On the robustness of the 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[3], Eric Guilyardi[4], Andrew T. Wittenberg[5], Michael J. McPhaden[6], and Giuliana Pallotta[1]

[1]Lawrence Livermore National Laboratory (DOE)
[2]NOAA Pacific Marine Environmental Laboratory
[3]Lawrence Livermore Natinal Laboratory
[4]IPSL /LOCEAN
[5]NOAA Geophysical Fluid Dynamics Laboratory
[6]NOAA/PMEL

November 23, 2022

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. Utilizing the recently developed CLIVAR ENSO Metrics Package (Planton et al., 2021), we estimate the ensemble size (N) needed to assess a model's ability to capture observed ENSO behavior. Using the larger ensembles available from CMIP6 and the CLIVAR Large Ensemble Project, we find that larger ensembles are needed to robustly capture baseline ENSO characteristics (N > 65) and physical processes (N > 50) than the background climatology (N [?] 12) and remote ENSO teleconnections (N [?] 6). While these results vary somewhat across metrics and models, our study highlights that ensembles are required to robustly evaluate simulated historical ENSO behavior, and provide initial guidance for designing model ensembles to reliably evaluate and compare ENSO simulations.

Hosted file

`paper_enso_supplement_v20210628_final_2.docx` available at https://authorea.com/users/541438/articles/600621-on-the-robustness-of-the-evaluation-of-enso-in-climate-models-how-many-ensemble-members-are-needed

Hosted file

`essoar.10507474.1.docx` available at https://authorea.com/users/541438/articles/600621-on-the-robustness-of-the-evaluation-of-enso-in-climate-models-how-many-ensemble-members-are-needed

On the robustness of the 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[1,3], Eric Guilyardi[4,5], Andrew T. Wittenberg[6], Michael J. McPhaden[2], Giuliana Pallotta[1]

[1]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, California, USA

[2]NOAA Pacific Marine Environmental Laboratory, Washington, USA

[3]Retired

[4]LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France

[5]NCAS-Climate, University of Reading, UK

[6]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

June 2021

Submitted to *GRL*

* Corresponding Author:

Dr. Jiwoo Lee

P.O. Box 808, L-103, Livermore, CA 94551, USA

lee1043@llnl.gov

Key Points

- We examine the performance of climate models in reproducing ENSO, including multiple ensemble members performed with individual models.
- There is broad range in the relative performance of models, with intrinsic variability influencing the robustness of many characteristics.
- We quantify the ensemble sizes required to characterize various important aspects of ENSO using large ensembles and Monte-Carlo sampling.

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 intrinsic 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.

Keywords

ENSO, CMIP6, Large Ensemble, CLIVAR ENSO Metrics, PCMDI Metrics Package (PMP), Monte-Carlo Sampling

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. Utilizing the recently developed CLIVAR ENSO Metrics Package (Planton et al., 2021), we estimate the ensemble size (N) needed to assess a model's ability to capture observed ENSO behavior. Using the larger ensembles available from CMIP6 and the CLIVAR Large Ensemble Project, we find that larger ensembles are needed to robustly capture baseline ENSO characteristics (N > 65) and physical processes (N > 50) than the background climatology (N 12) and remote ENSO teleconnections (N 6). While these results vary somewhat across metrics and models, our study highlights that ensembles are required to robustly evaluate simulated historical ENSO behavior, and provide initial guidance for designing model ensembles to reliably evaluate and compare ENSO simulations.

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 1–3°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 inter-model 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).

