R., Gleckler, P. J., Bonfils, C. J., & Taylor, K. E. (2019). Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dynamics*, *52*(7), 4057–4089. https://doi.org/10.1007/s00382-018-4355-4

Lee, J., Sperber, K. R., Gleckler, P. J., Taylor, K. E., & Bonfils, C. J. (2021). Benchmarking performance changes in the simulation of extratropical modes of variability across CMIP generations. *Journal of Climate*, *34*(17), 6945–6969. https://doi.org/10.1175/JCLI-D-20-0832.1

Leith, C. E. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, *102*(6), 409–418. https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2

Maher, N., Matei, D., Milinski, S., & Marotzke, J. (2018). ENSO change in climate projections: Forced response or internal variability? *Geophysical Research Letters*, *45*, 11390–11398. https://doi.org/10.1029/2018GL079764

McPhaden, M. J., Santoso, A., & Cai, W. (2020). *El Niño Southern Oscillation in a changing climate* (Vol. 253). John Wiley & Sons. https://doi.org/10.1002/9781119548164

McPhaden, M. J., Zebiak, S. E., & Glantz, M. H. (2006). ENSO as an integrating concept in earth science. *Science*, *314*(5806), 1740–1745. https://doi.org/10.1126/science.1132588

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F., et al. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. https://doi.org/10.1175/bams-88-9-1383

Milinski, S., Maher, N., & Olonscheck, D. (2020). How large does a large ensemble need to be? *Earth System Dynamics*, *11*, 885–901. https://doi.org/10.5194/esd-11-885-2020

Orbe, C., Van Roekel, L., Adames, Á. F., Dezfuli, A., Fasullo, J., Gleckler, P. J., et al. (2020). Representation of modes of variability in six US climate models. *Journal of Climate*, *33*(17), 7591–7617. https://doi.org/10.1175/JCLI-D-19-0956.1

Pennell, C., & Reichler, T. (2011). On the effective number of climate models. *Journal of Climate*, *24*, 2358–2367. https://doi.org/10.1175/2010JCLI3814.1

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., et al. (2021). Evaluating climate models with the CLIVAR 2020 ENSO metrics package. *Bulletin of the American Meteorological Society*, *102*(2), E193–E217. https://doi.org/10.1175/BAMS-D-19-0337.1

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., et al. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, *29*(11), 4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1

Praveen Kumar, B., Vialard, J., Lengaigne, M., Murty, V. S. N., & McPhaden, M. J. (2012). TropFlux: Air-sea fluxes for the global tropical oceans—Description and evaluation. *Climate Dynamics*, *38*(7–8), 1521–1543. https://doi.org/10.1007/s00382-011-1115-0

Praveen Kumar, B., Vialard, J., Lengaigne, M., Murty, V. S. N., McPhaden, M. J., Cronin, M. F., et al. (2013). TropFlux wind stresses over the tropical oceans: Evaluation and comparison with other products. *Climate Dynamics*, *40*(7–8), 2049–2071. https://doi.org/10.1007/s00382-012-1455-4

Rayner, N. A. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Ropelewski, C. F., & Halpert, M. S. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, *115*, 1606–1626. https://doi.org/10.1175/1520-0493(1987)115<1606:GARSPP>2.0.CO;2

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble. Part 1: Model evaluation in the present climate. *Journal Geophysical Research: Atmospheres*, *118*, 1716–1733. https://doi.org/10.1002/jgrd.50203

Stevenson, S., Fox-Kemper, B., Jochum, M., Rajagopalan, B., & Yeager, S. G. (2010). ENSO model validation using wavelet probability analysis. *Journal of Climate*, *23*, 5540–5547. https://doi.org/10.1175/2010JCLI3609.1

Stevenson, S., Wittenberg, A. T., Fasullo, J., Coats, S., & Otto-Bliesner, B. (2021). Understanding diverse model projections of future extreme El Niño. *Journal of Climate*, *34*(2), 449–464. https://doi.org/10.1175/JCLI-D-19-0969.1

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian earth system model version 5 (CanESM5. 0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., et al. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

van Oldenborgh, G. J., Philip, S. Y., & Collins, M. (2005). El Niño in a changing climate: A multi-model study. *Ocean Science*, *1*(2), 81–95. https://doi.org/10.5194/os-1-81-2005

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. https://doi.org/10.1029/2019MS001683

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., et al. (2016). A global repository for planet-sized experiments and observations. *Bulletin of the American Meteorological Society*, *97*(5), 803–816. https://doi.org/10.1175/BAMS-D-15-00132.1

Wills, R. C., Battisti, D. S., Armour, K. C., Schneider, T., & Deser, C. (2020). Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *Journal of Climate*, *33*, 8693–8719. https://doi.org/10.1175/JCLI-D-19-0855.1

Wittenberg, A. T. (2009). Are historical records sufficient to constrain ENSO simulations? *Geophysical Research Letters*, *36*, L12702. https://doi.org/10.1029/2009GL038710

Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A., & Zeng, F. (2014). ENSO modulation: Is it decadally predictable? *Journal of Climate*, *27*(7), 2667–2681. https://doi.org/10.1175/JCLI-D-13-00577.1

Xie, P., & Arkin, P. A. (1997). Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, *78*(11), 2539–2558. https://doi.org/10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., et al. (2020). The Australian Earth System Model: ACCESS-ESM1. 5. *Journal of Southern Hemisphere Earth Systems Science*, *70*(1), 193–214. https://doi.org/10.1071/ES19035