Recently, an International CLIVAR Pacific Region Panel team 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 the aspect to focus on: 1) background climatology and basic ENSO characteristics, 2) ENSO's worldwide

teleconnections, and 3) ENSO's internal processes and feedbacks represented in historical simulations of GCMs. However, in a multi-model ensemble, it can be difficult to tease apart the role of internal variability (sampling variation) versus model formulation (different physical parameterizations, resolutions, dynamical cores, representations of fluxes between ocean and atmosphere, etc.) in generating inter-model 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, multi-millennium simulations have demonstrated that ENSO's characteristics (amplitude, spectrum, irregularity, and spatial pattern) can vary substantially on multidecadal and multi-centennial scales, purely due to internal variability (Stevenson et al., 2010; Wittenberg 2009; Wittenberg et al., 2014). Thus it is essential to account for this internal variability when evaluating or comparing models, by using a sufficient run duration and ensemble size 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; Pennel & Reichler, 2011; Wills et al., 2020), 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. 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 address the following question: *What is the minimum number of ensemble members needed to obtain robust results for characterising 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 1850 and run to 2014 with close to observed time-varying natural and anthropogenic forcings (Durack et al., 2018). We use all available historical members from 41 CMIP6 models (Table 1) and 2 models from the Single-Model Initial condition Large Ensembles (SMILEs) Project (Deser et al., 2020).

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 inter-model and inter-member 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 datasets (list of variables and epochs are provided in supplement, as Table S1). The analysis is conducted using the PCMDI Metrics Package (PMP, Gleckler et al., 2016) framework in which the CEM2021 is implemented. In the study of Planton et al. (2021), the CEM2021 metrics were applied to CMIP6 simulations using one ensemble member per model. In this study, we apply the CEM2021 metrics to all available ensemble members of the CMIP6 models, to assess the robustness of model skill.

To estimate the 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 initialization procedures, physical parameterizations, and forcings), to capture the ensemble spread caused by internal variability. The LEs for 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 SMILEs (models marked with asterisk in Table 1). For each LE model and metric, a random sample of N members (pseudo-ensemble or *PE*), with N ranging from 1 to the full ensemble size, is drawn from the ensemble. We generate 1000 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 the supplementary material.

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 Fig. 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 climate (Gleckler et al., 2008; Flato et al., 2014), indices of temperature and precipitation extremes (Sillmann et al., 2013; Kim et al., 2020), extratropical modes of variability (Lee et al., 2019, 2021), and ENSO (Bellenger et al., 2014; Planton et al., 2021). The color scale in Fig. 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. Fig. 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 datasets.

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 dataset, with vertical line markers showing the results for individual members. These plots collectively illustrate the inter-model skill differences, as well as the inter-member (internal) variability in the errors for each model, for those selected three example metrics (analysis for other metrics are available in the supplement, Fig. S1). For the *Equatorial SST Bias* metric (Fig. 2a), as well as others based on mean state characteristics (Fig. S1), the inter-member spread due to internal variability is very narrow. The internally-generated spread is larger for *ENSO Amplitude* (Fig. 2b), as large as 1 of inter-model spread in general. For *ENSO Asymmetry* (Fig. 2c), there are

some members that nearly match the observations while others differ strongly from observed (e.g., CanESM2). For metrics with such behavior, multiple members are needed to obtain an accurate assessment of skill relative to observations. Figure 2 also shows that the inter-member spread is model dependent.

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 four 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 dataset. The pseudo-ensemble means from the "without replacement" sampling results converges to the full ensemble mean. On the contrary, pseudo-ensemble 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 sample in subset) where at least 95% of the "with replacement" pseudo-ensemble means fall within 10% of the mean of the full ensemble. The N_{min} is estimated to be 1 for *Equatorial SST Bias*, and 8 for *ENSO Amplitude*, while entire ensemble size (32) is not large enough for *ENSO Asymmetry,* for the IPSL-CM6A-LR model and for the epoch lengths of the reference dataset.

We repeated the aforementioned analysis to estimate the N_{min} for individual metrics, and from the four large ensembles mentioned 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, 17-65 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* and *Diversity* metrics require the largest N_{min}, 65. For the metrics evaluating physical processes (cyan), the N_{min} varies across from 1-50. The two largest include the *SST-Taux Feedback* metric, examining 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 (5N-5S, 150W-90W).

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. Several ensembles exceeded 20 members (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). We then 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 $>$ 65) and physical processes (N $>$ 50) than they do for the background climatology (N 12) and ENSO related teleconnections (N 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. Considering the early studies of how climate change affects ENSO amplitude in the future were based on CMIP model simulations with far fewer than 10 and often only 1 ensemble member (e.g., Collins et al., 2010; Meehl et al., 2007; van Oldenborgh et al., 2005), increasing ensemble size would help strengthen robustness of the results.

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 is particularly designed to address the three basic science questions identified in Planton et al. (2021), and because each of them incorporate some of the baseline ENSO characteristics, our findings suggest that to fully address each question requires a substantial ensemble size (N $>$ 65 for *Performance* and *Process*, N 47 for *Teleconnection Metrics Collections* of CEM2021), reinforcing the importance of the large ensembles.

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 for processes are in effect (e.g., Atwood et al., 2017; Chen et al., 2017). One possible avenue of evaluation is to subsample simulated ENSO's variability that is consistent with the range of present observations as a basis for more rigorous assessment for GCMs. But given the role of multi-decadal ENSO variability possibly extending to much 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.

Acknowledgement

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. We thank the CLIVAR Pacific Panel members and ENSO experts for a number of fruitful discussions. We 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 altimeter products were produced by Ssalto/Duacs and distributed by Aviso+, with support from Cnes (https://www.aviso.altimetry.fr). 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. This is PMEL contribution no. 5276. We 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.

Data Availability Statement

Observation-based reference datasets are available at their providers' websites: AVISO (https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products.html), ERA-Interim (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim), GPCPv2.3 (https://psl.noaa.gov/data/gridded/data.gpcp.html) and TropFlux

(https://incois.gov.in/tropflux/). Simulations of CESM1-CAM5 and CanESM2 models for the SMILEs are available at https://www.cesm.ucar.edu/projects/community-projects/MMLEA/. CMIP6 simulations are available through the ESGF at https://esgf-node.llnl.gov/projects/cmip6/. The results of this study will be released through the DOE's Coordinated Model Evaluation Capabilities (CMEC) website (https://cmec.llnl.gov/) and during the meantime they are available upon request to the authors.

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*(7), 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

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.

Cravatte, S., Kessler, W. S., Smith, N., Wijffels, S. E., Ando, K., Cronin, M., ... & Wittenberg, A. (2016). First Report of TPOS 2020. GOOS-215, 200 pp. [Available online at 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*(3), e2019MS001895. https://doi.org/10.1029/2019MS001895

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

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

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*(5), 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

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2014). Evaluation of climate models. In *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., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, *113*(D6). https://doi.org/10.1029/2007JD008972

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.

Guilyardi, E., Capotondi, A., Lengaigne, M., Thual, S., & Wittenberg, A. T. (2020). ENSO Modeling: History, Progress, and Challenges. *El Niño Southern Oscillation in a Changing Climate*, 199-226. 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

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*. [Available online at 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

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*, 1-70. 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

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

Maher, N., Matei, D., Milinski, S., & Marotzke, J. (2018). ENSO change in climate projections: Forced response or internal variability? *Geophysical Research Letters*, 45, 11-390. https://doi.org/10.1029/2018GL079764

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.

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

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.

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

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

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

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

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

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

Table 1. List of models and ensemble sizes. Models having 20 or more "initial condition" ensemble members (i.e., varying initial condition but fixed physical parameterizations) are marked in **bold** and with an asterisk (*) and used for determining the required ensemble size in Section 3.2. Models marked with a hash (#) 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/.

Participation	Model	Members	Model	Members
CMIP6	ACCESS-CM2	3	GFDL-CM4	1
	ACCESS-ESM1-5	**30***	GFDL-ESM4	3
	AWI-CM-1-1-MR	5	GISS-E2-1-G	47#
	AWI-ESM-1-1-LR	1	GISS-E2-1-G-CC	1
	BCC-CSM2-MR	3	GISS-E2-1-H	25#
	BCC-ESM1	3	HadGEM3-GC31-LL	5
	CAMS-CSM1-0	3	HadGEM3-GC31-MM	4
	CanESM5	**65***	INM-CM4-8	1
	CanESM5-CanOE	3	INM-CM5-0	10
	CESM2	11	IPSL-CM5A2-INCA	1
	CESM2-FV2	3	**IPSL-CM6A-LR**	**32***
	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***
	CNRM-CM6-1	**29***	**MIROC6**	**50***
	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***	NESM3	5
	EC-Earth3-AerChem	2	**NorCPM1**	**30***
	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		
SMILEs	**CESM1-CAM5**	**40***	**CanESM2**	**50***

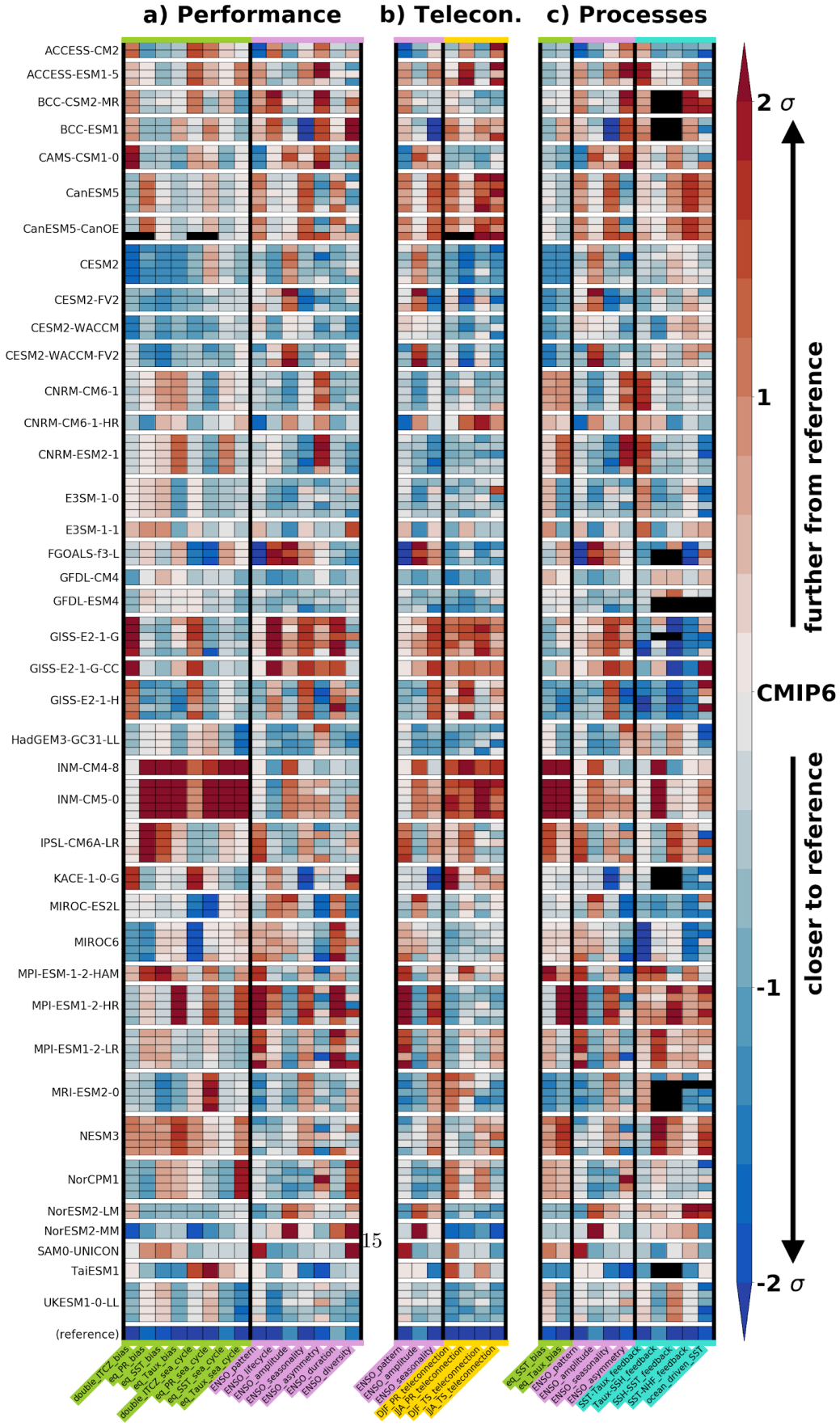

a) Performance
b) Telecon.
c) Processes
ACCESS-CM2
ACCESS-ESM1-5
BCC-CSM2-MR
BCC-ESM1
CAMS-CSM1-0
CanESM5
CanESM5-CanOE
CESM2
CESM2-FV2
CESM2-WACCM
CESM2-WACCM-FV2
CNRM-CM6-1
CNRM-CM6-1-HR
CNRM-ESM2-1
E3SM-1-0
E3SM-1-1
FGOALS-f3-L
GFDL-CM4
GFDL-ESM4
GISS-E2-1-G
GISS-E2-1-G-CC
GISS-E2-1-H
HadGEM3-GC31-LL
INM-CM4-8
INM-CM5-0
IPSL-CM6A-LR
KACE-1-0-G
MIROC-ES2L
MIROC6
MPI-ESM-1-2-HAM
MPI-ESM1-2-HR
MPI-ESM1-2-LR
MRI-ESM2-0
NESM3
NorCPM1
NorESM2-LM
NorESM2-MM
SAM0-UNICON
TaiESM1
UKESM1-0-LL
(reference)
double_ITCZ_bias
eq_PR_bias
eq_SST_bias
eq_Taux_bias
double_ITCZ_sea_cycle
eq_PR_sea_cycle
eq_SST_sea_cycle
eq_Taux_sea_cycle
ENSO_pattern
ENSO_lifecycle
ENSO_amplitude
ENSO_seasonality
ENSO_asymmetry
ENSO_duration
ENSO_diversity
ENSO_pattern
ENSO_amplitude
ENSO_seasonality
PR_teleconnection
DJF_PR_teleconnection
JJA_PR_teleconnection
DJF_TS_teleconnection
JJA_TS_teleconnection
eq_SST_bias
eq_Taux_bias
ENSO_pattern
ENSO_amplitude
ENSO_seasonality
ENSO_asymmetry
SST-Taux_feedback
Taux-SSH_feedback
SSH-SST_feedback
SST-NHF_feedback
pcean_driven_SST
2 σ
1
CMIP6
-1
-2 σ
further from reference
closer to reference

Figure 1. ENSO Metrics *portrait plot* for CMIP6 with results for *Performance*, *Teleconnection*, and *Processes* metrics defined as part of CEM2020 (Planton et al. 2021). Multiple realizations are shown as available, with a maximum of 5 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 non-dimensionally, as a difference from the multi-model mean error (MMME) computed from all CMIP6 divided by the inter-model standard deviation ( ) within each metric column. A displayed value of 0 (white) 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. 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 and highlighted according to their application (metrics collection or MC), evaluating background climatology (light green), basic ENSO characteristics (magenta), teleconnections (yellow), or physical processes (cyan).

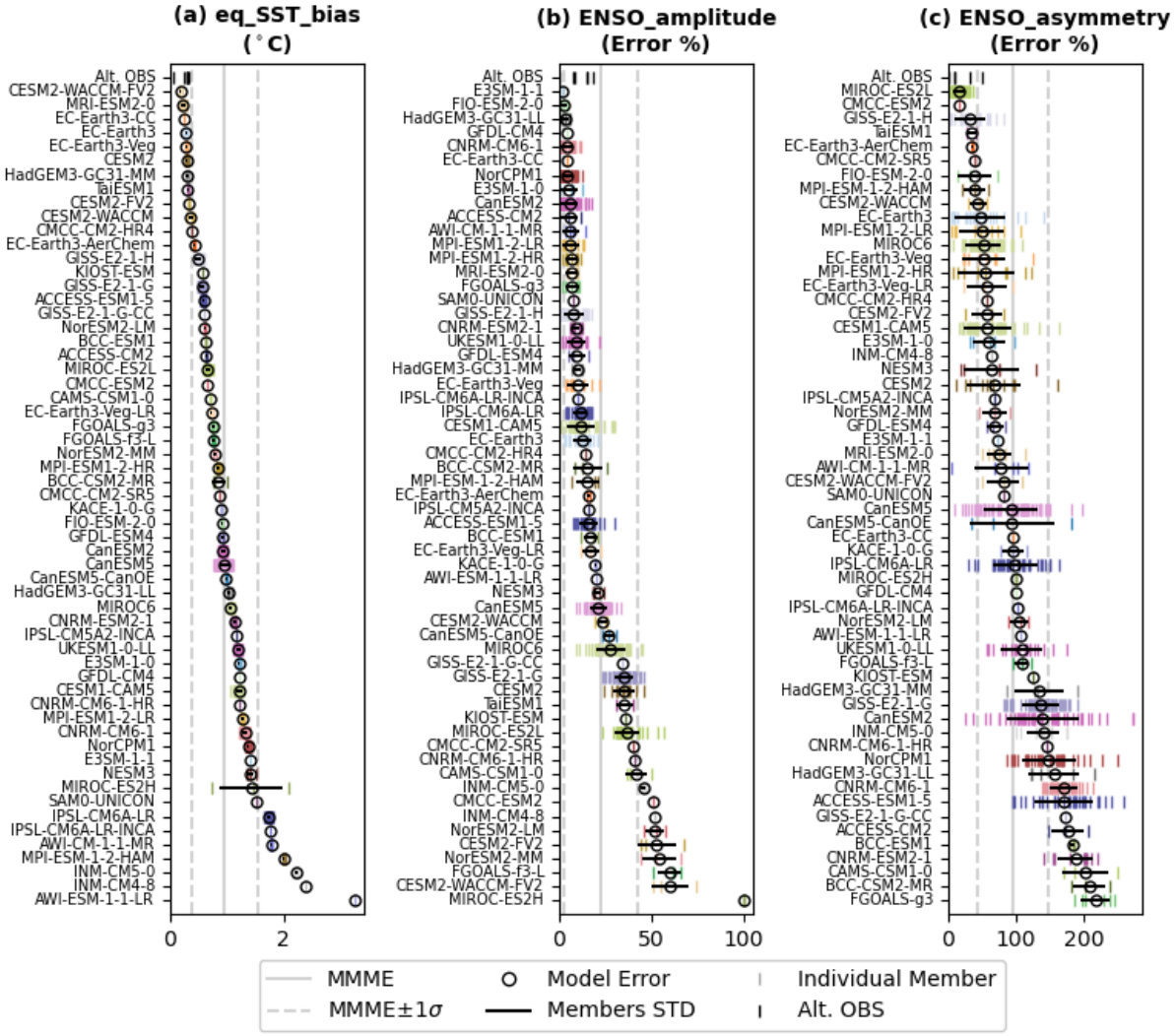
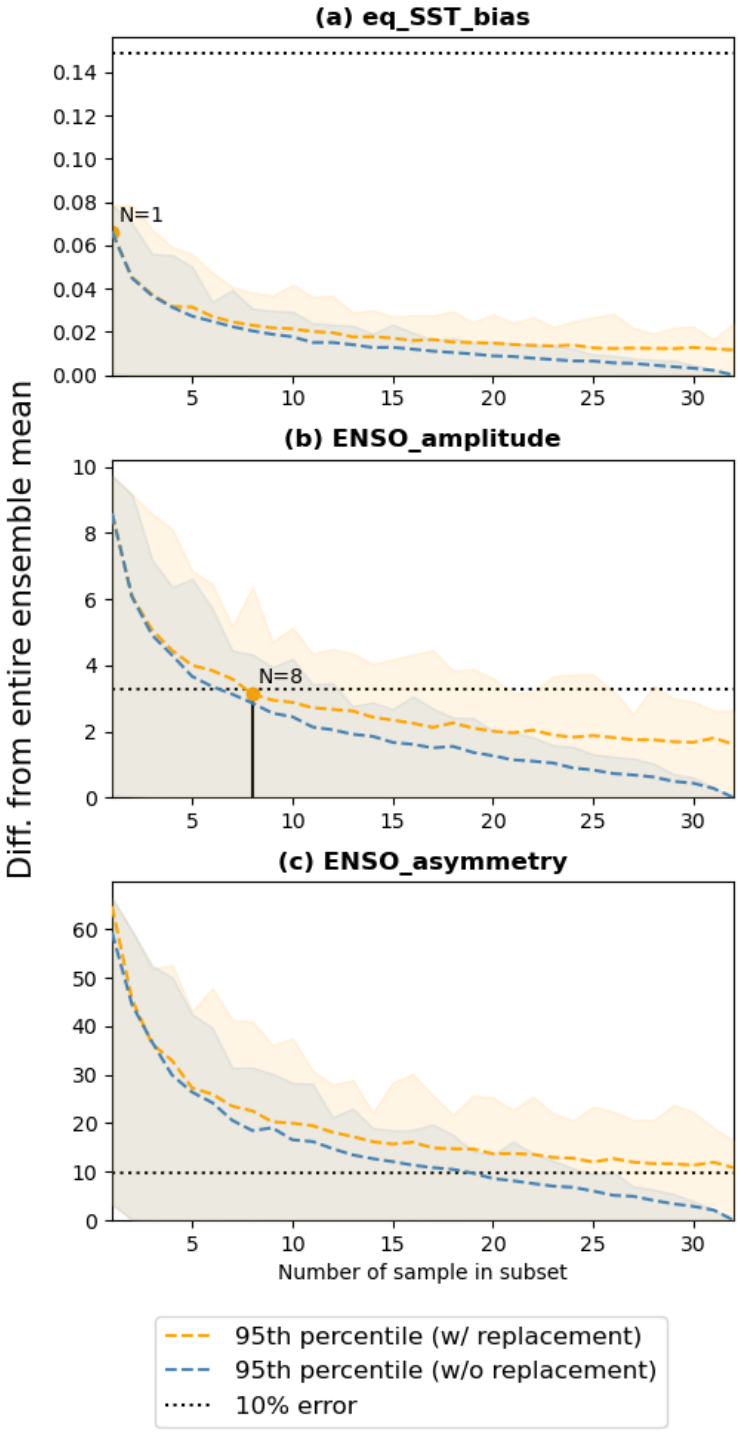


Figure 2. Error metrics calculated for ensemble simulations from the CMIP6 models and Large ensembles of CESM1-CAM5 and CanESM2. Lines represent standard deviation of error metrics for individual model ensembles, with circles denoting the average of all members for any given model. Three representative metrics are shown: (a) *Equatorial SST Bias*, (b) *ENSO Amplitude* and (c) *Asymmetry*, with results from other metrics in Supplemental Fig. 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 dashed lines are for multi-model mean error and its ±1 standard deviation, respectively. Error metrics calculated for alternative observation-based datasets (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 pseudo-ensembles sampled with (orange) or without (blue) replacement at different sample sizes (abscissa). Three representative metrics are shown: (a) *Equatorial SST Bias*, (b) *ENSO Amplitude* and (c) *Asymmetry*. Annotated N indicates the minimum ensemble size (N_{min}) for which at least 95% of the "with replacement" pseudo-ensemble 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 pseudo-ensemble without replacement (blue) converges toward the mean of the full ensemble.


(a) eq_SST_bias
N=1
(b) ENSO_amplitude
N=8
(c) ENSO_asymmetry
Diff. from entire ensemble mean
Number of sample in subset
95th percentile (w/ replacement)
95th percentile (w/o replacement)
10% error

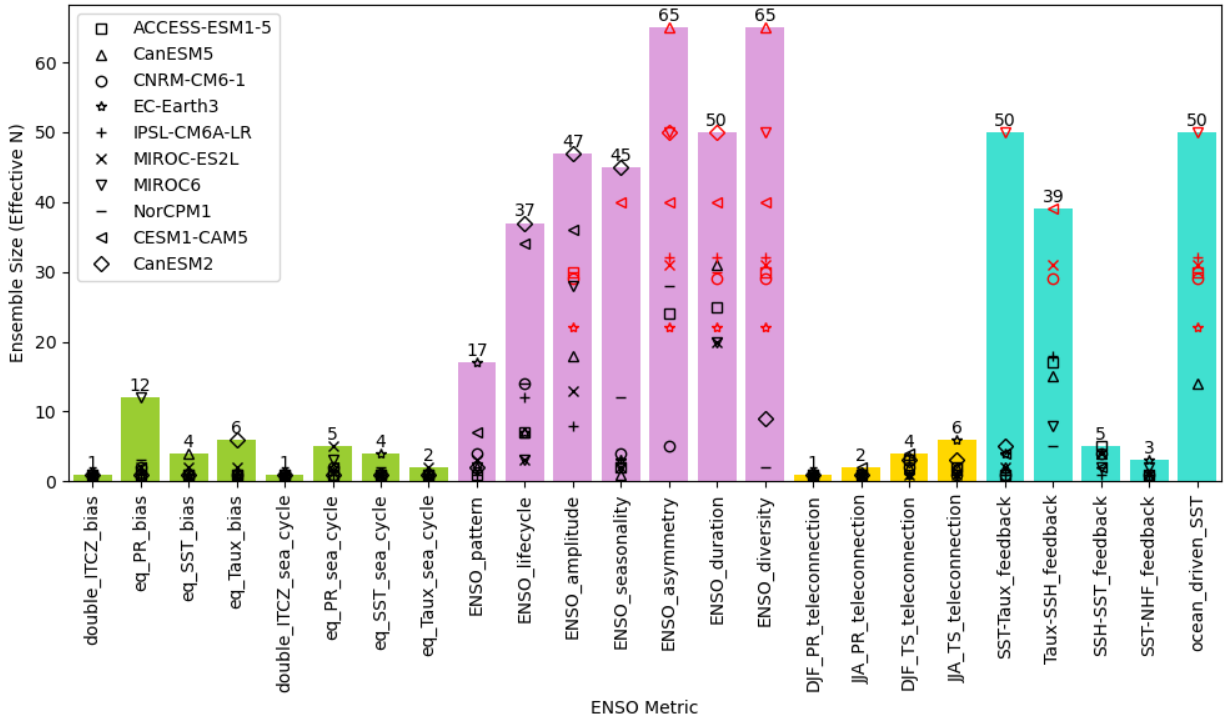


Figure 4. Minimum number of required ensemble members (N_{min} of Fig. 3) for individual metrics obtained from models with at least 20 initial condition ensemble members (see Table 1). Each vertical bar indicates the maximum N_{min} (across the 4 models) for the given metric. Metrics are listed along 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 ENSO characteristics (magenta), teleconnections (yellow), and physical processes (cyan).